

F

Face Recognition Models

The human face is a highly meaningful stimulus that provides us with diverse information for adaptive social interaction with people. Our ability to recognize faces is remarkably accurate and long lasting. We are also able to categorize people along a number of visual dimensions including sex, race, and age and can readily interpret facial expression. The challenges associated with encoding and interpreting this information have become evident over the last two decades as psychologists, computer scientists, and cognitive scientists have endeavored to formulate computational models of these processes. The resultant models give insight into the complexity of the problems solved by the human brain in perceiving, representing, and remembering faces. In this article, computational approaches to modeling the perception, categorization, and recognition of human faces will be presented. The properties of the human face as a visual stimulus are described first, followed by definitions of the relevant tasks we perform with faces. The steps involved in modeling these tasks are reviewed next, and representative approaches for modeling individual tasks are discussed. Finally, the article closes with a few open questions in face recognition modeling.

1. The Human Face as a Visual Stimulus

The human face is a complex three-dimensional object defined by the structure of the skull and by the shape, texture, and pigmentation of the overlying skin and tissue. All faces share a basic set of features (e.g., eyes, nose, and mouth, etc.) arranged in a well-defined configuration (eyes above the nose, etc). Individual faces comprise virtually limitless variations on this standard theme. To recognize an individual from a face, we must attend to the information that makes the face unique. To categorize a face we must extract and encode the information that a face shares with an entire category of faces (e.g., male faces), but which distinguishes the category from competing categories (e.g., female faces) (see *Face Recognition: Psychological and Neural Aspects*).

2. The Tasks

'Face recognition' models commonly encompass a range of tasks, including recognition, identification, verification, categorization, and the analysis of facial

expression (see *Facial Expressions*). Face recognition refers to the judgment of whether or not a particular face is 'known.' Face identification refers to the retrieval of information about the 'owner' of the face, such as a name or context of encounter. Face verification refers to a decision about whether a particular face image belongs to a particular individual. Is this person John Doe? Face verification is a common goal of face algorithms developed for security systems.

3. Modeling: A Step by Step Approach

Face recognition models involve: (a) preprocessing algorithms to encode facial 'features' and (b) the application of this information to solve particular tasks.

3.1 Preprocessing Algorithms

3.1.1 Aligning faces. All models that involve the analysis of a three-dimensional object from a two-dimensional image begin with the process of aligning the images into a common coordinate system. This facilitates feature extraction and comparison. Most current face recognition models operate effectively only with frontal images, tolerating only minimal changes in viewpoint. The alignment procedure employed in different models varies both in precision and in the degree of automaticity with which it is accomplished (i.e., by hand or by a computer algorithm). At the most basic level, alignment involves image translation, rotation, and scaling procedures implemented to assure that the eye levels are equivalent and that the centers of the foreheads correspond. More precise alignment is possible with *morphing* techniques that 'warp' individual faces into the 'average face' (Craw and Cameron 1991). To morph a face into another face (e.g., the average face), control points are located on the two faces (usually by hand). These consist of facial landmarks (e.g., corners of the eyes) and supplemental points (e.g., equally spaced points along the eyebrows). Using these points as a guide, each face is warped into the shape of the average face, yielding a correspondence of the control points across all faces. This alignment enables a separable encoding of the two-dimensional shape of the face and the image intensity information. Automated solutions to this correspondence problem have been implemented using all of the pixels or surface samples of

the face rather than just a subset (Beymer and Poggio 1996, Blanz and Vetter 1999). These algorithms employ elaborated optic flow computations and work well on sets of faces for which correspondence is relatively easy to establish, (e.g., faces without hair that are pre-aligned with the translation method). Though difficult to achieve, when successful, complete alignment provides a powerful basis for synthesizing faces with arbitrary shapes and faces composed of intensity composites of other faces (Blanz and Vetter 1999).

A different approach to alignment is represented by the work of Lades et al. (1993) who developed a face recognition algorithm base on the dynamic link architecture. This algorithm combines alignment with identification. The model operates by placing a deformable grid over the target face, sampling the face at the grid vertices. The sampling is done with a series of oriented Gabor wavelets, designed to emulate the orientation specific neurons of visual cortex. The connectors between the vertices are allowed to deform elastically, enabling a resampling of the image until the best fit is obtained. The deformation parameters of this fit serve as the face representation, which is matched to the faces in the database to identify the best match.

3.1.2 Encoding and representing faces. The information in the aligned faces must be quantified in a way that enables recognition, identification, verification, categorization, and the analysis of expression in the model. What are the features of the face? We commonly think of the features of a face as its eyes, nose, and mouth. Descriptions of these features, such as those an eyewitness might provide, are inadequate for communicating enough information about an individual face to distinguish it from competing candidates. Geometrical measures, e.g., distance between eyes, have proved similarly inadequate (Laughery et al. 1981). More recent models have employed relatively raw perceptual codes, including roughly aligned images and three-dimensional surfaces, including pigmentation information. Another code, common since the advent of morphing technology, involves a two-component separable encoding of the two-dimensional face shape and the image intensities. The 'shape' part of this code is defined as the deformation of the control points from the control points in the average face. The 'shape-free' part of the code consists of a 'shape standardized' two-dimensional array of image intensities created by warping an individual face into the shape of the average face.

In current computational and psychological models of face recognition, further analysis of these perceptual codes is carried out using a principal component analysis (PCA) (Sirovich and Kirby 1987; for a review see Valentin et al. 1994). In the US Government's tests of automatic face recognition algorithms between

1994–7, five of the seven algorithms tested used PCA. PCA is a statistical method for describing a set of correlated variables using a smaller number of uncorrelated or orthogonal variables. The uncorrelated variables are called eigenvectors or principal components (PCs), denoted u_i and play the role of 'features' for describing the faces. PCs can be considered features in the sense that any individual face, f , can be expressed as a linear combination of the PCs, $\sum_i w_i u_i$, where the weights are the dot products, $w_i = u_i^T f$, between the faces and PCs. Because PCA is applied usually to images/surfaces, the PCs are also images/surfaces. Thus, individual faces can be synthesized as a linear combination of the PC images/surfaces. In geometrical terms, the PCA creates a multidimensional space in which the PCs define the axes of the space and individual faces are points in the space. The coordinates of a face in the space are the weights that specify the face's value on each PC feature. Note also, that three-layer back propagation networks can extract facial features similarly when they are trained to reconstruct faces through a bottleneck of hidden units. The hidden units of these auto-encoders have been shown to derive rotated versions of the PCs space (Cottrell et al. 1987).

PCA has appeal as a psychological model of face perception and memory for several reasons. First, it is consistent with psychological theories that posit a 'face space' metaphor for human face memory (Valentine 1991). By this metaphor, faces can be thought of as points in a multidimensional space, with the distance between faces a measure of their similarity. At the center of the face space is the average or 'prototype' face. The prototype, a central concept in psychological studies of face recognition, is invoked to explain the role of face typicality in predicting recognition performance. Typical faces, thought to be close to the prototype, are recognized less accurately than distinctive faces. This occurs presumably due to the greater density of faces close to the prototype, causing more confusion among typical faces than among distinctive faces. The prototype is also used as a reference face in creating automatic caricatures. Caricatures can be created by 'moving a face' away from the average in the face space. This results in a more distinctive and recognizable version of the same face.

Second, the features that emerge from PCA are derived from the experience of the model. The role of experience in face recognition performance has been established perhaps most clearly in the phenomenon of the 'other-race effect'—the finding that people recognize faces of their own-race more accurately than faces of other-races. This effect is predicted when the PCA is applied to a majority of faces of one race, and a smaller number of faces of other races. Because PCA derives its features from the statistical structure of the input faces, the resultant features are most appropriate for describing the majority race of faces. Consequently, less distinct encodings of minority race faces

result because these faces are not well characterized by the features extracted primarily from the majority race of faces.

4. Tasks

4.1 Recognition

The quality of the stimulus representation determines the difficulty of the recognition or classification task. With a PCA-based representation, face recognition models can be implemented in a relatively simple way. A face is considered 'known' when an image of the individual was part of the input used to create the PCA space. The most common 'recognition' algorithm implements both recognition and identification. A target face is projected into the space and the distance to all other faces in the space is assessed. The nearest neighbor is chosen as the identity of the target face. Recognition can be implemented by setting a threshold distance, beyond which a target face is declared 'unknown.' An alternative and computationally more expedient algorithm for recognition assesses the representation error incurred by projecting the target face into the space. A threshold tolerance for error is used to determine whether a target face is known or novel.

4.2 Categorization

To categorize faces by sex, race, or age, individual exemplar faces must be assigned to different categories based on visually accessible facial features. Face categorization has been approached with supervised connectionist or neural network classifiers, such as the perceptron (see *Perceptrons*). The models use examples to learn the mapping between face representations and categories. Numerous sex categorization models have been implemented and have been found to perform at or near human performance levels. A similarly structured race classifier has also been implemented, though the imbalance of experience most people have for the faces of different races must be implemented also to model human performance accurately. Finally, little work has been done on categorizing faces by age, though two complementary models of facial aging make use of morphing and caricaturing techniques, respectively. The former simulates aging by morphing individual faces toward the average of older faces (Burt and Perrett 1995). The latter simulates aging by caricaturing the three-dimensional head structure relative to a mean of young adult faces. Surprisingly, this results in an aged face (O'Toole et al. 1997).

4.3 Facial Expression Analysis

Models for categorizing faces by expression have been implemented in ways similar to sex and race classifi-

cation models, but with somewhat less success. These models operate by mapping images of faces on to expression categories using supervised learning techniques. Representations have varied from aligned images, to PCAs of faces preprocessed by the Gabor wavelet filters described previously. Performance has been found to be well above chance, though still short of human performance on a similar task with similar stimuli. Facial expression analysis is currently a very active area of research and more published work on this problem is expected in the near future.

5. Open Questions

Despite the clear successes of face recognition models in the 1980s and 1990s, the problem of recognizing faces from different viewpoints remains an unsolved challenge for models. Though part and parcel of the larger unsolved inverse optics problem of computer vision, the domain of faces may be more accessible due to the specific nature of face recognition as a within category problem. Some promising lines of research have begun and may soon yield new insights into this difficult problem (Edelman 1999).

See also: Linear Algebra for Neural Networks; Object Recognition: Theories; Recognition Memory, Psychology of; Vision, Low-level Theory of; Visual Perception, Neural Basis of

Bibliography

- Beymer D, Poggio T 1996 Image representations for visual learning. *Science* **272**: 1905–9
- Blanz V, Vetter T 1999 A morphable model for the synthesis of 3D faces. *ACM SIGGRAPH Proceedings*, pp. 187–94
- Burt D M, Perrett D I 1995 Perception of age in adult Caucasian male faces: Computer graphic manipulation of shape and colour information. *Proceedings of the Royal Society London B Bio* **259**: 137–43
- Cottrell G W, Munro P, Zipser D 1987 Learning internal representations of gray scale images: An example of extensional programming. In: *Proceedings of the 9th Annual Cognitive Science Society*, Erlbaum, Hillsdale, NJ
- Craw I, Cameron P 1991 Parameterizing images for recognition and reconstruction. In: Mowforth P (ed.) *Proceedings of the British Machine Vision Conference*. Springer-Verlag, London
- Edelman S 1999 *Representation and Recognition in Vision*. MIT Press, Cambridge, MA
- Lades M, Vorbrüggen J C, Buhmann J, Lange J, von der Malsburg C, Wiskott R P, Konen R W 1993 Distortion invariant face recognition in dynamic link architectures. *IEEE Transactions on Computers* **42**: 300–11
- Laughery K, Rhodes B, Batten G 1981 Computer-guided recognition and retrieval of facial images. In: Davies G, Ellis H, Shepherd J (eds.) *Perceiving and Remembering Faces*. Academic Press, London
- O'Toole A J, Vetter T, Volz H, Salter E M 1997 Three-dimensional caricatures of human heads: Distinctiveness and the perception of facial age. *Perception* **26**: 719–32

ff

ff

ff

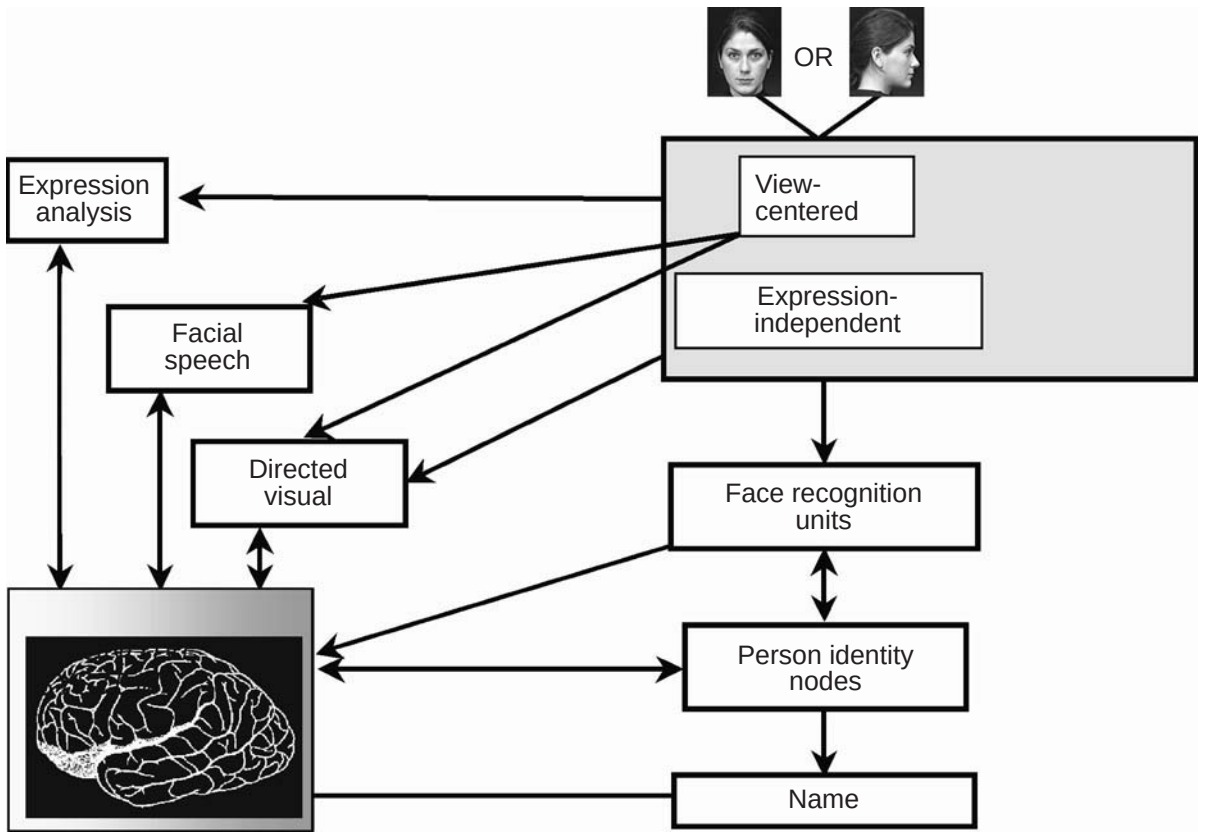


Figure 1

ff

ff

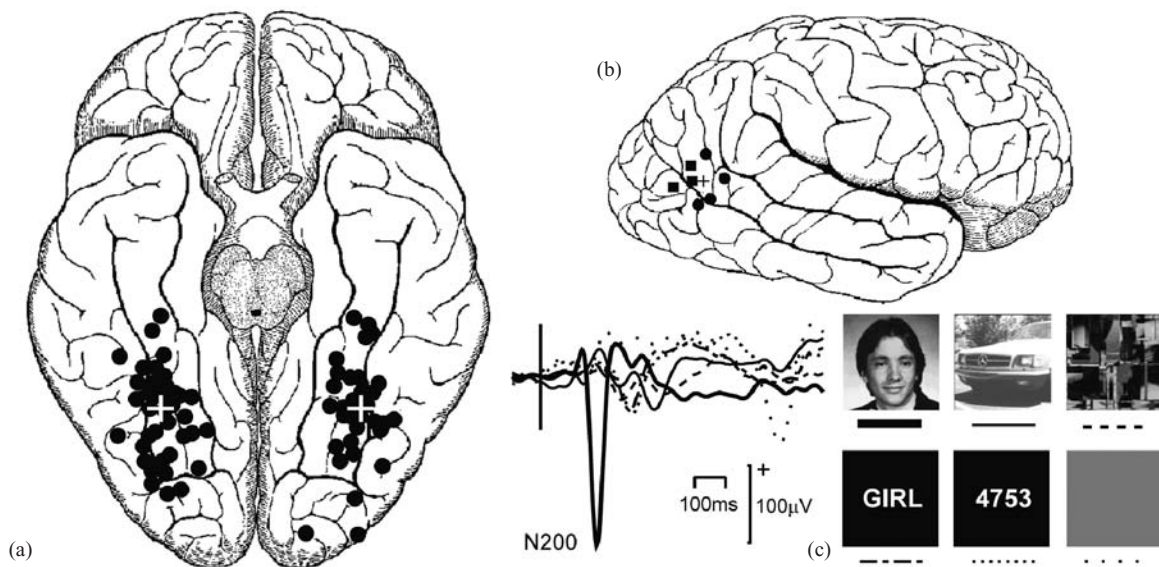


Figure 2

ffi

ffi

ffi

ffi

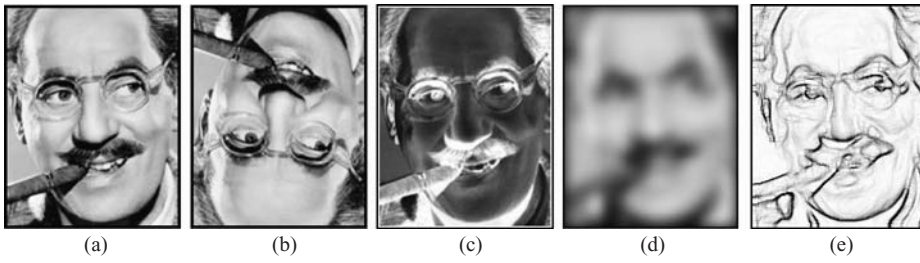


Figure 3

ff

ff

ff

ff
ff

ff

ff

- Farah M J, Rabiowitz C, Quinn G E, Liu G T 2000 Early commitment of neural substrates for face recognition. *Cognitive Neuropsychology* **17**: 117–23
- Gauthier I, Logothetis N K 2000 Is face recognition not so unique after all? *Cognitive Neuropsychology* **17**: 125–142
- Haxby J V, Hoffman E A, Gobbini M I 2000 The distributed human neural system for face perception. *Trends in Cognitive Science* **4**: 223–33
- Humphreys G W, Donnelly N, Riddoch M J 1993 Expression is computed separately from facial identity, and it is computed separately for moving and static faces: neuropsychological evidence. *Neuropsychologia* **31**: 173–181
- Meadows J C 1974 The anatomical basis of prosopagnosia. *Journal of Neurology, Neurosurgery and Psychiatry* **37**: 489–501
- Milders M V, Perrett D I 1993 Recent developments in the neuropsychology and physiology of face processing. *Baillieres Clinical Neurology* **2**: 361–88
- Morton J, Johnson M 1991 CONSPEC and CONLEARN: A two-process theory of infant face recognition. *Psychological Reviews* **98**: 164–81
- Puce A, Allison T, McCarthy G 1999 Electrophysiological studies of human face perception, III. Effects of top-down processing on face-specific potentials. *Cerebral Cortex* **9**: 445–58
- Sergent J, Ohta X, McDonald X 1992 Functional neuroanatomy of face and object processing. *Brain* **115**: 15–36
- Yin R K 1969 Looking at upside-down faces. *Journal of Experimental Psychology* **81**: 141–45

A. Puce

Facial Expressions

1. Introduction

This article describes one aspect of human communication and behavior—facial expressions. It describes the characteristics of a facial expression and its biological wiring, and then offers a selective history of research on facial expressions that emphasizes the nature vs. nurture debate over the origins of facial expressions of emotions. The article concludes with some current research issues and future directions of facial expression research.

2. Defining Facial Expressions

Of all the forms of human communication, which includes the written and spoken word, body language, and so forth, facial expressions are recognized as among the most salient and influential. Researchers reserve the term ‘facial expression’ for those recurring configurations of facial muscle movements that communicate some thought, emotion, or behavior. This is because not all recurring facial muscle configurations express specific messages. For example, some facial muscle actions that accompany spoken words—such

as raising one’s eyebrows when emphasizing a particular word—may modify those words, but are not messages in and of themselves (e.g., Ekman 1991).

The face can express various thoughts. For example, a person who raises the outer corner of one eyebrow may convey sophisticated skepticism. A person whose eyebrows are pulled up in the middle may convey sympathy for another. A wink can convey that one is kidding. Flashing both eyebrows upward may convey a greeting. Or, lowered eyebrows may convey uncertainty (Eibl-Eibesfeldt 1989). Researchers agree for the most part that these types of facial expressions are learned like language, displayed under conscious control, and their meanings are culturally specific that rely on context for proper interpretation (e.g., Bird-whistell 1970). Thus, the same lowered eyebrow expression that would convey ‘uncertainty’ in North America might convey ‘no’ in Borneo (Darwin 1872/1998).

The face can also express emotions. For example, humans express the emotion of happiness by raising lip corners into what is commonly called a smile. Humans can express sadness by frowning. Besides happiness and sadness, other emotions that seem to have specific facial expressions include anger, disgust, fear, and surprise, and to a lesser extent contempt, embarrassment, interest, pain, and shame (e.g., Ekman 1993, Izard 1991). What makes the facial expression of these aforementioned emotions different from other facial expressions is that there is evidence that these emotions are expressed and interpreted the same across all cultures (e.g., Ekman 1993, Izard 1971). This ‘universal’ production and perception across cultures suggests that those emotions and their specific facial expressions might be determined genetically, rather than socially learned. However, this claim is not without controversy (e.g., Russell 1994).

3. Neuroanatomy of Facial Expression

The idea that facial expressions can be both determined genetically, as in the case of some of the emotions, and learned socially, as in the case of all other facial expressions, is supported by an examination the neuroanatomy of the face. There appears to be two distinct neural pathways that mediate facial expressions, each originating in a different area of the brain; one area for the voluntary, willful facial actions (the cortical motor strip), and the second area for the more involuntary, emotional facial actions (subcortical areas; reviewed by Rinn 1984). This dual origin hypothesis is supported by clinical observations of patients who are paralyzed on one side of their face. When these patients were asked to pose a smile, they could only smile on half their face. Yet when these same patients felt the spontaneous emotion of enjoyment after being told a funny joke, they were able to smile on both sides of their face. Likewise, patients with lesions of the subcortical areas of the brain such

as the basal ganglia have difficulty showing spontaneous, emotional facial expressions; however, these patients are able to move their facial muscles on command. These facial action observations are so reliable that they serve as diagnostic criteria for brain lesions.

4. History of Facial Expression Research

The turbulent history of the systematic study of facial expressions began with the publication of Darwin's book *The Expression of the Emotions in Man and Animals* (1872/1998). In this book, Darwin proposed that humans across all cultures have particular and distinct facial expressions for particular emotions, and that these expressions are produced involuntarily as a result of that emotion. Darwin defined emotions as behavioral and physiological reactions that have helped humans and animals to survive the various life challenges they faced throughout their evolutionary history. For example, the fear reaction assisted humans and animals escape danger, the anger reaction assisted humans and animals to fight rivals, and so forth. Those who possessed these emotional reactions were more likely to live to reproductive age and therefore pass their genes to the next generation (see *Emotions, Evolution of*).

What Darwin argued (and elaborated by others, e.g., Ekman 1991, Izard 1991, Plutchik 1991) is that social animals, such as humans, must communicate these emotions to others in the group because emotions express imminent behavior, such as striking out in anger, fleeing in fear, and so on (see *Emotions, Psychological Structure of*). The facial expression of anger thus becomes the visual signal of this intention to strike. This signal allows others in the group to avoid this person, and thus avoid a potential fight (although others argue that these expressions would deprive an individual of a competitive advantage like the element of surprise prior to an attack, e.g., Fridlund 1994). These facial expressions of emotion were seen as vestiges of an entirely nonverbal human communication system that must have existed in extinct human forms such as Neanderthal, because only modern humans have the throat structures necessary to produce articulate speech. A further clue to this prehistoric human communication comes from the genetically closest living relatives of humans, the chimpanzees, who have a repertoire of facial expressions of emotion that parallel, but are not identical to, human facial expressions of emotion. Thus, current human facial expression of emotion represents more the communication methods of the past genetic history of the species, rather than its present conditions (Brown 1991).

However, much of the empirical work that followed Darwin's book failed to support his notion that there were particular facial expressions for particular emo-

tions. For example, when subjects were startled by firecrackers, embarrassed, or disgusted by having to decapitate a live rat, their most common facial expression shown across all these situations was a smile, even though these subjects were not experiencing positive emotion. Likewise, observations of people outside North America by social scientists cast further doubt; for example, the smile was observed as an expression of uncertainty in Africans but as an expression of sadness in Japanese women (see *Culture and Emotion*). Findings such as these—although there were scattered findings to the contrary—caused scholars to conclude that facial expressions did not provide accurate information as to emotional state (Birdwhistell 1970). Thus, by the early 1960s, social science seemed to conclude that all facial expressions—including facial expressions of emotion—were culturally relative, socially learned, and that there were no universals.

Despite this conclusion, two theorists revived Darwin's ideas about the evolutionary origins of facial expressions of emotion (Plutchik 1991, Tomkins 1962/1963). These researchers took photographs of people posing prototypical emotions such as anger, disgust, fear, happiness, and so on, and found that observers would agree as to which expression represented which emotion. Other researchers found similar results with various European, South American, African, and Asian cultures (e.g., Izard 1971). Proponents of the social learning/cultural relativism perspective counter-argued that the populations upon which this evidence for universality was based were mostly educated, and thus could have learned from various forms of media which expressions represented which emotions (e.g., Birdwhistell 1970). To parry this argument, researchers conducted similar studies with visually isolated peoples whom had limited contact with Westerners, and thus could not have learned these expressions from the media (e.g., the Sadong of Borneo, and Fore in New Guinea). These researchers found for the most part the same pattern of universal expression and recognition of facial expressions of emotion as in the Westernized peoples (e.g., Ekman 1993). Follow up research using a variety of methodological alterations to this basic paradigm found patterns consistent universality throughout the 1970s and 1980s (e.g., Izard 1991). Finally, parallel evidence in favor of universality came from observations of children who were born blind and deaf, and who could not have seen these facial expressions to learn how to express them. These children showed similar expressions of emotion as their sighted counterparts (Eibl-Eibesfeldt 1989).

However, proponents of Darwin's idea were still stuck with the findings that peoples of different cultures sometimes showed different expressions for a given emotion than North Americans. Ekman proposed that the reason this happened was that different cultures learned different rules to regulate their facial

expression of emotion—what he called ‘display rules’ (Ekman 1993). For example, Japanese culture has a display rule that prohibits expression of anger or disgust to higher status people, unlike North American culture. Researchers found that both groups showed facial expressions of disgust when viewing a gruesome film alone. But when in the presence of a high status person, the Japanese group hid their disgust feelings with a smile, whereas the Americans still showed disgust expressions. This concept of display rules seemed to account for why people smiled to such seemingly different events as the death of a loved one, confusion, uncertainty, startle, sexual excitement, disgust, and so on (Ekman 1991). Based on these findings, Ekman (1993) proposed his neurocultural theory of emotions. This theory argued that certain basic human emotions generated particular patterns of physiology and facial expressions, that these facial expressions were universal across all cultures, but that their ultimate expression was modified, exacerbated, suppressed, or masked by social learning processes dependent upon personal, family, or cultural display rules (see *Adulthood: Emotional Development*).

5. Current Facial Expression Research

By the early 1990s, a consensus seemingly emerged in the field of psychology that Darwin was correct after all—that some facial expressions of emotion were universal. This was not a peaceful consensus; social scientists who placed the uniqueness of culture at the forefront of any understanding of emotion were not convinced of universality, based on the observations described earlier (e.g., Russell 1994). Experimental psychology itself issued two challenges to universality in the early 1990s. One challenge suggested that all facial expressions were simply communicative gestures, that is, they are not the result of internal emotional states, but only the result of the social motives of the person within a particular context (the ‘behavioral ecology’ view; Fridlund 1994). The behavioral ecology view found that facial expression, particularly smiling, was related not to felt emotion, but to the presence of others. Proponents of universality counter-argued that not all smiles are the same. They demonstrated that only one type of smile, called the enjoyment smile, is related to the positive emotional experience of enjoyment, as measured through self-report or pattern of brain activity. This enjoyment smile looks different from other smiles in that only enjoyment smiles feature *orbicularis oculi* action (the muscles surrounding the eye that give a ‘crow’s feet’ appearance) along with zygomatic major action (the typical lip corner raising). Failing to note the distinction between enjoyment and other smiles may have been why other researchers found no relationship between smiling and positive emotion (Ekman and Rosenberg 1998).

A second challenge to universality attacked the concepts behind what was meant by universality, as well as the methods used to document universality. These methodological problems—such as biased response forms and preselected facial expressions—when added together may have conspired to bias observers’ judgments, causing them to artificially agree on which facial expression represented which emotion (Russell 1994). Prompted by these criticisms, ensuing experimental research corrected many of the proposed methodological shortcomings, and has so far reconfirmed support for the universality of facial expressions of emotion (Ekman 1994).

Although the issue of the biological vs. social origins of facial expressions of emotion is not fully resolved, what has been impressive is the amount of current research generated by the findings on universality of facial expressions of emotion (Ekman 1993). First, researchers have shown that people who pose and hold these universal facial expressions of emotion begin to experience the particular emotion they are posing—although researchers have debated the exact role of a facial expression in reflecting vs. causing an emotion (e.g., Buck 1988). Regardless, this phenomenon has enabled researchers to document physiologically specific patterns of arousal for specific emotions, and in more than one culture. Second, these universal facial expressions of emotion have been employed in studies of brain activity, leading researchers to discover that there are centers in the human brain that respond specifically to these expressions (e.g., the amygdala responds to fear expressions; Whalen 1998). Third, this work on the universal facial expressions has prompted researchers to examine their origins in children (e.g., Izard and Malatesta 1987). This work has shown that children as young as 12 months of age react differently to their mothers’ expressions of fear versus to happiness; a mother’s fear expression will stop a child’s risky behavior, whereas a mother’s happy expression will not (see *Emotions, Children’s Understanding of*).

These universal expressions of emotion have also shown utility as markers of social and psychological functioning. For example, the presence of enjoyment smiles on the part of a person who has survived the death of their romantic partner predicts successful coping with that traumatic loss. Schizophrenic patients tend to show different, and sometimes fewer or more disorganized facial expressions than normal patients (reviewed by Ekman and Rosenberg 1998). Mothers show different sorts of smiles to their difficult compared to their nondifficult children. The facial expression of disgust or contempt, but not anger, predicts marital divorce (Gottman 1994). Researchers found that these unbidden facial expressions of emotion can occur for very brief flashes, called ‘micro-expressions,’ that under certain circumstances can betray deception (Ekman 1991). Thus, current research on facial expressions has moved away from

documenting the existence of these emotional facial expressions and has moved toward examining the implications of the presence or absence of these facial expressions and their corresponding emotions on human social development, interaction, relationships, and psychopathology.

6. Future Directions of Facial Expression Research

Advances in technology will aid facial expression research by allowing researchers to quickly, validly, and reliably observe facial expressions. This will be helpful to the field because current work on facial expression is extremely time and labor intensive, or suffers from other experimental concerns. For example, visible scoring systems, that require close examination of videotape, can take 60 minutes to analyze one minute of behavior (e.g., Ekman and Rosenberg 1998, Izard and Malatesta 1987). Electromyographic techniques, that use electrodes on the face to measure the faint electrical impulses produced by muscle contractions, suffer from concerns about the salience of electrodes on a person's face affecting the behavior of that person (e.g., Fridlund 1994). In the future, computer based analysis programs will be developed to assess the specific muscle movements associated with facial expressions at a much faster, and more reliable way, than these older methods, without causing awareness on the part of the person being analyzed. This will have the effect of making research on facial expression more accessible to more researchers, which can only help the field progress more quickly than in the past.

There are many questions 'facing' future facial expression research, and space limitations permit only a description of a few. First, researchers will try to clarify the stimuli and processes by which social information elicits an emotion and its expression (the 'appraisal' process; Scherer et al. 2000), as well as the process by which people learn to control their emotional facial expressions. This inquiry might provide researchers with a gateway to understanding better the role of expression in the experience and management of emotions. It will also lead to understanding how perceptions of facial expressions may account for differences in social competence and functioning, or 'emotional intelligence' of adults and children. An offshoot of this work would explore whether professionals and lay people can be trained to improve their accuracy at interpreting emotional expressions, and the implications this has for their relationships. Second, researchers will move toward investigating more interactional research designs that place the facial expression of emotion back into the social context which it is typically embedded, to measure the consequences of such expression in the real world. Third, given that much of the previous work has been with

posed expressions, future work would employ more spontaneous facial expressions. Fourth, with the assistance of technology that allows noninvasive observation of the working brain (e.g., Positron Emission Tomography or functional Magnetic Resonance Imaging), researchers will continue to use facial expressions of emotion to map where in the brain the expressions are perceived as well as where they are generated (e.g., Whalen 1998). Fifth, an examination of these first four future directions will inevitably lead to a better understanding of individual differences in production, control, and recognition of facial expressions, of which there is little work at present. Finally, this work would need to be expanded to cultures other than Europe or North America to assess the relative universality of the process of emotion, its antecedents, attempts to control, and its effect on facial expression (e.g., Ekman 1993).

Research on facial expressions has both paralleled and driven changes in the general assumptions in the field of psychology. The finding that people of all cultures seemed to agree on which facial expressions represented which emotions pushed psychology toward re-examining the biological bases for behavior. But research on facial expressions will continue to be controversial because it exposes the strong feelings of those who believe in the power of social situations to mold all human behavior, expressive or not, and those who believe in the biological origins of some of those behaviors. Thus, debates over facial expressions are really debates about human nature—a debate that has tormented social science from time immemorial. What research on facial expression has done is to help move this debate away from an argument over political beliefs about human nature, and toward an argument over observable data.

See also: Adulthood: Emotional Development; Culture and Emotion; Emotion and Expression; Emotion: History of the Concept; Emotion in Cognition; Emotion, Neural Basis of; Emotional Inhibition and Health; Emotions and Health; Emotions, Children's Understanding of; Emotions, Evolution of; Emotions, History of; Emotions, Psychological Structure of; Emotions, Sociology of; Face Recognition Models; Face Recognition: Psychological and Neural Aspects; Infancy and Childhood: Emotional Development; Psychological Therapies: Emotional Processing

Bibliography

- Birdwhistell R L 1970 *Kinesics and Context*. University of Pennsylvania Press, Philadelphia, PA
- Brown D E 1991 *Human Universals*. McGraw-Hill, New York
- Buck R 1988 *Human Motivation and Emotion*. Wiley, New York
- Darwin C 1872/1998 *The Expression of the Emotions in Man and Animals*. Oxford University Press, New York
- Eibl-Eibesfeldt I 1989 *Human Ethology*. de Gruyter, New York

- Ekman P 1991 *Telling Lies*. Norton, New York
- Ekman P 1993 Facial expression and emotion. *American Psychologist* **48**: 384–92
- Ekman P 1994 Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin* **115**: 268–87
- Ekman P, Rosenberg E L (eds.) 1998 *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, New York
- Fridlund A J 1994 *Human Facial Expression: An Evolutionary View*. Academic Press, San Diego
- Gottman J 1994 *Why Marriages Succeed or Fail*. Simon & Schuster, New York
- Izard C E 1971 *The Face of Emotion*. Appleton-Century Crofts, New York
- Izard C E 1991 *Human Emotions*. Plenum Press, New York
- Izard C E, Malatesta C Z 1987 Perspectives on emotional development. I. Differential emotions theory of early emotional development. In: Osofsky J D (ed.) *Handbook of Infant Development*, 2nd edn. Wiley, New York, pp. 494–554
- Plutchik R 1991 *The Emotions: Facts, Theories, and a New Model*. University Press, New York
- Rinn W E 1984 The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions. *Psychological Bulletin* **95**: 52–77
- Russell J A 1994 Is there universal recognition of emotion from facial expression? A review of cross-cultural studies. *Psychological Bulletin* **115**: 102–41
- Scherer K R, Schorr A, Johnstone T 2000 (eds.) *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford University Press, New York
- Tomkins S S 1962/1963 *Affect, Imagery, Consciousness (Vol. 1, The Positive Affects, Vol. 2, The Negative Affects)*. Springer, New York
- Whalen P J 1998 Fear, vigilance, and ambiguity: Initial neuroimaging studies of the human amygdala. *Current Directions in Psychological Science* **7**: 177–88

M. G. Frank

Faction: Political

In its broadest construction, a political faction is any part of a political whole. The term has been applied to phenomena ranging from sets of people whose policy preferences tend to align to membership groups that undertake collective action.

In a most famous formulation Madison (Hamilton et al. 1961) defined a faction as a partisan political division of any size, although he qualified the definition to refer to such groups whose aims were 'adverse to the rights of other citizens, or to the permanent and aggregate interests of the community.' Madison viewed factions as the inevitable consequence of political liberty and social diversity, and urged the 'extended sphere' of a federal republic to contain 'the violence of faction.' Madison's treatment of the 'mischief of faction' is often given as evidence for

the American founders' distaste for such commonplace features of representative democracy as political parties and lobbying groups.

For current analytical purposes, Madison's definition of faction is overly broad, in encompassing political parties and interest groups, and overly restrictive in parsing motives. A good, if still broad, definition is that a political faction is a subsidiary part of a political institution, of a political party, a lobbying organization, or a legislature. Unlike political parties within a polity or within a legislature, factions commonly have no legal standing in the institution of which they are part, although they might play an important and enduring role in organizing it. Subgroups that have legal status within institutions are usually designated in other ways, as are the caucuses in the United States Congress and the international political parties in the International Typographical Union. Sometimes analysts identify as factions groupings that are evanescent and happenstantial, for instance a voting coalition on a particular legislative bill or party nomination. Most, however, limit the term to groupings that are enduring and widely recognized. The factional affiliations of Liberal Democrats standing for election to the Japanese Diet, for instance, were routinely reported in the newspapers. The factions within the Democratic Party in West Virginia, while not reported in the press, nevertheless persisted through changes in political administration and party leadership.

In any complex society, a tendency toward faction within political parties and interest groups would seem to be inevitable. Members of parties and groups might be united on the central purposes of the organization but still divided on secondary questions. For decades, the Democratic Party in the United States was loosely united on issues of economic policy but deeply divided into northern and southern factions on issues of race and federal powers, to the point where party allegiances were often superseded by a conservative coalition, an informal alliance of Republicans and southern Democrats, in the US Congress. A common purpose in clericalism still left the Christian Democrats, Italy's confessional party, vulnerable to factions arising from region and class.

Although they are more difficult to observe, similar factional divisions occur in many lobbying groups. At several points in its history, the American Farm Bureau Federation harbored two or three identifiable factions that reflected regional commodity interests. In the 1920s, the Anti-saloon League foundered on a factional division over the question of whether the League ought to emphasize the enforcement of Prohibition in the United States or the promotion of the cause of prohibition abroad.

Considerable evidence exists, however, that some political parties and some interest groups are more prone to faction than others. Political parties in Japan, Italy, and France in the Fourth Republic were

famously factionalized, while parties in Norway, Sweden, and the German Federal Republic were not.

Several conditions appear to promote a tendency toward factional strife in political parties. One factor is electoral domination: factionalization appears more likely in parties that face no real electoral challenge. The Italian Christian Democrats (DC) and the Japanese Liberal Democrats (LDP) are striking both for their enduring factional divisions and their long runs of electoral success—the DC participated in every Italian government from World War II into the 1990s, and the LDP controlled the Japanese Diet continuously from 1955 to 1993. In the American southern states, the electoral threat posed by the mountain Republicans produced a Democratic Party in North Carolina and Tennessee that was appreciably more cohesive than the Democratic Party in states like South Carolina, Florida, Mississippi, and Arkansas, where the Republican Party constituted barely even a nuisance. While counterexamples of dominant parties not rent by faction abound—the Social Democrats in Sweden, the Republicans in Vermont—a considerable body of theory, beginning with the work of Riker (1962) in fact predicts that parties that enjoy more than minimal majorities will be prone to factional disintegration (see *Minimum Winning Coalition, in Politics*).

A second influence on the propensity toward factionalization is the structure of political cleavage within the society (Lipset and Rokkan 1967). In Belgium, class divisions cut across parties that are organized primarily on the basis of language and religion. In Italy, economic and regional differences fueled factions within a confessional party whose *raison d'être* is the establishment of the Roman Catholic Church. In Sweden, by contrast, religious homogeneity and a dominant secularism caused religious disputes to recede well behind the class divisions that are the basis for the political parties (see *Cleavages: Political*).

A third influence on faction is the electoral system. The electoral laws place an upper bound on the number of political parties that can be sustained, a limit equal to one more than the number of candidates to be elected from the constituency (Cox 1997). Accordingly, the electoral laws govern the ease with which factional divisions can progress to outright schism and the formation of new political parties. In the Low Countries of Belgium and the Netherlands, highly proportional representation rules coupled with low election thresholds have produced party systems in which parties and factions morph each into the other with remarkable fluidity. Japan's system of multi-member districts with a single nontransferable vote fostered a half-dozen factions that slated candidates and oversaw the distribution of spoils. The hard limit of a system of single-member districts and plurality voting in the United States forced northern and southern Democrats to stick together, despite

their factional animosities. In the United States, the advent of nominating primaries—creating, effectively, single-member constituencies with plurality voting *within* parties—often limited the number of enduring party factions to two, as in North Dakota on the Republican side or in Louisiana on the Democratic side. In North Dakota, in fact, the more liberal wing of the Republican Party, the Nonpartisan League, split off and joined the Democrats in the 1950s, when two-party competition finally became viable.

The ease or difficulty of schism also contributes to the propensity of interest groups to be ridden with factions. As Hirschman (1970) put it, when exit is costly, members resort to voice. Interest groups that rely primarily on the expressive value of the group's purposes to motivate members seem especially prone to faction (Wilson 1974). Doctrinal purity matters when members are motivated primarily by doctrine. The National Organization for Women, for example, suffered through disputes between a faction that wished to pursue conventional lobbying for the Equal Rights Amendment and a faction that wished to promote social movement activities like direct action. Interest groups that rely upon expressive benefits, moreover, tend also to be prone to schism, because 'purposive' inducements to group involvement are so easy to provide. In the later part of the twentieth century, the 'public interest groups' in the United States multiplied like Protestant denominations, rapidly and schismatically. A dissident faction of the Sierra Club, for example, bolted the organization to form Friends of the Earth; a few years later, a dissident faction of Friends of the Earth created the Environmental Policy Center. In contrast, interest groups that attract members with relatively expensive material benefits like insurance, or interest groups that are able to secure membership through coercion, make it costly for dissidents to exit, channeling what might have been schism into faction. Factional fights within labor unions typically continue intramurally rather than extramurally; the costs a new organization would have to bear to win the right to represent workers in collective bargaining preclude exit. The American Medical Association (AMA) has endured despite tensions between specialists and general practitioners, between doctors in large practices and in small practices, between academics and clinicians, bound by the range of AMA services and by the state powers that have been delegated to the medical societies to regulate medical employment.

The potential for faction in interest groups has long been seen as an important limit on the power and influence of lobbying groups in American politics. Pluralist political scientists identified 'overlapping memberships,' that is, conflicting internal interests, as an important limit on the demands that interest groups might make. During the energy policy debates of the 1970s, for instance, the National Petroleum Refiners Association took no position on crude oil price

regulation, immobilized by conflicts between the small independent refiners and the 'majors.' Likewise, the potential for internal disagreement sidelined many of the large business associations, like the US Chamber of Commerce and the National Association of Manufacturers, in the debates over the extension of the Reciprocal Trade Agreements Act in the 1950s. 'Quasiunanimity,' Bauer et al. (1972) concluded, was a 'premise of action' in lobbying groups. A group that was too extreme in its demands risked disabling internal controversy.

In political parties, factions emerge and recombine to produce change in party systems, as has occurred recently in both Japan and Italy. But factional divisions also hobble parties in the achievement of their policy goals. The most dramatic example is surely the decades-long obstruction of civil rights legislation, a Democratic Party priority since the Truman administration, by the Party's southern Dixiecrat faction. But if factionalism is joined with limited party competition, many argue, the consequences are the more severe. When parties are isolated from the judgment of the electorate, factions dispute not policy direction but the division of the spoils. The extreme factionalization of the Liberal Democratic Party in Japan and the Christian Democratic Party in Italy contributed to the inefficiency, particularism, and corruption of those two dominant parties. But at least the Japanese and Italian factions were relatively enduring and organized. As Key (1950) argued of the Democratic Party in the American southern states, where commonly it was 'every man for himself' in elections, competition between ephemeral factions within a hegemonic party is no substitute for competition between two or more political parties. Political parties develop a stake in their policy reputations that factions commonly do not. Popular control in a democracy requires the clear electoral choices that political parties provide and party factions cannot.

See also: Cleavages: Political; Electoral Systems; Factionalism; Interest Groups; Interest Groups, History of; Interest: History of the Concept; Lobbying; Minimum Winning Coalition, in Politics; Party Systems; Political Parties

Bibliography

- Bauer R A, Pool I deS, Dexter L A 1972 *American Business and Public Policy*. Aldine-Atherton, New York
- Cox G W 1997 *Making Votes Count*. Cambridge University Press, New York
- Hamilton A, Madison J, Jay J 1961 In: Rossiter C (ed.) *The Federalist Papers*. New American Library, New York, no. 10
- Hirschman A O 1970 *Exit, Voice, and Loyalty*. Harvard University Press, Cambridge, MA

- Key V O Jr 1950 *Southern Politics*. Knopf, New York
- Laver M, Schofield N 1990 *Multiparty Government*. Oxford University Press, New York
- Lipset S M, Rokkan S 1967 Cleavage structures, party systems, and voter alignments. In: Lipset S M, Rokkan S (eds.) *Party Systems and Voter Alignments*. Free Press, New York
- Mayhew D R 1986 *Placing Parties in American Politics*. Princeton University Press, Princeton, NJ
- Ramseyer J M, Rosenbluth F McC 1993 *Japan's Political Marketplace*. Harvard University Press, Cambridge, MA
- Riker W 1962 *The Theory of Political Coalitions*. Yale University Press, New Haven, CT
- Wilson J Q 1974 *Political Organizations*. Basic Books, New York

J. M. Hansen

Factionalism

Factionalism refers to dissension between rival subgroups—factions—within a larger social unit. Factionalism can take many forms and has been observed in all parts of the world. It is a basic political process dynamically related to social change.

1. Factions and Organizations

Factionalism, regardless of where it is classified on the conflict continuum, is conflict between factions. Factions are coalitions of persons or subgroups that compete over specific issues within a larger organization or community. The central focus of a faction is the leader who coordinates its activities and recruits its members. Ties between leader and followers are usually personal, although some followers may recruit others on behalf of the leader. The issues about which factions compete are diverse. They generally concern scarce resources the control over which provide power chances, such as economic assets, office and new laws. But they may also involve honor, ideology and behavioral norms. Loosely structured, factions are non-corporate groups that generally dissolve when the particular issues that gave rise to them are resolved. But if the issues remain unresolved, factions may acquire a range of cultural trappings such as property, symbols, ideology and bureaucratic organization. They can then evolve into permanent corporate groups, such as ritual moieties, political parties or other formal associations, for which the term faction is no longer appropriate.

Factionalism takes place within the framework of an established social entity, whether village, school, political party, office, club, kin group, etc., that have clear norms of behaviour. These norms generally include notions of unity, consensus, and cooperation.

The covert maneuvering of factions bent on achieving their own, often self-serving aims, contradict these norms. Hence factionalism has a pejorative connotation. It is seen as subverting organizational rules and goals. Faction members are consequently viewed as disloyal, obstructive persons whose pursuit of narrow, short-term advantage endangers the wider, long-term goals of the organization. The divisiveness inherent in factionalism also hampers the day to day affairs of the organization or community that depend on cooperation. Furthermore, factionalism jeopardizes the good name and image of unity and harmony it seeks to project to the outside world.

2. *The Study of Factionalism*

The study of factions and factionalism developed slowly, shadowing theoretical shifts in the social sciences. Although Linton (1936) long ago suggested that factions presented an interesting but unexplored field, little was done until the 1950s. Political anthropology until then was dominated by the functionalist paradigm elaborated in *African Political Systems* (Fortes and Evans-Pritchard 1940). This viewed politics as maintaining order through consensus, harmony, and balanced opposition. The political groups on which functionalists focused were enduring units, corporate groups. Conflict, if examined, was viewed as reinforcing the social structure. Loosely structured, temporary coalitions such as factions patently did fit into this conception of politics. The theoretical hegemony of Africanist political anthropologists began to be challenged in the 1950s. Not surprisingly, Raymond Firth, given his interest in individual choice, had for some time been uncomfortable with the functionalist paradigm. He and colleagues, who had observed factionalism in Indian communities, were the first to examine factionalism theoretically (Firth 1957). They treated factions as informal counterparts of more formal political formations whose members were recruited according to structurally diverse principles. They also noted that factions tended to become activated on specific occasions and not as regularly recurring features. Other studies of factions and factionalism swiftly followed (Siegel and Beals 1960, Boissevain 1964, 1974, Nicholas 1965, Bailey 1969, Thoden van Velzen 1973, Bujra 1973, Alavi 1973). Most employed a transactionalist perspective, viewing political activity as an arena in which entrepreneurs transact personal relations for political and economic gain. The study of factionalism culminated in the late 1970s with the volume edited by Silverman and Salisbury (1977). This is still considered the definitive work on the subject. Following its publication and the demise of functionalist social science—in which transactionalists played a significant role—academic debate moved on to puzzles related to symbolic, cognitive, and discursive approaches to politics. Factions and factionalism have become accepted concepts

whose characteristics are no longer debated. They are also proving useful to related disciplines (see Brumfiel and Fox 1994).

3. *Issues*

Much of the work on factionalism in the 1960s, although using a transactionalist perspective, continued to be strongly influenced by functionalism. Factionalism was seen as occurring as a result of rapid sociocultural change. Change was seen as coming from the surrounding environment, not as the result of tensions inherent in the community or society in which factionalism arose. Factionalism was generally viewed as occurring because the system's equilibrium was disturbed, and it operated to restore its dynamic equilibrium. Factions were regarded as structurally similar and in balanced opposition. These viewpoints were debated and successfully challenged during the 1970s.

Many of the societies in which factionalism was observed were in fact not subject to rapid social change. But if factionalism was not always a 'result' of change, it seemed always to be 'about' change: changes in power resources, ideology, rules and/or ways of doing things. Factions competed about who was to be boss, about which normative concepts were to be followed, about whose will was to prevail, and thus about which rules were to be followed. Moreover, a closer examination of rival factions revealed that far from being similar or evenly matched, they differed in their access to resources, strategy, tactics, internal organization, ideology, and social composition.

4. *Factions: Structure, Symmetry, and Balance*

Factions often form in opposition to or in defense of some issue or some pre-existing source of power and authority within a community or organization. The distribution of such resources is binary. Some persons have more and some have less. Those with more normally constitute the local establishment that coalesces around a leader or a dominant personage such as a headman, mayor, parish priest, or club president. They and their supporters form the 'establishment' faction. Those who are dissatisfied with the establishment's exercise of its power constitute a category from which a rival or 'opposition' faction can be recruited.

Reports also mention conservative and progressive factions. Because the local establishment normally defends the 'status quo,' from which it derives its superordinate position, it is often labeled as conservative. Since the opposition faction challenges the established defenders of the 'status quo' it is labelled progressive. These labels often do not reflect reality, as when a progressive opposition faction defeats its rival and becomes the dominant, establishment faction.

There is a further reason why opposition factions come to be regarded as progressive. Since they are weaker, they are perhaps more receptive to new resources if and when these become available in the wider society. With these they can challenge their rivals. In rapidly changing societies these may include new government and commercial offices, new laws and new ideologies. Such new resources rapidly tend to change the balance of power. Factionalism thus does not necessarily result from the availability of new resources, as some authors have suggested. Rather, new resources are used in ongoing competitions for power and prestige and tend to escalate the conflict. The use of new resources is not random. A faction will use new resources when it seems likely that they will strengthen its position. It is then labelled progressive, in the sense of favoring change of the 'status quo.'

There is evidence that conflict groups—whether faction, ritual moiety, or political party—differ organizationally. When the opposition consists merely of a category of persons disgruntled with the dominant power elite, they are obviously less organized than the establishment. The internal structure of the local establishment faction, whose members are used to networking to maintain their position, will normally have a more developed exchange circuit than the opposition faction. But if a conflict between the two persists over time, the opposition faction may well become better organized than its rival may. Good organization is a valuable resource and one of the ways a weaker faction can successfully challenge its rival. It is thus more open than the establishment faction to organizational innovation. Because of its superior resources, the dominant faction also tends to be more wasteful. It does not need to husband its resources to the extent that its weaker rival does.

Like most coalitions, factions have core and peripheral members. Core members cluster around the leader and have multiple links to each other. The peripheral members are often linked only to the leader or to a single member of the core group. Where there is a strong core, the faction often acquires some of the characteristics of corporate groups noted previously.

Factions are not necessarily ideologically neutral, as some authors have suggested. The differences between establishment and opposition, like those between conservative and progressive, are not random. They have ideological implications. There is evidence that opposition factions recruit more support than their rivals from weaker or even marginal social categories do. Since opposition leaders usually lack the network ties and material resources that establishment leaders use to recruit followers, they cannot afford to be too particular about the nature of their support. The strength of a faction is usually a function of its size. Just as they often turn to new ideologies and tactics—which because they are new are often viewed as socially unacceptable and subversive—opposition leaders also recruit supporters from among those who are less

influential or are regarded as social or morally inferior. Followers are followers.

Opposition factions consequently often develop or adopt an overarching ideology or symbol to bind their heterogeneous members into a unity. They also often align themselves with political parties that defend the interests of their socially weaker supporters. Most often, opposition factions seek links to parties that embody an emancipatory ideology. Their establishment rivals develop relations with political parties representing vested interests in the wider society. Political parties, on the other hand, also consciously make use of local factions to recruit support at the grass roots level.

5. Conclusion

If factionalism is not necessarily a product of social change, it appears to be always about change. Factions are coalitions that compete for power to determine, and thus to change, what is to be accepted as normal. Rival factions, because they have different access to power chances, are not evenly matched, structurally similar groups. Their asymmetry is fundamental to understanding the nature of factionalism and its dynamic for long term change.

Structural asymmetry and competition for power are also characteristic of class-based conflict groups. This suggests that the line of cleavage between faction, class or party cuts across moral categories and socio-economic classes, not at right angles, as most functionalist and class-oriented analysts postulate, but diagonally. Where the line approaches the vertical, forming conflicting coalitions with a generally even spread across socio-economic classes, it is reasonable to speak of factionalism, in the case of face-to-face groups, and party conflict in the case of conflict on a broader scale. Where the line of cleavage approaches the horizontal, forming conflict groups that are more clearly differentiated according to socio-economic criteria, the term class conflict seems appropriate. But in every case the axis of cleavage must be determined by empirical investigation. It should not be taken for granted.

See also: Charisma: Social Aspects of; Conflict and Conflict Resolution, Social Psychology of; Conflict: Anthropological Aspects; Conflict/Consensus; Conflict: Organizational; Conflict Sociology; Faction: Political; Groups, Sociology of; Issue Networks: Iron Triangles, Subgovernments, and Policy Communities; Leadership, Psychology of; Solidarity, Sociology of

Bibliography

Alavi H 1973 Peasant classes and primordial loyalties. *Journal of Peasant Studies*. 1: 232–62

- Bailey F G 1957 *Caste and the Economic Frontier*. Manchester University Press, Manchester, UK
- Bailey F G 1969 *Stratagems and Spoils*. Basil Blackwell, Oxford, UK
- Boissevain J 1964 Factions, parties and politics in a Maltese village. *American Anthropologist*. **66**: 1275–87
- Boissevain J 1974 *Friends of Friends: Networks, Manipulators and Coalitions*. Basil Blackwell, Oxford, UK
- Brumfiel E M, Fox J W (eds.) 1994 *Factional Competition and Political Development in the New World*. Cambridge University Press, Cambridge, UK
- Bujra J 1973 The dynamics of political action: A new look at factionalism. *American Anthropologist*. **75**: 132–52
- Firth R 1957 Introduction to factions in Indian and overseas Indian societies. *British Journal of Sociology*. **8**: 291–5
- Fortes M, Evans-Pritchard E E (eds.) 1940 *African Political Systems*. Oxford University Press, London
- Linton R 1936 *The Study of Man*. D. Appleton-Century, New York
- Nicholas R W 1965 Factions: A comparative analysis. In: Banton M (ed.) *Political Systems and the Distribution of Power*. Tavistock, London
- Siegel B, Beals A R 1960 Pervasive factionalism. *American Anthropologist*. **62**: 395–417
- Silverman M, Salisbury R F (eds.) 1977 *A House Divided? Anthropological Studies of Factionalism*. Memorial University, Newfoundland, Canada
- Thoden van Velzen H U E 1973 Coalitions and network analysis. In: Boissevain J, Mitchell J C (eds.) *Network Analysis: Studies in Human Interaction*. Mouton, The Hague, The Netherlands

J. Boissevain

Factor Analysis and Latent Structure, Confirmatory

Confirmatory factor analysis (CFA) is a quantitative data analysis method that belongs to the family of structural equation modeling (SEM) techniques. CFA allows for the assessment of fit between observed data and an *a priori* conceptualized, theoretically grounded model that specifies the hypothesized causal relations between latent factors and their observed indicator variables. Because population-level equivalence between data and model cannot be proven with sample data, CFA should be viewed as a mainly disconfirmatory technique. That is, CFA facilitates the statistical rejection—or, at best, a very tentative retention—of a specific theory regarding the factor(s) responsible for the observed relations in the data. If, on the other hand, the investigator's intentions are a mostly ungrounded exploration of relations suggested by the data, classical exploratory factor analysis is the more appropriate approach. In this entry, typical steps in a CFA are introduced theoretic-

ally and via example: from model specification and identification, to parameter estimation, data-model fit assessment, and potential model modification. Applied and methodological references are provided for a more in-depth study of CFA and SEM techniques in the social and behavioral sciences.

1. Overview

The term 'factor analysis' describes a host of methods, all of which have the purpose of facilitating a better understanding of the latent, unobserved variables (factors) that underlie a set of directly measurable and observed variables. These factors are often believed to represent constructs, psychological or otherwise, that have a direct bearing on the measured variables; as such they are assumed to motivate (and in turn be inferable from) the pattern of correlations or covariances among those observed variables. In the late 1960s, works by Karl Jöreskog (e.g., 1966, 1967) articulated a method for confirmatory factor analysis (CFA), an application of normal theory maximum likelihood estimation to factor models with specific theoretical latent structures. Such structures could include the *a priori* specification of the number of factors, their orthogonality or obliquity, and which variables had zero and nonzero relations with those factors. This distinguishes CFA from well-known exploratory factor analysis (e.g., Gorsuch 1983, Mulaik 1972) wherein the number and nature of the factors emerge from the observed variables' data through a mathematical algorithm, largely blind to any substantive theory. Most crucial in Jöreskog's CFA work was the provision for a formal statistical χ^2 -test of the fit between the pattern of relations among the measured variables and the theorized factor model, thereby facilitating the disconfirmation or tentative confirmation of an hypothesized factor model. Soon after, Jöreskog and others put forth a more general framework for the integration of measured and latent variables into complex causal networks, serving as the foundation for what is often known as structural equation modeling (SEM). CFA, which may be considered a special case within the more general SEM framework, is the focus of the current entry.

2. CFA Model Specification, Identification, and Parameter Estimation

Suppose an educational researcher wishes to investigate the possibility of a low positive relation between reading and mathematics ability for fifth grade students, which is measured by standardized tests such as the Stanford Achievement Test or the Iowa Test of

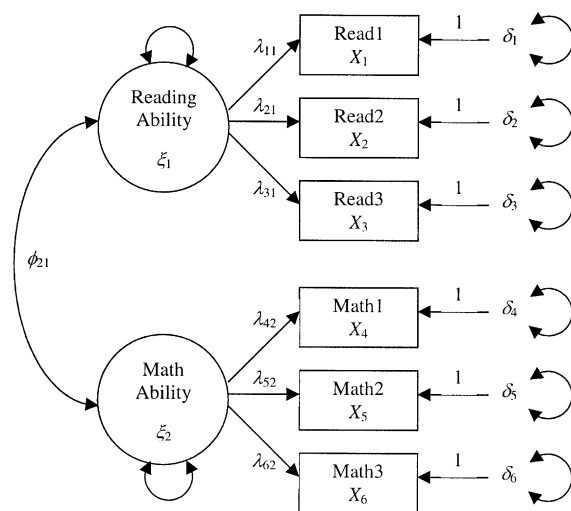


Figure 1
Hypothetical CFA model of reading and mathematics ability

Basic Skills. The model shown in Fig. 1 might be hypothesized. Measured variables X_1 through X_6 , shown in rectangles, are believed to be caused by the latent factors ξ_1 and ξ_2 , shown in circles. Here, ξ_1 and ξ_2 represent true latent (unobserved) reading and mathematics ability, respectively, with X_1 through X_3 being standardized reading test measures (Read1 through Read3) and X_4 through X_6 being standardized mathematics test measures (Math1 through Math3). Table 1 includes a simulated variance–covariance matrix for the observed variables X_1 through X_6 based on test data from $n = 1,200$ fifth graders. Our example's focus is the noncausal covariance between reading and mathematics ability, ϕ_{21} . In general, a

covariance is indicated by a two-headed arrow connecting the two constructs, and—because a variance is a covariance of a variable (observed or latent) with itself—it, too, is depicted by a two-headed arrow from the variable to itself.

A factor's hypothesized causal impact on its measured indicator variables is symbolized by an arrow from the factor to the variable with magnitude λ_{ij} , where i denotes the observed variable and j denotes the latent factor. Note that such a model explicitly posits the factors as causing the variables, rather than the variables causing the factors; the latter type of model, in which the factor is characterized as emergent rather than latent, is much less common and beyond the scope of this entry. In many cases, there is no arrow from a factor to a variable, such as from ξ_1 to X_4 ; this implies that Reading Ability has no theoretical causal bearing on the Math1 variable. Finally, to the extent that the factors do not perfectly explain each variable, a residual term, δ , is included as an influential contributor (with its variance shown by a two-headed arrow from δ to itself). This residual might consist of variable-specific measurement error as well as other influences. Thus, each observed variable is the sum of two parts, that attributable to the common factor(s) and that residual part specific to the variable.

The causal relations of the hypothesized model shown in Fig. 1 may be expressed as a system of six regression-like structural equations:

$$X_1 = \lambda_{11}\xi_1 + 0\xi_2 + \delta_1 \quad (1)$$

$$X_2 = \lambda_{21}\xi_1 + 0\xi_2 + \delta_2 \quad (2)$$

$$X_3 = \lambda_{31}\xi_1 + 0\xi_2 + \delta_3 \quad (3)$$

$$X_4 = 0\xi_1 + \lambda_{42}\xi_2 + \delta_4 \quad (4)$$

$$X_5 = 0\xi_1 + \lambda_{52}\xi_2 + \delta_5 \quad (5)$$

$$X_6 = 0\xi_1 + \lambda_{62}\xi_2 + \delta_6 \quad (6)$$

Table 1
Simulated data and selected parameter estimates for the reading and mathematics ability model in Fig. 1

Variance/covariance matrix ($n = 1200$)						
	X_1	X_2	X_3	X_4	X_5	X_6
	129.96					
	79.75	192.65				
	694.20	871.11	12038.48			
	307.03	391.75	3402.04	9876.38		
	230.99	415.74	2476.89	4815.24	12126.41	
	37.85	53.36	416.71	740.98	656.84	135.722
Standardized factor loadings and indicator reliability estimates						
ξ_1	0.70*	0.73*	0.79*	0	0	0
ξ_2	0	0	0	0.75*	0.60*	0.85*
R^2	0.49	0.53	0.62	0.56	0.35	0.73
Data-model fit indices						
$\chi^2 = 16.98$, $df = 8$, $p = 0.030$				CFI = 0.996	SRMR = 0.016	RMSEA = 0.031

Note. * $p < 0.05$

Equivalently, these equations can be represented in matrix form, as in

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \end{bmatrix} \quad (7)$$

That is,

$$\mathbf{X} = \mathbf{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta} \quad (8)$$

where $\mathbf{X}$ is a column vector of observed variables, $\mathbf{\Lambda}$ is a matrix of factor loadings, $\boldsymbol{\xi}$ is a vector of latent constructs, and $\boldsymbol{\delta}$ is a column vector of residuals.

The implication of Fig. 1 and the accompanying structural equations is that the population variance-covariance matrix for the X variables, $\boldsymbol{\Sigma}$, is a function of (1) the λ_{ij} loadings in matrix $\mathbf{\Lambda}$, of (2) the variances and covariance between the latent factors in a matrix $\boldsymbol{\Phi}$, and of (3) the variances and covariances among the residuals in a matrix $\boldsymbol{\Theta}_\delta$ (note that in Fig. 1 all residual covariances are zero as implied by the absence of two-headed arrows between the δ terms). More specifically, if all model parameters (loadings, variances, and covariances) are contained in a single column vector $\boldsymbol{\theta}$, the population variance-covariance matrix of the observed variables that is implied by the model and its parameters, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, is given by

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathbf{\Lambda}\boldsymbol{\Phi}\mathbf{\Lambda}' + \boldsymbol{\Theta}_\delta \quad (9)$$

A vector of parameter estimates $\hat{\boldsymbol{\theta}}$ can be derived so that the model-implied variance-covariance matrix $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ is as similar as possible to the observed variance-covariance matrix $\mathbf{S}$ provided that model identification has first been ensured. To this end, each parameter in a model must be expressible as a function of the variances and covariances of the observed variables. When a system of such relations can be uniquely solved for the unknown parameters, the model is just-identified. When multiple such expressions exist for one or more parameters, the model is over-identified; in this case, a best-fit (although not unique) estimate for each parameter is derived. If, however, at least one parameter cannot be expressed as a function of the observed variables' variances and covariances, the model is under-identified and some or all parameters cannot be estimated on the basis of the data alone. This under-identification might be the result of the researcher attempting to impose a model that is too complex relative to the number of variances and covariances of the observed variables. Additionally, empirical under-identification might arise when unfortunate estimates for select parameters (e.g., values of zero for factor covariances) render subsets of

model parameters inestimable. Fortunately, in most CFA applications, it suffices to ensure that (1) the number of parameters to be estimated, p , does not exceed the number of variances and covariances of the observed variables, c , and (2) each latent factor has an assigned unit of measurement. To accomplish the latter condition for the model in Fig. 1, we set the factor variances to unity (alternatively, for each of the two factors we could have specified one of the factor loadings to equal unity, thereby setting each factor's units equal to that observed variable). The model in Fig. 1 is over-identified with $c = 6(7)/2 = 21$ non-redundant observed (co)variances and $p = 13$ parameters to be estimated: one covariance between the two latent factors, six variances of the error terms associated with the observed variables, and six factor loadings.

Given that a model is just- or (preferably) over-identified, sample estimates can be obtained through a variety of estimation methods. These include maximum likelihood and generalized least squares, both of which assume multivariate normality and are asymptotically equivalent, as well as asymptotically distribution-free estimation methods that generally require a substantially larger sample size. These methods iteratively minimize a function of the discrepancy between $\mathbf{S}$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, where $\mathbf{S}$ is the unrestricted variance-covariance matrix of the observed X variables and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the model-implied variance-covariance matrix reproduced from the iteratively changing parameter estimates. The standardized maximum likelihood estimates of key parameters in the reading and mathematics ability model are presented in Table 1. Before focusing on our example's main parameter estimate (ϕ_{21}), however, we should consider whether or not there is any evidence suggesting data-model misfit, any statistical—and theoretically justifiable—rationale for modifying the hypothesized model, or any indication of factor unreliability.

3. Data-Model Fit Assessment and Model Modification

One of the advantages of CFA is the ability it offers to assess the quality of the fit of the data to the model. A multitude of measures exists that assist the researcher in deciding whether to reject or tentatively retain an *a priori* specified over-identified model (see Marsh et al. 1988, Tanaka 1993). In general, measures to assess the fit between the variances and covariances observed in the data and those implied by the model can be classified into three categories: absolute, parsimonious, and incremental. Absolute fit indices are those that improve as the discrepancy between $\mathbf{S}$ and $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ decreases, that is, as the number of parameters, p , approaches the number of nonredundant observed variances and covariances, c . Examples of such measures include the model χ^2 statistic that tests the

stringent null hypothesis $H_0: \Sigma = \Sigma(\theta)$, the standardized root mean-square residual (SRMR) that roughly assesses the average standardized discrepancy between observed and model-implied variances and covariances, and the goodness-of-fit index (Jöreskog and Sörbom 1996) designed to evaluate the amount of observed variance-covariance information that can be accounted for by the model.

Parsimonious fit indices take into account not just the overall absolute fit, but also the degree of model complexity required to achieve that fit. Indices such as the adjusted goodness-of-fit index (Jöreskog and Sörbom 1996), the parsimonious goodness-of-fit index (Mulaik et al. 1989), and the root mean-square error of approximation (RMSEA, Steiger and Lind 1980) indicate greatest data-model fit when data have reasonable absolute fit and models are relatively simple. Finally, incremental fit indices such as the normed fit index (Bentler and Bonett 1980) and the comparative fit index (CFI, Bentler 1990) gauge the data-model fit of an hypothesized model relative to that of a more restrictive baseline model with fewer parameters.

The three types of fit indices together help the researcher to converge upon a decision regarding the CFA model's acceptability. In our example, the significant $\chi^2 = 16.98$ ($df = c - p = 8$, $p < 0.05$) indicates that the observed variance-covariance matrix would occur rarely if our model correctly depicted the true population relations. This absolute fit statistic, however, is notoriously sensitive to very small and theoretically trivial model mis-specifications (e.g., slight amounts of error covariance) under large sample conditions. As such, other fit indices are generally preferred for model evaluation. According to Hu and Bentler (1999), CFI values of 0.96 or greater together with SRMR values less than 0.09 (or with RMSEA values less than 0.06) point to acceptable data-model fit. The indices in Table 1 suggest no appreciable data-model inconsistency given the Hu and Bentler cut-off criteria.

After the data-model fit has been assessed, a decision about that model's worth must be reached. Acceptable fit indices usually lead to the conclusion that no present evidence exists warranting a rejection of the model or the theory underlying it. This is not to say that the model and theory have been confirmed, much less proven as correct; rather, the current factor model remains as one of possibly many that satisfactorily explain the relations among the observed variables. On the other hand, when fit indices indicate a potential data-model misfit, one might be reluctant to dismiss the model entirely. Instead, attempts are often made to modify the model *post hoc* so that acceptable fit indices can be obtained. Such modifications could include the addition of cross-loadings, allowing a given variable to load on multiple factors, or of error covariances in which variables' residuals are allowed to covary in order to reflect some

potential variable relation above and beyond that motivated by the factor structure itself.

Most CFA software packages will facilitate such model 'improvement' by providing modification indices (Lagrange multiplier tests) indicating what changes in the model could reap the greatest increase in absolute fit, that is, decrease in the model χ^2 statistic. While such indices constitute a potentially useful tool for remedying incorrectly specified models, it seems imperative to warn against an atheoretical hunt for the model with the best fit. Many alternative models exist that can explain the observed data equally well; hence, attempted modifications must be based on a sound understanding of the specific theory underlying the model. Furthermore, when modifications and re-analyses of the data are based solely on data-model misfit information, subsequent fit results might be due largely to chance rather than true improvements to the model. Modified structures therefore should be cross-validated with an independent sample whenever possible. If a new sample is not available but the initial sample is large enough, one can randomly split the sample into calibration and validation subsamples and compute Cudeck and Browne's (1983) cross-validation index. When the initial sample is too small, Browne and Cudeck (1989) also offered a single sample alternative, an estimate of the expected value of the cross-validation index. From the analysis of the model in Fig. 1, none of the modification indices suggested changes to the model that would result in a significant improvement in data-model fit (i.e., a significant decrease in the model χ^2 statistic); thus, we did not report them in Table 1.

Before drawing conclusions regarding the relation between reading and mathematics ability, the question of the 'quality' (i.e., reliability) of the factors should be addressed. Traditionally, this has been accomplished (a) by focusing on the reliability of scores from individual indicator variables, R^2 (i.e., the proportion of variability in an observed variable that can be accounted for by the underlying factor) or (b) by assessing the reliability of scores from linear composites of the indicator variables (e.g., Miller 1995, Raykov 1997). More recently, Hancock and Mueller (2001) suggested a measure (coefficient H) computed from the standardized factor loadings that can be used to assess the reliability of a latent construct itself as reflected by scores from multiple observed variables. One of coefficient H 's advantages over traditional construct reliability measures is that it is never less than the best indicator variable's reliability (R^2), thereby drawing information from all indicators in a manner commensurate with their own ability to reflect the construct. For the current example, $H = 0.79$ and $H = 0.82$ for the reading and mathematics ability constructs, respectively, while the R^2 values for the respective factor's indicator variables in Table 1 range between 0.49–0.62 and 0.35–0.73. Thus, the two factors exhibit reasonable and satisfactory levels of

construct reliability since about 80 percent of their variance is explainable by their respective indicators. Table 1 also lists the standardized factor loadings, that is, the estimates of the λ_{ij} parameters in the Λ matrix that indicate the strength and direction of the *a priori* specified causal influences of the latent factors on the observed variables. All factor loadings are positive and significantly different from zero.

Finally, our primary research question can be addressed. The CFA estimate of the correlation between the two latent constructs reading and mathematics ability is 0.51 and significantly different from zero. This estimate indicates a low to moderate positive association between the two constructs of interest, as hypothesized.

4. Conclusion

Confirmatory factor analysis has become established as an important analysis tool for many areas of the social and behavioral sciences. It belongs to the family of structural equation modeling techniques that allow for the investigation of causal relations among latent and observed variables in *a priori* specified, theory-derived models. The main advantage of CFA lies in its ability to aid researchers in bridging the often-observed gap between theory and observation. For example, an instrument might be developed by creating multiple items for each of several specific theoretical constructs (Fig. 1). Instead of analyzing data with an exploratory factor analysis (where each item is free to load on each factor) and potentially facing a solution inconsistent with initial theory, a CFA can give the investigator valuable information regarding the fit of the data to the specific, theory-derived measurement model (where items load only on the factors they were designed to measure), and point to the potential weakness of specific items. CFA is best understood as a process, from model conceptualization, identification and parameter estimation, to data-model fit assessment and potential model modification. As opposed to exploratory methods, CFA's strength lies in its disconfirmatory nature: models or theories can be rejected, but results might also point toward potential modifications to be investigated in subsequent analyses.

Numerous CFA examples exist in the various disciplines covered in this encyclopedia. The reader might consult (a) in education: O'Grady (1989), Stevens (1995); (b) in psychology: Asmundson et al. (2000), Byrne (1993); (c) in sociology: Alsup and Gillespie (1997), Mulvey et al. (1992); and (d) in health: Galassi et al. (1992), Windle and Dumenci (1999). Some of the more widely used and state-of-the-art SEM software packages for conducting CFA are LISREL (Jöreskog & Sörbom 1996), EQS (Bentler 1997), and AMOS (Arbuckle 1999). Popular textbooks on CFA and more general SEM include Bollen (1989), Hayduk (1987), Loehlin (1998), and Mueller

(1996). Social science journals that publish many CFA and SEM applications and methodological developments include: *Educational and Psychological Measurement*, *Multivariate Behavioral Research*, *Journal of Experimental Education*, *Psychological Methods*, *Sociological Methodology*, *Sociological Methods and Research*, and *Structural Equation Modeling: A Multidisciplinary Journal*.

See also: Factor Analysis and Latent Structure: Overview; Latent Structure and Casual Variables

Bibliography

- Alsup R, Gillespie D F 1997 Stability of attitudes toward abortion and sex roles: A two-factor measurement model at two points in time. *Structural Equation Modeling: A Multidisciplinary Journal* **4**: 338–52
- Arbuckle J L 1999 *AMOS 4.0*. SmallWaters Corporation, Chicago
- Asmundson G J G, Frombach I, McQuaid J, Pedrelli P, Lenox R, Stein M B 2000 Dimensionality of posttraumatic stress symptoms: A confirmatory factor analysis of DSM-IV symptom clusters and other symptom models. *Behaviour Research and Therapy* **38**: 203–14
- Bentler P M 1990 Comparative fit indexes in structural models. *Psychological Bulletin* **107**: 238–46
- Bentler P M 1997 *EQS 5.7*. Multivariate Software, Inc., Encino, CA
- Bentler P M, Bonett D G 1980 Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin* **88**: 588–606
- Bollen K A 1989 *Structural equations with latent variables*. Wiley, New York
- Browne M W, Cudeck R 1989 Single sample cross-validation indexes for covariance structures. *Multivariate Behavioral Research* **24**: 445–55
- Byrne B M 1993 The Maslach Burnout Inventory: Testing for factorial validity and invariance across elementary, intermediate, and secondary teachers. *Journal of Organizational and Occupational Psychology* **66**: 197–212
- Cudeck R, Browne M W 1983 Cross-validation of covariance structures. *Multivariate Behavioral Research* **18**: 147–67
- Galassi J P, Schanberg R, Ware W B 1992 The Patient Reactions Assessment: A brief measure of the perceived quality of the patient-provider medical relationship. *International Journal of Psychology* **27**: 639
- Gorsuch R L 1983 *Factor Analysis*, 2nd edn. Lawrence Erlbaum Associates, Hillsdale, NJ
- Hancock G R, Mueller R O 2001 Rethinking construct reliability within latent variable systems. In: Cudeck R, du Toit S, Sörbom D (eds.) *Structural Equation Modeling: Present and Future—A Festschrift in Honor of Karl Jöreskog*. Scientific Software International, Lincolnwood, IL, pp. 195–216
- Hayduk L A 1987 *Structural Equation Modeling with LISREL: Essentials and Advances*. Johns Hopkins University Press, Baltimore, MD
- Hu L, Bentler P M 1999 Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal* **6**(1): 1–55
- Jöreskog K G 1966 Testing a simple structure hypothesis in factor analysis. *Psychometrika* **31**: 165–178

- Jöreskog K G 1967 Some contributions to maximum likelihood factor analysis. *Psychometrika* **32**: 443–82
- Jöreskog K G, Sörbom D 1996 *LISREL 8*. Scientific Software International, Chicago
- Loehlin J C 1998 *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis*, 3rd edn. Lawrence Erlbaum Associates, Mahwah, NJ
- Marsh H W, Balla J R, McDonald R P 1988 Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin* **103**: 391–410
- Miller M B 1995 Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal* **2**: 255–73
- Mueller R O 1996 *Basic Principles of Structural Equation Modeling: An Introduction to LISREL and EQS*. Springer-Verlag, New York
- Mulaik S A 1972 *The Foundations of Factor Analysis*. McGraw-Hill, New York
- Mulaik S A, James L R, Van Alstine J, Bennett N, Lind S, Stilwell C D 1989 Evaluation of goodness-of-fit indexes for structural equation models. *Psychological Bulletin* **105**: 430–45
- Mulvey P W, Miceli M P, Near J P 1992 The Pay Satisfaction Questionnaire: A confirmatory factor analysis. *Journal of Social Psychology* **132**: 139–42
- O'Grady K E 1989 Factor structure of the WISC-R. *Multivariate Behavioral Research* **24**: 177–93
- Raykov T 1997 Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement* **21**: 173–84
- Steiger J H, Lind J M 1980 *Statistically Based Tests for The Number of Common Factors*. Paper presented at the annual meeting of the Psychometric Society. Iowa City, IA
- Stevens J J 1995 Confirmatory factor analysis of the Iowa Test of Basic Skills. *Structural Equation Modeling: A Multidisciplinary Journal* **2**: 214–31
- Tanaka J S 1993 Multifaceted conceptions of fit in structural equation models. In: Bollen K A, Long J S (eds.) *Testing Structural Equation Models*. Sage, Newbury Park, CA, pp. 10–39
- Windle M, Dumenci L 1999 The factorial structure and construct validity of the Psychopathy Checklist-Revised (PCL-R) among alcoholic inpatients. *Structural Equation Modeling: A Multidisciplinary Journal* **6**: 372–93

R. O. Mueller and G. R. Hancock

Factor Analysis and Latent Structure: IRT and Rasch Models

In education and the social sciences, we often ask subjects to respond to a set of items (questions, statements, or tasks) on survey forms, self-report inventories, and mental tests, that are coded as discrete—often dichotomous—variables. In many settings it is natural to think, in analogy with factor analysis (see *Factor Analysis and Latent Structure: Overview; Factor Analysis and Latent Structure, Confirmatory*), that there is one or more continuous latent variables for each subject—such as political agency, extrovertedness, or ability in some area of mathe-

matics—that can be measured or estimated using the positive and negative responses to these items. Item Response Theory (IRT), and specializations such as the Rasch Model, refer to a set of statistical modeling and estimation tools for making inferences about one or more continuous latent variables from this kind of multivariate discrete data.

1. Introduction: the Rasch Model

As an example, consider a test of transitive reasoning in school children (e.g., Sijtsma and Verweij 1999). An item may consist of three rods, A, B, and C, of differing lengths. The child is given evidence that $\text{Length}(A) < \text{Length}(B)$, and $\text{Length}(B) < \text{Length}(C)$, and is asked to deduce a relationship between $\text{Length}(A)$ and $\text{Length}(C)$; his or her answer is scored 1 if it is correct and 0 if it is incorrect. Let us say that J such items, involving different types of objects and attributes (balls of varying weight, disks of varying area, etc.) and different numbers of objects, make up a test given to N school children. If the scored responses were continuous we might build a two-way additive ANOVA model in which child i 's response to item j , X_{ij} , is the sum (or difference) of the child's achieved transitive reasoning level, and the item's innate difficulty, as determined by the attribute on which objects are compared and the number of objects being compared.

However, since the responses are dichotomous $X_{ij} = 0$ or 1, we need a model specifying the probability that $X_{ij} = 1$, in terms of the child's transitive reasoning ability and the item's difficulty. One of the earliest such models, developed by the Danish mathematician Georg Rasch (Rasch 1960), specifies

$$P_j(\theta_i) \equiv P[X_{ij} = 1 | \theta_i, \beta_j] = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)},$$

$$i = 1, \dots, N, j = 1, \dots, J \quad (1)$$

where θ_i represents child i 's transitive reasoning level, and β_j represents the difficulty of the item: as the latent variable θ_i increases, the probability of getting the item right increases, and as the difficulty parameter β_j increases, the probability of getting the item right decreases. The Rasch model, together with similar models using different forms for $P_j(\theta_i)$ developed by Jane Loevinger (Loevinger 1948) and Frederic M Lord (Lord 1952), became the core of what is now item response theory.

On the logit scale, the Rasch model has exactly the two-way additive ANOVA structure we sought:

$$\text{logit } P_j(\theta_i) = \ln \frac{P_j(\theta_i)}{1 - P_j(\theta_i)} = \theta_i - \beta_j.$$

We further assume local independence among item responses, i.e., all X_{ij} are conditionally independent,

given θ s and β s. Letting x_{ij} denote the observed values of X_{ij} , we may write the likelihood for the Rasch model as

$$P[X_{ij} = x_{ij}, i = 1, \dots, N, j = 1, \dots, J | \theta_1, \dots, \theta_N; \beta_1, \dots, \beta_J] = \prod_{i=1}^N \prod_{j=1}^J P_j(\theta_i)^{x_{ij}} [1 - P_j(\theta_i)]^{1-x_{ij}}. \quad (2)$$

An up-to date account of the Rasch model and its various extensions is given by Fischer and Molenaar (1995), to whom we also refer for historical notes and references.

The Rasch model in Eqn. 1 has the interesting invariance property that the odds ratio comparing two items j and k is

$$\frac{P_j(\theta_i)}{1 - P_j(\theta_i)} \cdot \frac{1 - P_k(\theta_i)}{P_k(\theta_i)} = \exp \beta_k - \beta_j$$

independent of θ_i . Hence in our example, the ‘difficulty’ of two items involving different transitive reasoning settings can be compared, in principle, using any convenient sample of subjects regardless of their transitive reasoning level (in practice, subjects still contribute at least ‘small sample bias’ to the estimated β s). Similarly, the odds ratio comparing two subjects is independent of the item difficulty parameters. These two invariance properties are instances of ‘specific objectivity,’ a property Rasch considered important in defining a measurement model and from which Eqn. 1 can be derived (see Fischer and Molenaar 1995). The Rasch model is also closely related to the Bradley–Terry model of paired comparisons. Consider two items j and k for which $\beta_j < \beta_k$. For any subject i who got only one item right, it follows from Eqn. 1 that the probability that the easier item was correct is

$$P[X_{ij} = 1 | X_{ij} + X_{ik} = 1, \theta_i, \beta_j, \beta_k] = \frac{\exp(\beta_k - \beta_j)}{1 + \exp(\beta_k - \beta_j)} \quad (3)$$

independent of θ_i ; this is exactly the Bradley–Terry model with parameters $\phi_j = -\beta_j$. The lack of dependence on θ_i on the right can also be used to construct empirical tests of the Rasch model, since the probability in Eqn. 3 can be estimated without assuming the Rasch model and should be the same in any subpopulation of subjects in which the Rasch model applies.

2. Estimation in the Rasch Model

Two estimation methods are traditionally associated with the Rasch model, joint maximum likelihood (JML) and conditional maximum likelihood (CML).

One can maximize Eqn. 2 jointly in the θ s and β s, by numerically solving the ‘normal equations’ obtained by setting all partial derivatives of the logarithm of Eqn. 2 equal to zero; the maximizing values $\hat{\theta}_i$ and $\hat{\beta}_j$ are called joint maximum likelihood estimators (JMLEs). However, the JMLEs turn out to be *inconsistent*, or asymptotically biased: suppose we are primarily interested in estimating β_j s; it need not be the case that each $\hat{\beta}_j$ converges to β_j as we increase the subject sample size N , keeping the number of items J fixed. Indeed, as N increases, more θ_i s are also added to the model: intuitively, some information must be expended on estimating them, and there is not enough information left over to improve the estimates of the β_j s. One solution to this problem is to exploit the fact that in the Rasch model, a sufficient statistic for each θ_i is subject i ’s total number correct score, $X_{i+} = \sum_{j=1}^J X_{ij}$. If the likelihood in Eqn. 2 is divided by the joint likelihood for the observed x_{i+} s, the resulting conditional likelihood contains only the β_j parameters, and the conditional maximum likelihood estimators (CMLEs) $\hat{\beta}_j$ obtained by maximizing this new likelihood are consistent (asymptotically unbiased) as N grows and J stays fixed. For this reason, CML estimates are usually preferred over JML estimates. Additional details are provided by Andersen (1980), whose earlier work demonstrated the inconsistency problem with JML, and established consistency of CML for the Rasch model.

Another approach to solving the inconsistency problem with JML is to assume the θ_i are independent random effects following a common (discrete or continuous) distribution $f(\theta | \lambda)$. Integrating over θ_i for each subject yields the marginal likelihood

$$\prod_{i=1}^N \int \prod_{j=1}^J P_j(\theta_i)^{x_{ij}} [1 - P_j(\theta_i)]^{1-x_{ij}} f(\theta_i | \lambda) d\theta_i \quad (4)$$

and maximizing this with respect to β_j s and λ (using an E-M algorithm, for example) yields what are called maximum marginal likelihood (MML) estimates $\hat{\beta}_j$ and $\hat{\lambda}$. These MML estimates are also consistent (asymptotically unbiased). The method based on Eqn. 4 can be interpreted as an empirical Bayes method; it also links the Rasch model directly with other latent variable approaches such as factor analysis, where the latent variable θ_i is treated as an unobserved random variable, or equivalently as completely missing data (see *Factor Analysis and Latent Structure: Overview*).

Finally, it can be shown that under Eqn. 4, the marginal distribution of the data follows a log-linear model for the 2^J table of the form

$$\ln p(x_{i1}, \dots, x_{iJ}) = \alpha - \sum_{j=1}^J \beta_j x_{ij} + \sum_{k=0}^J \gamma_k \chi(x_{i+} = k)$$

where $\chi(S) = 1$ if statement S is true, and 0 otherwise. This is the log-linear model of quasisymmetry; the connection between the Rasch model and the quasi-

symmetry model has been independently discovered innumerable times in the literature. Lindsay et al. (1991) give a reasonably complete account of the connection with quasisymmetry and consequences for the identifiability of the distribution $f(\theta|\lambda)$ in a semiparametric formulation of Eqn. 4.

3. Parametric IRT

Other parametric IRT models have been developed for situations in which the Rasch model does not fit, while retaining the ideas of dichotomous ($X_{ij} \in \{0, 1\}$) or at least polytomous ($X_{ij} \in \{0, 1, \dots, K-1\}$) item scores, a unidimensional (or low-dimensional) latent variable θ_i characterizing subjects, a low-dimensional parameter β_j characterizing items, and local (conditional) independence of all item responses X_{ij} given all parameters. An accessible review of developments in parametric IRT in the 1960s, 1970s and 1980s is given by Hambleton et al. (1991).

A simple generalization of the Rasch model that is often useful for dichotomously-scored items,

$$P_j(\theta_i) = \frac{\exp[\alpha_j(\theta_i - \beta_j)]}{1 + \exp[\alpha_j(\theta_i - \beta_j)]} \quad (5)$$

is called the two-parameter logistic (2PL) model. In the 2PL model, the slope parameters α_{jk} are called 'discrimination parameters,' and they control the Fisher information for estimating the θ s. Again, the location parameters β_j play the role of 'difficulty parameters'; they also determine the value of θ maximizing the Fisher information. The three-parameter logistic (3PL) model adds a nonzero lower asymptote to Eqn. (3), to model random or exchangeable response behavior by low- θ subjects. Further generalizations, discussed in the next several paragraphs, are treated at length by Fischer and Molenaar (1995) and van der Linden and Hambleton (1997).

Additional covariates and other structures can be incorporated into these models as regression terms. For example, in the linear logistic test model (LLTM), the difficulty parameters β_j in the Rasch model are rewritten as linear combinations of K basic parameters η_k with weights q_{jk} , and

$$\text{logit } P_j(\theta_i) = \theta_i - \sum_{k=1}^K q_{jk} \eta_k.$$

The matrix $Q = [q_{jk}]$ is usually obtained as a regression model matrix from *a priori* analysis of the items; for example, in the transitive reasoning example described in Sect. 1, the q_{jk} might indicate what kind and how many objects were used for each item. By contrast, multidimensional compensatory IRT models decompose the unidimensional θ_i parameter into an item-

dependent linear combination of underlying traits, e.g.,

$$\text{logit } P_j(\theta_i) = \sum_{a=1}^D \alpha_{ja} \theta_{ia} - \beta_j.$$

For example, mathematics word problems may involve both a math component, θ_{i1} , and a verbal component θ_{i2} in different proportions in different problems, as determined by the discrimination parameters α_{ja} . In some settings, the α_{ja} may be fixed *a priori*, like the q_{jk} in the LLTM model, and in other settings they may be estimated analogously to factor loadings in Factor Analysis. Multiplicative or conjunctive IRT models combine unidimensional models for components of response in a product $P_j(\theta_i) = \prod_a^D P_{ja}(\theta_{ia})$ where $P_{ja}(\theta_{ia})$ are parametric unidimensional dichotomous response functions. The usual interpretation is that the $P_{ja}(\theta_{ia})$ are probabilities of correctly performing component skills or subtasks, which must be done in conjunction in order to generate a correct response to the item itself; this model is sometimes referred to as the multicomponent latent trait model (MLTM).

Probit versions of all the above models, in which the logit function is replaced with $\Phi^{-1}(p)$, the inverse standard normal distribution function, are also in use. Aside from a trivial rescaling of the parameters, there is little difference in practice between the probit and logit models; the choice depends mostly on computational expedience.

IRT model can also be extended to handle polytomous or multcategory items, $X_{ij} \in \{0, 1, 2, \dots, K-1\}$, in much the same way that logistic regression models are extended polytomous data: in the 'graded response model' (GRM) the cumulative logit, $\text{logit } P[X_{ij} > c | \theta_i, \beta_j]$, is linear in θ_i ; in the 'sequential model' (SM) the continuation-ratio logit, $\text{logit } P[X_{ij} > c | X_{ij} > c-1, \theta_i, \beta_j]$, is linear in θ_i ; and in the 'partial credit model' (PCM) the adjacent-category logit, $\text{logit } P[X_{ij} = c+1 | X_{ij} \in \{c, c+1\}, \theta_i, \beta_j]$, is linear in θ_i . Again, probit reformulations of these models are straightforward. When the discrimination parameters are all non-negative, the cumulative response curves in any of these models is nondecreasing (Hemker et al. 1997), a condition referred to as 'monotonicity.' By contrast, IRT formulations of direct-response probabilistic unfolding models, also known as proximity models or parallelogram models, have unimodal response curves that peak near the location β_j of the item.

As was the case for the Rasch model discussed in Sect. 2, JML estimates for IRT models are relatively easy to set up but generally inconsistent (asymptotically biased); CML methods can be used when simple sufficient statistics are available. MML methods based on applying an E-M algorithm to Eqn. 4, with $P_j(\theta_i)$ redefined according to one of the above

First level: $X_{ij} \sim P(\theta_i; \beta_j)^{X_{ij}} [1 - P(\theta_i; \beta_j)]^{(1-X_{ij})}$,

$$i = 1, \dots, N; j = 1, \dots, J$$

Second level: $\theta_i \sim f_i(\theta|\lambda_f)$, each i

$$\beta_j \sim g_j(\beta|\lambda_g), \text{ each } j$$

Third level: $\lambda_f \sim \phi_f(\lambda_f)$

$$\lambda_g \sim \phi_g(\lambda_g)$$

Figure 1

Two-way hierarchical structure for N individuals and J dichotomous response variables. Factors in the first level are independent and are multiplied together to form a likelihood for the response data matrix $[X_{ij}]$. Factors in the second and third levels are also independent, and covariates and special structure for the latent variables and item parameters may be introduced in these levels

parametric forms, are most common, however. The recent development of Markov Chain Monte Carlo (MCMC) integration methods for Bayesian statistics (see *Markov Chain Monte Carlo Methods*) provides an alternative to these methods. Patz and Junker (1999) sketch a general MCMC methodology for parametric IRT models, and Johnson and Albert (1999) survey a range of related applications from a working Bayesian statistician's perspective.

The IRT toolbox has been pushed forward in recent years by the needs of both large-scale educational assessment surveys (e.g., Johnson et al. 1994) and cognitively based assessment models (e.g., Nichols et al. 1995). To accommodate finer structure in the latent space, as well as to incorporate subject and item covariates in the model, it is useful to reformulate the IRT framework as a two-way hierarchical Bayes structure for N individuals and J response variables, as in Fig. 1. In addition to the item response functions $P_j(\theta_i) = P(\theta_i; \beta_j)$ in Fig. 1, f_i and g_j are prior distributions on the latent variables θ and item parameters β respectively (both of which may be multidimensional), and λ_f and λ_g represent sets of hyperparameters needed to specify these prior distributions. The model in Fig. 1 is expressed for dichotomous items but can be generalized easily to polytomous items or combinations of item types, incomplete data, and so forth (e.g., Patz and Junker 1999). The MML formulation dis-

cussed in the previous paragraph is obtained by taking $f_i(\theta)$ to be a single population distribution not dependent on j for the latent trait, and taking g_j to be flat priors not depending on j for the item parameters.

4. Nonparametric IRT

Nonparametric IRT refers to several related methodologies for working with item response data without fully committing to any particular well-known family of parametric item response theory models. Some topics in nonparametric IRT are considered in van der Linden and Hambleton (1997) and Boomsma et al. (2001), and several current issues are surveyed in a special issue of *Applied Psychological Measurement* (Junker and Sijtsma 2001).

Modern nonparametric scaling methods include the 'essential independence' (EI) approach of Stout and his students and colleagues (e.g., Stout et al. 1996), and the Mokken scaling approach (e.g., Hemker et al. 1995); both methods are related to Cronbach's alpha bound on reliability. Mokken scaling is very good at finding groups of items that are highly discriminating among subjects; Stout's methods tend to identify groups of items that narrowly satisfy LI/EI, M, and U better. Ramsay (e.g., Ramsay 1996) has developed computational tools for estimating item response functions nonparametrically, and for visualizing the dimensionality of θ in terms of the surface in the J -dimensional unit cube generated by the joint likelihood for $X_{i1}, \dots, X_{iJ}$ as θ_i varies.

5. Some Applications

Parametric and nonparametric IRT methods are routinely used to assess the quality of individual items and sets of items in educational measurement work (Linn 1989) and to elucidate experimental design and statistical estimation issues when performances of examinees who sit for different versions of the same exam must be compared; such work goes under the name test equating. A similar problem arises in the scoring of sequentially-designed, computerized adaptive tests (e.g., Sands et al. 1997). IRT modeling has also elucidated research into the assessment of sociological bias (differential item functioning) on standardized test items (Holland and Wainer 1993) and the detection and diagnosis of subject outliers (Meijer 1996). Related work includes the study of rater effects (e.g., Patz et al. 2000) and matrix-sampling and the incorporation of group effects and other item and subject covariates in educational assessment survey work (Johnson et al. 1994).

In the broader social sciences and related areas, IRT has also played a substantial role. The Rasch model has long been applied in social survey work; see, for example, Duncan (1985). The Rasch model and other

IRT models can also be applied to model subject heterogeneity in estimating a closed population from multiple-recapture census data (see *Censuses: History and Methods*); Fienberg et al. (1999) recently surveyed this methodology and compared Bayesian IRT models with standard log-linear models in this context. Applications to multiple outcomes of designed experiments and to panel data are indicated in the volumes by Fischer and Molenaar (1995) and van der Linden and Hambleton (1997). Sijtsma and Verweij (1999) apply nonparametric IRT to scale construction in developmental psychology; and Nichols et al. (1995) collect many applications of IRT and related Bayesian inference network models (see *Bayesian Graphical Models and Networks*; Mislevy 1996) in cognitive diagnosis. Applications of parametric and nonparametric IRT to psychiatric scales include Gibbons et al. (1985), Santor et al. (1995) and Kim and Pilkonis (1999). In biostatistics Legler and Ryan (1997) use IRT to model multiple physical outcomes in the study of birth defects.

6. Summary

Item response theory (IRT) has grown from its roots in postwar mental-testing problems, through intensive use in educational measurements in the 1970s, 1980s, and 1990s, to become a mature statistical toolkit for modeling of multivariate discrete response data using subject-level latent variables. Applications of IRT can be found throughout the social sciences and related areas, from education, psychology, economics, and demography to medical research.

Most parametric IRT models would be recognizable by modern statisticians as mixed-effects multivariate generalized linear models, but IRT has benefited from interaction with all parts of the statistical community. Almost any assessment phenomenon—from between-subject dependence due to institutional or sociological factors, to behavioral aspects of raters, to the analysis of item responses in terms of requisite subject or item features—can be expressed in the hierarchical mixture/Bayes modeling framework of Fig. 1, because of its conceptual simplicity. Recent advances in computation, and MCMC methods in particular, have made it possible to estimate a vastly wider variety of these models than would have been imaginable even in the early 1990s however the more complicated models typically also make higher sample-size demands. Nonparametric IRT approaches also rely on computationally intensive methods, including spline and kernel smoothing and bootstrap techniques, to estimate and test probability inequalities, stochastic ordering properties, and similar features of the models. Questions motivating both parametric and nonparametric IRT modeling inevitably involve identifying the phenomena that are worth detailed modeling, and seeing if the computational and data collection

machinery can be pushed to be informative about reasonable models of these phenomena.

Hambleton et al. (1991) provide a fairly straightforward introduction to IRT in educational testing; Andersen (1980) presents the Rasch model in the context of other statistical models for the social sciences; and Fischer and Molenaar (1995) and van der Linden and Hambleton (1997) provide modern technical accounts. A broad range of current research issues in IRT is collected in Boomsma et al. (2001).

Bibliography

- Andersen E B 1980 *Discrete Statistical Models with Social Science Applications*. North-Holland, New York
- Boomsma A, Snijders T A B, van Duijn M A J (eds.) 2001 *Essays on Item Response Theory*. Springer-Verlag, New York
- Duncan O D 1985 Some models of response uncertainty for panel analysis. *Social Science Research* **14**: 126–41
- Fienberg S E, Johnson M S, Junker B W 1999 Classical multi-level and Bayesian approaches to population size estimation using multiple lists. *Journal of Royal Statistical Society: Series A* **162**: 383–405
- Fischer G H, Molenaar I W (eds.) 1995 *Rasch Models: Foundations, Recent Developments, and Applications*. Springer-Verlag, New York
- Gibbons R D, Clark D C, von Ammon Cavanaugh S, Davis J M 1985 Application of modern psychometric theory in psychiatric research. *Journal of Psychiatric Research* **19**: 43–55
- Hambleton R K, Swaminathan H, Rogers H J (eds.) 1991 *Fundamentals of Item Response Theory*. Sage, Newbury Park, CA
- Hemker B T, Sijtsma K, Molenaar I W 1995 Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement* **19**: 337–52
- Hemker B T, Sijtsma K, Molenaar I W, Junker B W 1997 Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psymka* **62**: 331–47
- Holland P W, Wainer H 1993 *Differential Item Functioning*. Erlbaum, Hillsdale, NJ
- Johnson V E, Albert J 1999 *Ordinal Data Modeling*. Springer-Verlag, New York
- Johnson E G, Mislevy R J, Thomas N 1994 Theoretical background and philosophy of NAEP scaling procedures. In: Johnson E G et al. (eds.) *Technical Report of the NAEP 1992 Trial State Assessment Program in Reading*. OERI, US Dept. of Ed., Washington, DC, Chap. 8, pp. 133–46
- Junker B W, Sijtsma K 2001 Nonparametric IRT in action: An overview of the special issue. *Applied Psychological Measurement* **25**:
- Kim Y, Pilkonis P A 1999 Selecting the most informative items in the IIP scales for personality disorders: An application of item response theory. *Journal of Personality Disorders* **13**: 157–74
- Legler J M, Ryan L M 1997 Latent variable models for teratogenesis using multiple binary outcomes. *Journal of the American Statistical Association* **92**: 13–20
- Linn R L (ed.) 1989 *Educational Measurement*, 3rd edn. Macmillan, New York
- Lindsay B, Clogg C C, Grego J 1991 Semiparametric estimation in the Rasch model and related exponential response models,

- including a simple latent class model for item analysis. *Journal of the American Statistical Association* **86**: 96–107
- Lord F M 1952 A theory of test scores. *Psychometric Society*, New York
- Loevinger J 1948 The technique of homogeneous tests compared with some aspects of 'scale analysis' and factor analysis. *Psychological Bulletin* **45**: 507–30
- Meijer R R 1996 Person-fit research: An introduction. *Applied Measurement in Education* **9**: 3–8
- Mislevy R J 1996 Test theory reconceived. *Journal of Educational Measurement* **33**: 379–416
- Nichols P D, Chipman S F, Brennan R L (eds.) 1995 *Cognitively Diagnostic Assessment*. Erlbaum, Hillsdale, NJ
- Patz R J, Junker B W 1999 Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics* **24**: 342–66
- Patz R J, Junker B W, Johnson M S submitted The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*
- Ramsay J O 1996 A geometrical approach to item response theory. *Behmka* **23**: 3–17
- Rasch G 1960 Probabilistic models for some intelligence and attainment tests. University of Chicago Press, Chicago
- Sands W A, Waters B K, McBride J R 1997 *Computerized Adaptive Testing: From Inquiry to Operation*. American Psychological Association, Washington, DC
- Santor D A, Zuroff D C, Ramsay J O, Cervantes P, Palacios J 1995 Examining scale discriminability in the BDI and CES-D as a function of depressive severity. *Psychiatry Assessment* **7**: 131–9
- Sijtsma K, Verweij A C 1999 Knowledge of solution strategies and IRT modeling of items for transitive reasoning. *Applied Psychological Measurement* **23**: 55–68
- Stout W, Habing B, Douglas J, Kim H R, Roussos L, Zhang J 1996 Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement* **20**: 331–54
- van der Linden W J, Hambleton R K (eds.) 1997 *Handbook of Modern Item Response Theory*. Springer-Verlag, New York

B. W. Junker

Factor Analysis and Latent Structure: Overview

The terms factor analysis and latent structure refer to two aspects of essentially the same problem. Both are concerned with statistical problems in which some of the variables are latent, meaning that they are unobservable. Information about the latent variables has, therefore, to be obtained indirectly through indicators, also known as manifest or observable variables. The terminology of the subject reflects the diverse origins of the subject accumulated over almost a century. 'Factor' in this context is a synonym for latent variable. From a theoretical perspective, the only distinctive thing about a latent variable model is

the presence of unobservable variables. In all other respects the normal methods of statistics apply.

Indeed, a latent variable problem can be regarded as a standard statistical problem in which the data on the latent variables are missing. There is no essential difference, for purposes of analysis, between a problem in which some of the data are never obtained and one in which some are lost.

1. Introduction

Latent variables arise mainly, but not exclusively, in the social sciences. This is because social science often deals in concepts which are constructs rather than the directly measurable variables which are typical of physical sciences. The earliest example, and still one of the most important, is general intelligence or 'g.' This goes back to Spearman (1904) and was constructed to describe the variation among individuals in what appeared to be common to a wide range of mental tests. Psychology and sociology abound in such latent variables. Attitudes as well as abilities are all spoken of in the discourse of these subjects as things which occur in varying amounts and which, therefore, appear in the theory as quantitative variables. In economics, variables like business confidence play a similar role. It could be justly claimed that the aspiration of these subjects to be regarded as sciences depends on the success with which latent variables can be quantified.

Latent variables can be classified into several types. Many, like intelligence, are conceived of as continuous in which case we are looking for a scale on which individuals can be located. In other contexts it is more appropriate to think of the latent variable as categorical. In that case individuals are supposed to belong to one of several classes which may or may not be ordered. What is true of the latent variables is, of course, true for the manifest variables and the only essential difference between the various methods is in the types of variables for which they are appropriate.

A convenient way of displaying the relationship between the methods is to introduce the fourfold classification of Table 1. We classify the manifest and the latent variables as metrical or categorical. In the former case we mean that they take values on some

Table 1
Classification of latent variable methods

	Manifest variables	
	Metrical	Categorical
Latent variables		
Metrical	Factor analysis	Latent trait analysis
Categorical	Latent profile analysis	Latent class analysis

continuous or discrete scale of measurement; in the latter, that they fall into categories.

This classification is not exhaustive. It does not, for example, include cases where both kinds of variable occur among the manifest (or latent) variables. Neither does it distinguish between continuous and discrete metrical variables or between ordered and unordered categorical variables. Nevertheless, it does cover the most widely used methods on which this overview focuses.

2. Historical Perspective

Factor analysis was invented by Spearman (1904) for the specific purpose of operationalizing the notion of general intelligence. He remained wedded to the view that there was a single dominant latent variable of human ability which he named 'g' which accounted for most of the variation in observed performance. Thurstone (1947) emphasized the multifactorial approach and, for that purpose, generalized Spearman's treatment to allow for several factors. Factor analysis was ahead of its time in two senses. First, it was a sophisticated multivariate technique introduced long before the statistical technology required was available. In consequence, its development was somewhat idiosyncratic and its separation from the statistical mainstream gave it an alien appearance which persists to this day and still causes suspicion and misunderstanding. Second, it is a computer intensive technique whose full potential could only be realized with the arrival of the powerful electronic computers since the 1980s. This lack distorted the early development of the subject by the need to concentrate on the search for computational shortcuts.

Latent structure analysis, which covers the remaining cells of Table 1, was introduced by Lazarsfeld after the Second World War, with a view to applications in sociology. A book-length treatment, by Lazarsfeld and Henry (1968), has held the field for many years as the definitive source of information but has now been superseded by a new generation of books of which Heinen (1996) is both recent and comprehensive.

Latent structure analysis came on the scene half a century after factor analysis, and in a different disciplinary context. For that reason there was, until very recently, little cross-fertilization between the two fields. When viewed from a level of sufficient generality, the methods are essentially the same.

In the postwar era more powerful methods became available but the whole field largely has remained a set of distinct subspecialisms each with its own literature and technology. It is over 40 years since Anderson (1959) pointed out that all latent methods shared a common conceptual framework but his insight has been slow to yield its full potential. The classification of methods as given in Table 1 was made the basis of

a common approach in Bartholomew (1984) the culmination of which is to be seen in Bartholomew and Knott (1999).

3. The Basic Conceptual Framework

First, we describe the key ideas in nonmathematical terms and then, in the following section, express them more precisely. The central idea on which the treatment is based is that of a probability model in which all variables are represented by random variables. This is so natural to the modern statistician as to pass without comment, but this was not so at the time when factor analysis was being developed. Even today the notion is not always clear. A statistical model is, essentially, a statement about the joint distribution of a set of random variables. Model-building is the process of specifying that distribution.

The discussion is based on a simple example. Consider the case of arithmetical ability and suppose that this varies from one individual to another along a one-dimensional scale but cannot be observed directly. Test items are chosen which depend on this particular skill, and are administered to a large sample of subjects. Our problem is to see whether our hypothesis about the existence of a single latent variable is correct and, if so, to provide a means of locating subjects on that scale on the evidence of their test score.

A simple intuitive approach, which has been used in the educational world for generations, is to add up, or average, the scores for any individual and treat the average as an approximation to the value of the underlying latent variable for that individual. From this it is a short step to a rudimentary model according to which the observed score is equal to the 'true score' plus a random deviation. The deviations will tend to cancel out on addition. All latent variable models can be regarded as elaborations of this basic idea.

The first step is to specify how the observed score should vary when the value of the latent variable is fixed. The next is to deduce what the joint distribution of the scores ought to be when there is no variation. The final step is to introduce the distribution of the latent variable(s)—the true scores. Together these give the overall distribution of the scores which can then be compared with the observed distribution. If the two are in agreement the model is supported, if not it must be replaced with something better.

The second stage in the model building exercise is more problematical. By definition, a latent variable can never be observed so its distribution cannot be observed either. The choice of prior distribution is thus a matter of convention. If it is chosen to be normal, that is equivalent to agreeing to adopt a scale which renders the distribution normal.

All of the models covered by Table 1 have these two elements: (a) a statement of how the scores vary when

the latent variable(s) is fixed and (b) a convention about the form of the distribution of the latent variable in the population from which the sample of those tested has come. Once these two things are known probability theory provides the means to go in the reverse direction and deduce how the latent variable would vary if the observed score were to be fixed.

4. Formal Statement of the Conceptual Framework

The last section set out the key ideas underlying all latent variable models. Here they are expressed more formally in mathematical terms. The nonmathematical reader who is interested only in specific models may pass immediately to the next section.

Denote the set of indicators by the p -vector $\mathbf{x}$. Its elements $x_1, x_2, \dots, x_p$ may be continuous or categorical. The latent variables are denoted by the q -vector $\mathbf{z}$. In the example above there was only one latent variable, arithmetical ability, but in general, there may be several.

Any model must specify, completely or partially, the joint probability distribution of $\mathbf{x}$ and $\mathbf{z}$ denoted by $f(\mathbf{x}, \mathbf{z})$. The distinctive thing about this kind of model is that the values of $\mathbf{z}$ can never be observed. For this reason everything must be expressed in terms of the distribution of the observable variables $\mathbf{x}$. The relationship between the two is

$$f(\mathbf{x}) = \int f(\mathbf{x}, \mathbf{z}) d\mathbf{z} \quad (1)$$

Here and subsequently equations will be written as if the variables were all continuous—hence the integral in (1). This is purely for convenience and the translation needed to cover other cases does not affect the logic of the argument. The next step is to write

$$f(\mathbf{x}, \mathbf{z}) = f(\mathbf{z})f(\mathbf{x}|\mathbf{z}) \quad (2)$$

This decomposition provides us with the two elements to which the description of the last section referred—namely a statement of how the manifest variables would vary if $\mathbf{z}$ were known, given here by $f(\mathbf{x}|\mathbf{z})$, and the prior distribution of $\mathbf{z}$. If both parts of (2) can be specified, $f(\mathbf{x})$ can be determined and thus the way is open to estimate any parameters of the model from a likelihood function constructed from $f(\mathbf{x})$. To scale $\mathbf{z}$, that is to find value for $\mathbf{z}$ for an individual with a given $\mathbf{x}$, the posterior distribution given by

$$f(\mathbf{z}|\mathbf{x}) = f(\mathbf{z})f(\mathbf{x}|\mathbf{z})/f(\mathbf{x}) \quad (3)$$

is needed. A scale value, or factor score, for a particular $\mathbf{z}$ can then be found by using some measure of location of its posterior distribution such as $E(z_i|\mathbf{x})$.

This program is deceptively simple. It merely maps out some of the consequences of regarding $\mathbf{x}$ and $\mathbf{z}$ as

random variables. To turn it into a model we have to specify $f(\mathbf{z})$ and $f(\mathbf{x}|\mathbf{z})$. Here we face a fundamental problem because the decomposition of (3) is not unique—there are infinitely many possible pairs each leading to the same $f(\mathbf{x})$ and hence indistinguishable empirically from one another.

The choice of $f(\mathbf{z})$ is resolved by appealing to the fact established above that $\mathbf{z}$ is a latent variable and hence that it can be constructed to have any convenient distribution. For instance, as a matter of convention one may choose the z s to be standard normal and mutually independent.

The core of the model-building exercise centers on the choice of the conditional distribution $f(\mathbf{x}|\mathbf{z})$. This has to be considered in the context of individual problems but there is one further simplification which applies to all models. This simplification involves making explicit what is already implicit in our formulation. The fact that the x s are supposed to be indicators of $\mathbf{z}$ implies that the x s will be mutually correlated—because of their common dependence on some, at least, of the z s. Hence, if $\mathbf{z}$ were to be fixed, the source of those correlations would vanish. Conditional on $\mathbf{z}$ the x s would then be independent. If this were not so it would be inferred that there was at least one other z , not already included, giving rise to the correlations remaining. Sufficient z s therefore have to be included in the model to eliminate the correlations after conditioning of $\mathbf{z}$. It then follows that

$$f(\mathbf{x}|\mathbf{z}) = \prod_{i=1}^q f_i(x_i|\mathbf{z}) \quad (4)$$

This is often referred to as the assumption of conditional (or local) independence but it is not an assumption in the usual sense. It is, rather, a statement of what is meant by saying that $\mathbf{z}$ explains the interdependence of the x s. It is, essentially, a definition of q as the smallest integer for which an equation of the form (4) can be constructed.

How this may be done in particular cases will be shown in the following section, but there is an important general principle which offers a fruitful way forward. Distributions of the exponential family are widely used in statistical modeling because they lead to estimates with highly desirable properties such as sufficiency. They provide a rationale for many of the most important models used in statistics, such as generalized linear models (see *Neural Networks and Related Statistical Latent Variable Models*).

Essentially the same properties can be utilized here. They are set out in the chapter on the General Linear Latent Variable Models in Bartholomew and Knott (1999). The essential idea is to make what is called the ‘natural’ or ‘canonical’ parameter of the one-parameter exponential family distribution a linear function of the latent variables. If that is done it turns out that the posterior distribution of $\mathbf{z}$ given $\mathbf{x}$ of (3) depends

on the x s only through q linear combinations of the x s which we may write

$$X_j = \sum_{i=1}^p \alpha_{ij} x_i \quad (j = 1, 2, \dots, q) \quad (5)$$

This is a remarkable result which justifies the common practice of 'estimating' latent variables from sums, or weighted sums, of the indicators. Technically speaking the X s are sufficient for $\mathbf{z}$ in the sense that, if we know $\mathbf{X}$, there is no further information in the data about $\mathbf{z}$.

5. Special Cases

5.1 Factor Analysis

This is the oldest and most widely used latent variable method. It is usually written

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{z} + \mathbf{e} \quad (6)$$

where $E(\mathbf{z}) = E(\mathbf{e}) = \mathbf{0}$, $\text{cov}(\mathbf{e}) = \boldsymbol{\Psi}$, a diagonal matrix, $\text{cov}(\mathbf{z}, \mathbf{e}) = \mathbf{0}$ and $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$. An equivalent way of writing it, which conforms more closely to the general treatment above, is to specify the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$ as

$$\mathbf{x} | \mathbf{z} \sim N(\boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{z}, \boldsymbol{\Psi}) \quad (7)$$

which makes it clear that $\mathbf{z}$ influences $\mathbf{x}$ only through its mean which is linear in $\mathbf{z}$. The marginal distribution of $\mathbf{x}$ is $N(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})$ and it is from this that the parameters $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ must be estimated. The matrix $\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$ is sometimes referred to as the covariance structure.

Fitting the model consists in determining that $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ which make the fitted and observed covariances of the x s as close as possible. This can be done without invoking the distributional assumptions at all since the covariance structure derived from (6) is the same, regardless of the assumptions about the form of the distributions of $\mathbf{z}$ and $\mathbf{e}$. However, without some such assumption nothing can be said about goodness of fit or the sampling variation of the estimators.

Posterior analysis allows the distribution of $\mathbf{z}$ given $\mathbf{x}$ to be determined, which for the factor model is

$$\mathbf{z} | \mathbf{x} \sim N(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}), (\mathbf{I} + \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})) \quad (8)$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$. As predictions of $\mathbf{z}$ given $\mathbf{x}$ we can use the mean values of this distribution, replacing the parameters by sample estimates.

5.2 Interpretation and Rotation

Interpretation of factors involves procedures which are common to all linear latent variable models. They

are most familiar in the context of factor analysis so, although treated here, their generality must be emphasised. The problem may arise in two forms. If there is some idea of what factors to expect (as in the arithmetical ability example above) the question is how to recognize them. If, on the other hand, the analysis is purely exploratory, further guidance is needed on how to assign meaning to any factor that is found. The essential idea is to look at the relationships between the latent and manifest variables. This is facilitated by noting that the element λ_{ij} of $\boldsymbol{\Lambda}$ may be interpreted as the correlation between x_i and z_j . Manifest variables which are highly correlated with z_1 say, have a lot in common with z_1 which, in turn is highly influential in determining those x s. It must then be asked what it is that those x s have in common with one another which they also share with z_1 . Interpretation is often facilitated by a process known as rotation. Thus far the treatment has supposed that there was only one solution to the problem of fitting the model. In fact there are infinitely many models which all predict the same covariance structure. These solutions can be generated from one another by a procedure known (because of its geometrical interpretation) as rotation. Each such solution represents an equivalent way of describing the latent space, but some may be easier to interpret than others. For example, a solution for $\boldsymbol{\Lambda}$ which has, at most, one nonzero value in each row implies that the x s have been divided into disjoint groups, each group depending on a single latent variable. That latent variable is then interpreted in the light of what that subset has in common.

5.3 The Latent Class Model

If the latent space consists of a finite (usually small) set of classes and if the manifest variables are also categorical the bottom right cell of Table 1 applies. The commonest case is where the manifest variables are binary. If there are c classes the prior distribution $f(\mathbf{z})$ becomes a discrete probability distribution over the classes, η_j being the probability of belonging to class j . In this case these probabilities can be treated as unknown parameters to be estimated. For convenience, the two values which each x_i takes are coded as 1 and 0, the conditional distributions $f_i(x_i | z)$ consists of two probabilities: $\Pr\{x_i = 1 | z\}$ and $\Pr\{x_i = 0 | z\}$. The natural choice for the distribution of x_i is the Bernoulli distribution written as

$$f_i(x_i | j) = \{\pi_i(j)\}^{x_i} \{1 - \pi_i(j)\}^{1-x_i} \\ x_i = 0, 1 \quad j = 1, 2, \dots, c \quad (9)$$

The joint probability distribution of $\mathbf{x}$ is then

$$f(\mathbf{x}) = \sum_{j=1}^c \eta_j \prod_{i=1}^p \{\pi_i(j)\}^{x_i} \{1 - \pi_i(j)\}^{1-x_i} \quad (10)$$

This can be used to construct the likelihood function from which the parameters may be estimated.

Posterior analysis for this case is concerned with allocating individuals to latent classes after $\mathbf{x}$ has been observed and this is done easily by substituting into (3). This gives us the probability that an individual with observed vector $\mathbf{x}$ belongs to each of the classes.

The number of parameters in this model increases rapidly with c and serious problems of identifiability arise if there are more than four or five classes. Also the standard errors of the parameter estimates increase rapidly as c increases.

5.4 Latent Trait Models

These models were devised, primarily, for use in educational testing where the latent trait refers to some ability. There is, thus, usually one latent variable, representing the ability, and many indicators. These are often binary, corresponding to the response to the item being 'right' or 'wrong.' This has become a specialized field, often referred to as Item Response Theory (IRT) with a literature and notation of its own (see *Factor Analysis and Latent Structure: IRT and Rasch Models*). The model may also be used in other fields, such as sociology, where it may be more appropriate to introduce several latent variables.

A latent trait model with binary x s is similar to a latent class model. The prior distribution is now continuous and will usually be taken to be standard normal. The response variables will still be Bernoulli random variables but now they depend on the continuous latent variable(s). Since the Bernoulli distribution is a member of the exponential family the appropriate form for $\pi_i(z)$ turns out to be

$$\text{logit } \pi_i(\mathbf{z}) = \alpha_{i0} + \alpha_{i1}z_1 + \alpha_{i2}z_2 + \cdots + \alpha_{iq}z_q \quad (11)$$

Other versions of the model in which the logit on the left hand side of (11) is replaced by $\Phi^{-1}(\cdot)$, the inverse standard normal distribution function, are widely used and give similar results to (11) but they lack the sufficiency properties. If $j = 1$ and the α_{ij} s are the same for all i , the model is a random effects version of the Rasch model (see, e.g., Bartholomew 1996). In the latter the trait values are taken as parameters rather than random variables so, in the strict sense of classical inference, the Rasch model is not a latent variable model.

5.5 Other Models

The latent profile models of Table 1 have been omitted because they have found relatively little practical application. It has also been shown by Molenaar and

von Eye (1994) that they are virtually indistinguishable from factor models because they have an equivalent covariance structure. Neither have we considered hybrid models in which manifest and latent variables may be mixed type—discrete or continuous. The advantage of our general approach is that it lends itself to the treatment of all such models. This approach has been exemplified by Moustaki (1996) and is capable of further development.

6. Relationship between Latent Variables

In much sociological research the interest is not so much in the latent variables as individual entities, but in the relationships between them. The interrelationship of manifest variables is a major subject of investigation in social science using techniques such as path analysis, regression, log-linear analysis, and graphical models. It is natural to extend these models to include latent variables. This has given rise to linear structural relations modeling which can be implemented using widely available software such as LISREL, AMOS, and EQS (see *Factor Analysis and Latent Structure, Confirmatory; Latent Structure and Causal Variables*).

In essence, such models have two parts: (a) a measurement model which supposes that each latent variable is linked to its own set of indicators through a factor model and (b) a structural model which specifies the (linear) relationships among the latent variables. Such a model is fitted by matching the observed and theoretical covariance matrices. A much more general framework which allows a wider range of models is provided by Bartholomew and Knott (1999). Although these methods are very widely used, serious questions have been raised about the identifiability of the models (Anderson and Gerbing 1988, Croon and Bolck 1997, Bartholomew and Knott 1999). These authors have suggested that it may be better to separate the measurement and structural parts of the analysis. This can be done by constructing 'estimates' of the latent variables and then exploring their interrelationships by the traditional methods used for manifest variables.

7. Future Developments

The advent of massive computer power has changed the practice of multivariate analysis radically, and of latent variable analysis in particular. The limiting factor is no longer computing power but of getting data of sufficient quality and quantity to fit the very complicated models which the theory provides and computers can handle. The precision with which models with many parameters can be estimated is

often very low unless the sample size runs into thousands. This makes it all the more important to estimate the sampling variability of the estimates of the parameters on which the interpretation depends. Traditionally this has been done by finding asymptotic standard errors but these can be very imprecise. However, it is now possible to supplement these results by resampling methods such as the bootstrap.

There are other fields in which latent variable models are used which currently exist in isolation. An obvious generalization is to latent time series. Some work has been done for the case where the latent process is a Markov chain. In this area the term 'hidden' is used instead of 'latent' which helps to conceal the family connections (see *Neural Networks and Related Statistical Latent Variable Models*) (for an introduction see MacDonald and Zucchini 1997). An application in a more traditional time series context will be found in Harvey and Chung (2000). Also, there is work by economists on unobserved heterogeneity as it is called which, essentially, involves the introduction of latent variables into econometric models.

Bibliography

- Anderson T W 1959 Some scaling models and estimation procedures in the latent class model. In: Grenander U (ed.) *Probability and Statistics*. Wiley, New York
- Anderson J C, Gerbing D W 1988 Structural equation modeling in practice: a review and recommended two-step approach. *Psychological Bulletin* **103**: 411–23
- Bartholomew D J 1984 The foundations of factor analysis. *Biometrika* **71**: 221–32
- Bartholomew D J 1996 *The Statistical Approach to Social Measurement*. Academic Press, San Diego, CA
- Bartholomew D J, Knott M 1999 *Latent Variable Models and Factor Analysis*, 2nd edn. Arnold, London
- Croon M, Bolck A 1997 On the use of factor scores in structural equations models. *Technical Report 97.10.102/7*. Work and Organization Research Centre, Tilburg University
- Harvey A, Chung C-H 2000 Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society A* **163**: 303–39
- Heinen T 1996 *Latent Class and Discrete Latent Trait Models: Similarities and Differences*. Sage, Thousand Oaks, CA
- Lazarsfeld P F, Henry N W 1968 *Latent Structure Analysis*. Houghton Mifflin, New York
- MacDonald I L, Zucchini W 1997 *Hidden Markov and other Models for Discrete-valued Time Series*. Chapman and Hall, London
- Molenaar P C W, von Eye A 1994 On the arbitrary nature of latent variables. In: von Eye A, Clogg C C (eds.) *Latent Variables Analysis: Applications for Developmental Research*. Sage, Thousand Oaks, CA
- Moustaki I 1996 A latent trait and latent class model for mixed observed variables. *British Journal of Mathematical and Statistical Psychology* **49**: 313–34
- Spearman C 1904 General intelligence, objectively determined and measured. *American Journal of Psychology* **15**: 201–93

Thurstone L L 1947 *Multiple Factor Analysis*. University of Chicago Press, Chicago

D. J. Bartholomew

False Memories, Psychology of

A false memory is a mental experience that is mistakenly taken to be a veridical representation of an event from one's personal past. Memories can be false in relatively minor ways (e.g., believing one last saw the keys in the kitchen when they were in the living room) and in major ways that have profound implications for oneself and others (e.g., mistakenly believing one is the originator of an idea or that one was sexually abused as a child). False memories arise from the same processes as do true memories and hence their study reveals basic mechanisms of memory. This article describes empirical research investigating false memories and a theoretical approach—the source monitoring framework—for integrating the findings and guiding further investigation.

1. Selected Early Research

Psychologists have long been interested in memory distortions. A classic example from the 1930s is Bartlett's report of studies in which he told people a folk tale from an unfamiliar culture and asked them to recall it from memory. He noted that the memory errors people made brought the story more in line with their own culturally determined expectations. Another is Carmichael, Hogan, and Walter's demonstration that how forms are labeled influences how they are remembered. For example, people's reproductions from memory of two circles connected by a straight line were more likely to include a curved line if the form had been labeled as 'eyeglasses' than if it had been labeled as 'dumbbell.' Still another is work in the late 1950s and early 1960s by Deese (1959) and by Underwood (1965) demonstrating that people are very likely to falsely remember an item (e.g., needle) if they had earlier been presented with related items (e.g., thread, haystack, sharp).

In the 1970s, studies from a number of laboratories highlighted the fact that such memory distortions do not only occur for unfamiliar materials such as folk tales, or lists of words or pictures—they are a byproduct of our everyday understanding of, and memory for, information and events. For example, Bransford and Johnson's (1973) work on comprehension and memory demonstrated that recalling information often depends on engaging constructive processes by which information is related to prior knowledge or schemas at encoding. For example, memory for a sentence such as 'the haystack was

important because the cloth ripped' was much better when the idea of a parachute had been recently activated than when it had not. However, this work also showed that people are more likely to falsely claim that stories included information that was only inferred based on prior schemas or knowledge. People who heard a story that included the information that 'the spy threw the secret document into the fireplace just in time because 30 seconds longer would have been too late' were later likely to claim that 'the spy burned the document.' Thus distorted 'remembering' of information that might be, but is not necessarily, true (e.g., the spy could have been hiding the document in a fireplace that was not lit) is the potential downside of the intelligent, active processing of information.

Such findings led Johnson and Raye (1981) to suggest that memory distortions like those described above reflect errors arising from imperfect reality monitoring processes that are an integral part of remembering: that is, people sometimes confuse the information they generate during the initial encoding, the retention interval, or subsequent remembering of an event for information that came from the event. Reality monitoring is a special case of the more general ongoing function of source monitoring—making attributions about the origin of activated information in mental experience (Johnson et al. 1993). Thus, source monitoring errors include both confusions between internal and external sources and between various external sources (e.g., attributing something that was imagined to actual perception, an intention to an action, something read in a tabloid to a television program, an incident that occurred in place A or time A to place B or time B).

Of course, how serious any of these memory distortions are depends on what is at stake. A clear case where it matters is eyewitness testimony. In the 1970s, E. Loftus launched an influential research program investigating the impact on event memory of information that is suggested between the time of a witnessed event and the time people are required to 'testify.' In one study, people saw a film of an automobile accident and were asked some questions, including 'About how fast were the cars going when they smashed into each other?' Other people were asked the same question except that the words 'smashed into' were replaced with the word 'hit.' On a memory test a week later, the people in the first group were more likely to mistakenly say 'yes' to a question about whether they had seen broken glass in the film. To take another example from the Loftus laboratory, after seeing a film, participants were asked 'How fast was the white sports car going when it passed the barn while traveling along the country road?' Other participants were asked the same question without mention of the barn. In fact, there had been no barn in the film, but those people who had the barn mentioned in the question were later more likely to say that they had seen a barn in the film. Both of these examples likely

involve source monitoring errors—the former a confusion between a self-generated inference based on the supplied information that the cars 'smashed into' each other and the latter a confusion between what was read and what was seen.

In sum, generation, elaboration, and integration of information across individual experiences from different sources reflects associative, imaginative, and reasoning processes that are necessary for all higher-order, complex thought. But, this very capacity for creativity makes us vulnerable to having false memories.

2. *Source Monitoring*

Perhaps the most comprehensive theoretical account of false memories to date is provided by the source monitoring framework (SMF) proposed by the present author and her colleagues. According to the SMF, it is not the case that memories are found (or not found), but rather that mental experiences are attributed to memory (or not) by ongoing judgment processes. The SMF further highlights several key aspects of these memory attributions:

(a) Memory attributions are based on various qualitative characteristics of the mental experience. For example, perceptual, spatial, temporal, or emotional details typically are taken as evidence that a mental experience reflects a true memory.

(b) Memory attributions are influenced by the embeddedness of the mental experience. Embeddedness depends on such factors as the availability of supporting memories, consistency with knowledge and beliefs (e.g., plausibility), and coherence of the information, and agreement with the reports of others about the event.

(c) Memory attributions are made according to flexible criteria (which qualities are considered and how they are weighted, how much evidence of any given type is needed). Hence, what may be taken for a memory under one set of circumstances might not be under another.

(d) Goals, beliefs, and motivational and social factors influence what characteristics are looked for, how much embedding occurs, and which criteria are applied.

According to the SMF, false memories occur because the mental experiences arising from events of different sorts (e.g., imagination and perception) overlap in characteristics (they are imperfectly differentiated) and because the processes that make judgments about these mental experiences are also imperfect (i.e., they not only operate on imperfectly differentiated data, they are not always fully engaged, they sometimes overweight nondiagnostic evidence or employ inappropriate criteria, they are subject to social influences, etc.). As do most theoretical accounts of false memories, the SMF assumes that the as-

sociative, constructive, elaborative, and integrative processes that are central to human perception and thought contribute information that may become a candidate for misattribution. However, the SMF has emphasized that these processes alone do not produce false memories. Rather, false memories arise when the products of these processes are taken *to be* memories. That is, it is not the fact that associations and other elaborations occur that produces false memories—for example one can make an inference and later attribute it to an inference, or imagine an event and later correctly attribute it to imagination. Survival would be difficult, if not impossible, if there were no differences between the memory records of things that happened and things we imagined, or between activities we observed or participated in and activities we only read or heard about. Fortunately, the imperfect system people have is usually good enough, and perhaps even has some advantages over a perfect system (e.g., life satisfaction may be higher when we remember things as better than they were, social relations may be better when we agree on a common account of events, or generalization between similar situations may be faster when we are not concerned about where inferences come from).

Some theorists (e.g., Gardiner, Jacoby, and Kelley) have proposed that there are two distinct types of memory experience, ‘Knowing and Remembering,’ or ‘Familiarity and Recollection,’ and have shown that errors often result from misattribution of familiarity which can sometimes be corrected with recollection. Within the SMF, these would represent ends (or a coarse categorization) of a continuum of mental experiences. Also, in the SMF, both true and false memories can be associated with a sense of familiarity and with a sense of recollection.

3. Investigating True and False Memories

Evidence for the propositions in Sect. 2 comes from laboratories in many countries working in many domains (the SMF can be thought of as a way of unifying this evidence), including work on familiarity-based attributions, cryptomnesia, the phenomenal qualities of memories, suggestibility, rehearsal, schema effects in memory, the impact of emotion, motivational and social effects in memory, interpersonal reality monitoring, and imagination. Several illustrative findings are briefly outlined here.

3.1 Memory Characteristics

When participants are asked to make ratings of the qualitative characteristics of their memories, memories for real events tend to have more perceptual and emotional detail than memories for imagined events. Moreover, when they are asked to justify why they

believe their own (or someone else’s) memory they invoke these characteristics or are influenced by them. Manipulating these characteristics influences the accuracy of source monitoring. For example, if participants have imagined a word in another person’s voice they are later likely to claim that person said the word. If they have seen a magnifying glass, they are likely to say they have seen a physically similar item that they only imagined (e.g., a lollipop). Furthermore, features are ‘borrowed’ or cumulated across modalities. For example, hearing the sound of a dog barking increases subsequent claims that an imagined dog was seen. In some experiments, people are as likely to claim to ‘remember’ their false as their true memories and to be as confident (or more so) about them. However, in several experiments using more detailed rating categories (e.g., asking for specific ratings of visual and auditory detail, emotional qualities) false memories were on average somewhat less vivid or detailed than true memories.

3.2 Evidence and/or Criteria used in Source Monitoring

Attribution processes are subject to task and social demands. If people are induced to, or have the opportunity to more carefully evaluate their memories, the probability of false memories decreases. Ways to decrease false memories include avoiding leading questions, asking people to explicitly consider the possible sources of their memories rather than simply asking about one source, having them rate the qualitative characteristics of their memories, and public as opposed to private remembering. Presumably, these manipulations decrease the degree that people will rely on a single type of information (e.g., familiarity or perceptual detail) and increase consideration of, and retrieval of, additional potentially useful information.

3.3 Rehearsal and Imagination

Thinking about events can increase the vividness of imagined events, the probability people will claim imagined events occurred, and their confidence that an event occurred. Imagining events can also make similar perceived events seem to have occurred more frequently. Furthermore, inducing subjects to generate images of suggested information increases later source misattributions.

3.4 Attention

Emotional self-focus or distraction during encoding or remembering can result in decreases in source accuracy without necessarily decreasing memory that an event

occurred. Such findings emphasize that memory 'strength' (in the sense of probability of recognition or recall) is not a perfect predictor of source memory. People can have strong memories for information for which they have only vague source cues, or strong-but-wrong source cues.

4. Are Some Memories Immune to Distortion?

In 1977 Brown and Kulick proposed that there is a class of 'flashbulb memories'—accurate, long lasting, and detailed—for highly significant, emotional events (e.g., where one was when one heard that Kennedy was assassinated or that the Challenger space shuttle exploded). However, subsequent research has shown that, like ordinary memories, these memories are subject to distortion.

The question of whether some memories might be immune to distortion resurfaced in the 1980s and 1990s, prompted in part, by high profile legal cases that involved either children's testimony in sexual abuse cases (e.g., *New Jersey vs. Michaels*), or adults' testimony about recovering previously repressed memories that they were sexually abused (e.g., *Ramona vs. Ramona*; *Martinelli vs. Diocese of Bridgeport*), or witnessed crimes (e.g., *People vs. Franklin*). Such cases greatly broadened the interest among the research community in the processes of memory distortion and the number of published papers related to memory distortion increased dramatically.

A central issue in discussions about these cases has been to what extent the findings and principles arising from laboratory research apply to real life, especially to traumatic memories. Some clinicians and authors of popular self help books encourage individuals experiencing psychological problems to freely imagine what might have happened to them as children and to think about why such events might have been likely. Although intended to help clients, the potential consequences of such practices are clearly problematic from the perspective of the SMF. Such practices encourage the client to generate details that may not be accurate, embed false information in a web of supporting knowledge and beliefs, and to adopt weak criteria for what constitutes evidence for a memory. At the same time, the conditions of laboratory studies cannot mirror those of real-life traumatic situations. Thus some have suggested that experimental results and some theoretical ideas do not always apply. They claim that certain classes of traumatic memory, including childhood sexual abuse, are unlikely to be false because the associated emotion is evidence of their authenticity.

Although there are obvious ethical limits to the types of laboratory situations that researchers can set up, researchers have demonstrated that false memories can be created for entire, complex, salient, and

emotionally significant events—inducing both children and adults to believe that they remember autobiographical events that never happened. The general procedure in some of these studies is to obtain reports of actual events from a relative of the participant and then query the participants about these events. Embedded in the list of real events is a false event that did not happen to the participant. After reading the accounts and, especially after being encouraged to think about the events, some participants claim to remember the false event. For example, Loftus and Pickrell induced adult participants to remember being lost in a shopping mall as a child. Hyman and colleagues induced participants to remember being taken as a child to the hospital for an ear infection. These false memories can contain quite specific details not provided in the relative's account. Similarly, Ceci and his colleagues have found that some children will provide compellingly vivid accounts of complex events (e.g., having their finger caught in a mousetrap), particularly after repeated questioning about the event. Spanos and colleagues demonstrated that adult participants could also be led to report memories from infancy. For example, participants were told that, because they have good visual skills, they were likely to have been born in a hospital that hung mobiles over their cribs and then they were 'age regressed' to recover the memories. Groups that were hypnotized and groups that were not both reported a high incidence of 'infant memories.' Spanos and colleagues suggest that memories of past-life identities, UFO abductions, and satanic ritual abuse are similarly likely to be created in a social context where authoritative or trusted sources make such events seem plausible, create the expectation that the individual has had such experiences, and use techniques such as generating imagery or repeated questioning to facilitate recovery of such hidden memories. Ofshe has also argued that induction of hypnotic or dissociative states, in combination with specific suggestions from authority figures, can result in false memories and provided a chilling account of a criminal investigation in which leading questioning techniques induced a man to confess to the sexual abuse of his children.

Of course, it is quite difficult to isolate particular factors contributing to false autobiographical memories in these relatively complex, emotionally salient situations. Nevertheless, overall the results of these studies appear to reflect the processes proposed and they increase confidence in the generality of laboratory findings. Repeated questioning or thinking about an event increases the details that are remembered or confidence in the memory. Encouraging participants to embed a 'memory' in other supporting personally relevant details increases false memories. Individuals with high imagery ability seem to be more susceptible to induced false memories, presumably because they embellish more or create representations that are more like perceptions. In some studies, individuals who

score high on tests of hypnotizability or dissociative experiences are also more susceptible, perhaps because these individuals are high in imagery ability or easy to coax into using lax criteria for assuming something is a memory.

Studies that induce false autobiographical memories highlight the fact that remembering often takes place in a social context that can be a source of false information, can provide occasions for repeating and embellishing 'memories,' and can coax people to ignore the doubt that is an important cue during normal source monitoring. Therapy is itself a social context that potentially can have these characteristics. In 1995, Poole in the US and Lindsay in Canada, in collaboration with UK researchers Memon and Bull, reported results of a survey of therapists in the US and UK: 25 percent of the sampled therapists who work with adult female clients believe 'that recovering memories is an important part of therapy, think they can identify clients with hidden memories during the initial session, and use two or more techniques to help such clients recover suspected memories of [childhood sexual abuse].' They note that such techniques may be used with caution by many therapists, but that other therapists may be unfamiliar with the cognitive literature on processes of memory distortion, and may underestimate their own influence. As professional books and journals and, especially, the popular press, have provided wider exposure to theoretical issues and research findings, therapists are likely becoming better informed for making judgments about which techniques to use. Importantly, however, not only therapists, but also police officers, lawyers, social workers, parents, and many other trusted authorities (e.g., talk show hosts, Internet sites) are in a position to influence how people search and evaluate their memories. There is no reason to believe that memories cannot be accurate (they often are quite accurate in both laboratory and more naturalistic studies), and appropriate cueing and some conditions of social remembering can facilitate memory as well. Nevertheless, the potential for memories to be false is clear and the mechanisms of distortion are generally understood. Recent work has been directed at developing interview techniques for reducing false memories in both children and adults.

5. The Prefrontal Cortex and Source Monitoring

Based on the evidence that memory involves constructive and reconstructive processes (including the self-generation of retrieval cues, setting criteria, and evaluation of activated information), one would expect the region most associated with reflective or executive processes—the prefrontal cortex (PFC)—to be important for episodic memory, especially distinguishing between true and false memories. And, indeed, damage to PFC often produces memory deficits, including deficits in self-generated retrieval

necessary for autobiographical recall, increases in false recognition of related distracters, and, in some cases, clinically significant confabulation. Source memory deficits (and resulting false memories) are more likely in children (whose frontal lobes are slow to develop) and in older adults (who are likely to show increased neuropathology in PFC with age). PFC dysfunction may play a role in schizophrenia, which sometimes includes severe source monitoring deficits in the form of delusions. Furthermore, in cognitive tasks given to neurologically intact young adult participants, increasing the cognitive load (having the participant simultaneously engage in a secondary task) disrupts performance more on tasks requiring source identification than on tasks simply requiring old/new recognition. Presumably, the extra load taxes the recruitment, organization, and integration of cognitive component processes necessary for source identification that are subserved by PFC.

A few studies have directly examined the brain activity associated with true and false memories. These studies show that there is overlap in the brain regions activated for correct recognition of old items and recognition of semantically related lures, as would be expected based on the behavioral evidence of their confusion. However, evidence also suggests that the similarity of the brain activity associated with true and false memories depends on how people are tested (e.g., with targets and lures randomly intermixed or in separate lists). This is consistent with the idea that what people are looking for as they are remembering will influence the qualities of their memories.

6. Conclusions

Autobiographical memories are narratives that are influenced by expectations before events ever happen and by rumination after the fact. These narratives are also influenced by other experiences (photographs, other people's accounts, and even unrelated events) and our goals and motives at the time of remembering. False memories arise from the same encoding, rehearsal and source monitoring (memory attribution) processes that produce true memories; thus one can never be absolutely sure of the truth of any particular memory. Of course, in many contexts, minor (and even major) distortions are of little practical consequence. There may even be some advantages to certain kinds of false memories (e.g., remembering a vacation as having been more pleasant than it was). However, there are other contexts in which it does matter whether memory is accurate—when distorted memories result in consequences that are clearly undesirable from the personal (e.g., mistakenly believing one was a victim), social (e.g., giving erroneous testimony about someone's actions), or professional (e.g., appropriating someone else's ideas) perspectives. Furthermore, certain conditions of psychopathology or

brain damage result in clinically significant increases in false memories and beliefs—confabulations and delusions. Hence understanding the mechanisms underlying true and false memories can potentially help reduce their occurrence under everyday circumstances (e.g., in therapy, in police interrogations, in scholarship) and help specify more completely and assess cognitive dysfunction in clinical populations.

See also: Comprehension, Cognitive Psychology of; Confessions: Psychological and Forensic Aspects; Constructivism in Cognitive Psychology; Elaboration in Memory; Emotion in Cognition; Eyewitness Memory: Psychological Aspects; Inferences in Discourse, Psychology of; Interference and Inhibition, Psychology of; Memory and Aging, Cognitive Psychology of; Memory: Autobiographical; Memory Development in Children; Memory for Meaning and Surface Memory; Memory Retrieval; Mood-dependent Memory; Reconstructive Memory, Psychology of; Schemas, Frames, and Scripts in Cognitive Psychology

Bibliography

- Bartlett F C 1932/1995 *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press, New York
- Bottoms B, Goodman G (eds.) 1996 *International Perspectives on Child Witnesses: Psychological Research and Law*. Sage, Thousand Oaks, CA
- Bransford J D, Johnson M K 1973 Considerations of some problems of comprehension. In Chase W (ed.) *Visual Information Processing*. Academic Press, New York, pp. 383–438
- Brown R, Kulik J 1977 Flashbulb memories. *Cognition* 5: 73–99
- Bruck M, Ceci S J 1999 The suggestibility of children's memory. *Annual Review of Psychology* 50: 419–39
- Carmichael L, Hogan H P, Walter A A 1932 An experimental study of the effect of language on the reproduction of visually perceived form. *Journal of Experimental Psychology* 15: 73–86
- Ceci S J, Crotteau Huffman M L, Smith E, Loftus E F 1994 Repeatedly thinking about non-events. *Consciousness and Cognition* 3: 388–407
- Christianson S-A (ed.) 1992 *The Handbook of Emotion and Memory: Research and Theory*. Erlbaum, Hillsdale, NJ
- Conway M A (ed.) 1997 *Recovered Memories and False Memories*. Oxford University Press, Oxford, UK
- Deese J 1959 On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology* 58: 17–22
- Hyman I E, Husband T H, Billings F J 1995 False memories of childhood experiences. *Applied Cognitive Psychology* 9: 181–97
- Johnson M K, Hashtroudi S, Lindsay D S 1993 Source monitoring. *Psychological Bulletin* 114: 3–28
- Johnson M K, Raye C L 1981 Reality monitoring. *Psychological Review* 88: 67–85
- Johnson M K, Raye C L 1998 False memories and confabulation. *Trends in Cognitive Sciences* 2: 137–45
- Lindsay D S, Read J D 1994 Psychotherapy and memories of childhood sexual abuse: A cognitive perspective. *Applied Cognitive Psychology* 8: 281–338
- Loftus E F 1997 Creating false memories. *Scientific American* 277: 70–5
- Loftus E F, Pickrell J E 1995 The formation of false memories. *Psychiatric Annals* 25: 720–5
- Ofshe R J 1992 Inadvertent hypnosis during interrogation: False confession due to dissociative state: Mis-identified multiple personality and the satanic cult hypothesis. *International Journal of Clinical & Experimental Hypnosis* 40: 125–56
- Pezdek K, Banks W P (eds.) 1996 *The Recovered Memory/False Memory Debate*. Academic Press, San Diego, CA
- Poole D A, Lindsay D S, Memon A, Bull R 1995 Psychotherapy and the recovery of memories of childhood sexual abuse: US and British practitioners' opinions, practices, and experiences. *Journal of Consulting and Clinical Psychology* 63: 426–37
- Read J D, Lindsay D S (eds.) 1997 *Recollection of Trauma: Scientific Evidence and Clinical Practice*. Plenum, New York
- Schacter D L (ed.) 1995 *Memory Distortion: How Minds, Brains, and Societies Reconstruct the Past*. Harvard University Press, Cambridge, MA
- Spanos N P, Burgess C A, Burgess M F, Samuels C, Blois W O 1998 Creating false memories of infancy with hypnotic and non-hypnotic procedures. *Applied Cognitive Psychology* 13: 201–18
- Tulving E, Craik F I M (eds.) 2000 *The Oxford Handbook of Memory*. Oxford University Press, New York, Chap. 10, 12, 14, 24
- Underwood B J 1965 False recognition produced by implicit verbal responses. *Journal of Experimental Psychology* 70: 122–9

M. K. Johnson

Familial Studies: Genetic Inferences

The twentieth century may be described as the century of genetics. It began with the rediscovery of Mendel's work by Johannsen (1903) and ended with the completion of the Human Genome Project (HGP). *Nature* (2001, 409) published the *Initial sequencing and analysis of the human genome* by The Genome International Sequencing Consortium and *Science* (2001, 291) published similar sequencing by Celera. The twenty-first century, particularly its first half, will see a spate of work related to gene manipulation and genetic disorders. We can confidently look forward to substantial developments in our understanding of our species.

We fear that even in the post-HGP era, some of the twentieth-century errors relating to inferences from genetic analyses of familial data will be repeated and some more will be added. Indeed, McGuffin et al. (2001) say 'The most solid genetic findings about individual differences in human behavior come from quantitative genetic research such as twin and adoption studies that consistently converge on the conclusion that genetic variation makes a substantial contribution to the phenotypic variation for all behavioral domains.' The conceptual and mathematical errors in behavior genetic models, therefore, need to

be known widely. Behavioral traits are also described as 'complex' traits (McGuffin et al. 2001, Carroll 2001). Feldman and Lewontin (1975) define a complex trait as one whose frequency of occurrence cannot be analysed in terms of simple genetic hypothesis. We prefer Mather's (1949) description of such a trait as 'polygenic' as it indicates clearly the basic genetic hypothesis. This article is concerned with the inheritance of a polygenic trait. It will: (a) explain the nature of the statistical model used for the statistical analysis of a polygenic trait, (b) show that heritability is not defined in the presence of assortative mating, (c) explain Galton's ideas about nature-nurture contributions and their extension in the twentieth century, (d) clarify some confusions about correlation and regression, (e) discuss heritability, and (f) show that behavior geneticists use incorrect formulas in model fitting. As an example, we choose Devlin et al. (1997) because it was published in *Nature*, and show that the formulas they used were incorrect. We state why Fisher's (1918) formulas are wrong and point out the deficiencies of Jinks and Fulker (1970). We explain why phenotypic, that is, observed variance of a trait cannot be partitioned into genetic and environmental components. Finally, we state how the research in genetic disorders should progress.

1. Mathematical Models in Biology and Medicine

There is a long tradition of using mathematical models in medicine and related areas. For example, Cullen (1983) divides his book *Mathematics for the Biosciences* into four major areas: (a) physiology and medicine, (b) ecology and population, (c) fisheries and oceanography, and (d) genetics. Roach (1984) in *Mathematics in Medicine and Biomechanics* has chapters on blood modeling, oxygen transport, nerve modeling, etc. Mathematical models have also been used in epidemiology, neurophysiology, cardiovascular simulation, tubercular skeleton, body fluids, kidneys, etc. All these models are type I models and have proved very useful in medical research. Recently, type II models to estimate the heritability of a trait have been introduced in familial aggregation studies. We give a brief description.

A Type II model is also known as the Components of Variance model. Its basic structure is: observed value = constant + linear function of *random* variables + residual. Residuals are assumed to have 0 mean and σ^2 variance (the assumption of normality of residuals is required for tests of significance). Here, we sample the variables at random from all the variables that affect a disorder. The problem is that we may not know all variables that affect it. Therefore, random selection of variables is very important. Our interest lies not in the mean effect of a factor but in the contribution it makes to the total variance. Thus, the aim is to divide the total (phenotypic) variance of a

disorder into the sum of variances due to factors. Numerical analyses for type I and type II models are similar and this may cause confusion in the minds of some nonspecialist researchers. We emphasize that no treatment of a disorder can be deduced from a type II model.

2. Heritability

In genetics, narrow heritability (h^2) and broad heritability (H^2) are strictly defined. h^2 = additive variance/phenotypic variance and H^2 = genetic variance/phenotypic variance. Additive variance is the variance of additive values and genetic variance is the variance of genetic values. The genetic value of an allele at a locus is defined as the regression of the phenotypic values on genotype and the additive value is defined as the regression of genetic values on genotype. Thus, both additive value and genetic value at a locus *are statistical constructs*. We can calculate phenotypic variance directly from phenotypic values. We cannot, however, calculate additive or genetic variance directly, as neither additive nor genetic values are known (Capron et al. 1999). Geneticists have devised various methods for estimating h^2 and H^2 . For example, if we know *beforehand* that a trait is under genetic control, then assuming random environment and random mating with respect to a trait, h^2 = corr. (mid-parent, child).

Jensen (1969) used the concept of heritability for analysis of IQ data. Since then an enormous literature on heritability has built up. Layzer (1999) questions the assumption that phenotypic measurement = genetic component + an environmental component, as 'every phenotypic trait is a *function* of genetic and nongenetic variables.' Feldman and Lewontin (1975) said: 'As we show below, the partition of the causes of variation is really illusory and the analysis of variation can not really separate the variation that is the result of genetic segregation. *The genetic variation depends on the distribution of environments and the environmental variance depends on the distribution of genotypes*' (italics added).

We agree with these criticisms but the nature of our criticism is slightly different. We believe that heritability is not defined when the population mates assortatively with respect to the trait. Fisher (1918) showed that assortative mating will introduce association between similar alleles and, thereby, destroy independent segregation of alleles. This association will also destroy the linearity of regression on which the definition of additive values is based. We do not know how to define additive value at a locus when a population is at equilibrium under assortative mating. Moreover, it can be shown that assortative mating will also create association between additive and dominance deviations at different loci. Thus, in the presence of assortative mating narrow heritability is not

defined. Fisher (1951) did not approve of the concept of heritability.

Nonrandom environment also causes problems. The trait is then associated with environmental variables. It can be shown, if assortative mating continues, it will create association between genetic alleles and environmental factors. The additive and genetic values cannot now be defined. Fisher may have known this and assumed random environment (see Kempthorne 1969). Moreover, few of us are capable of identifying all aspects of environment. To equate environment in genetic studies with the number of cars, or the number of rooms, or the number of books in a family home, etc. is simplistic (see Sect. 10). If a researcher wishes to study the effect of books, cars, or rooms on a trait, they should use a Type I model. Heritability analysis is a genetic cul-de-sac. It suggests no remedy for a disorder. With high H^2 one can only argue incorrectly, as Jensen (1969) did, that intervention programmes are useless.

Concerning comparisons between two groups using heritability, we agree with Feldman and Lewontin (1975) that 'the concept of heritability is of no value for the study of differences in measures of human behavioral characters between groups' (see also Thoday 1973).

3. Type II Model Fitting on Twin Data

This section is in two parts: (a) Galton's ideas, and (b) later attempts to divide the phenotypic variance into nature and nurture components.

3.1 Galton's Nineteenth-century Ideas

Galton was not aware of Mendel's work. His main interest lay in Eugenics, that is, the improvement of the white 'racial' stock. He invented the concept of correlation. He was well-off and endowed the Galton Laboratory for National Eugenics to propagate eugenics. Karl Pearson was its first Director. He devised the commonly used formula for product moment correlation following Galton's concept of correlation. Galton thought that the difference between concordances of monozygotic (MZ) and dizygotic (DZ) twins on a trait gave an indication of genetic (nature) effects on that trait. Galton's views have had a powerful effect in genetics and social sciences. His *law of familial regression* has been used to 'prove' that a trait is determined genetically.

Galton's method of concordance is explained in textbooks on genetics (e.g., Snustad et al. 1997 and Hartl and Clark 1997, etc.). In exercise 15-6 Levitan (1988, p. 361) asks: 'What can be said about the relative roles of heredity and environment in the following studies?' and lists a number of studies with respective MZ and DZ concordances. The first on the list is

tuberculosis (TB). The ratio of MZ and DZ concordant twins are 202/381 and 187/843. Perhaps he expects the students to compare the two concordances and conclude that heredity plays a major role in causing TB. This, indeed, was the 'scientific' view of the researchers in the field of TB based on analysis of familial data until the late 1950s (Capron et al. 1999). We now know that this is not the case.

Vogel and Motulsky (1996) give a similar table (6.22, p. 237) with larger sample sizes. They also calculate the statistics 'MZ higher than DZ.' Its highest value, 2.9, is for leprosy and the second highest, 2.56, is for TB. They say, 'In Western Europe, for example, leprosy disappeared during the seventeenth and eighteenth centuries without any therapy, only due to improvement in living conditions. *There was probably little or no influence of genetic changes*' (p. 237, italics added). They then claim that 'Analysis of discordance can shed some light on genetic versus environmental factors in disease.' They are obviously wrong. The only reasonable conclusion from these data is that the concordance method has no value in genetics. Note that heritability, as defined above, cannot be estimated from concordances.

3.2 Nature–Nurture Methods in the Twentieth Century

Researchers in twin studies soon realised the deficiencies of Galton's method and devised new methods to estimate nature–nurture contributions. Holzinger (1929) gave an index $H^2 = (r_{mz} - r_{dz}) / (1 - r_{dz})$ in terms of intraclass correlations between MZ and DZ twins. Nichols (1965) proposed an 'improved' index $I = 2(r_{mz} - r_{dz}) / r_{mz}$. In these formulas r_{mz} and r_{dz} are the phenotypic correlations between MZ and DZ twins, respectively. Obviously, neither of these formulas estimates h^2 or H^2 . Jensen (1967) realised this and published his formula for estimating broad heritability, H^2 from twin data in the PNAS. He began by stating a complex 'generalized formula' but did not say where it came from. That formula has *no* theoretical basis. After 13 steps he stated the formula $H^2 = (r_{mz} - r_{dz}) / (1 - \rho_{dz})$ where ρ_{dz} is the genetic correlation between DZ twins. Comparison of Jensen's and Holzinger's formulas shows that Jensen had replaced the phenotypic DZ correlation r_{dz} with the genetic DZ correlation ρ_{dz} . We do not know the value of the genetic correlation ρ_{dz} . Jensen (1967) got himself into a circular situation.

How did he remove the circularity? He claimed that we could find ρ_{dz} ($= \rho_{oo}$, the genetic correlation between offsprings) by using the formula $\rho_{oo} = (1 + \rho_{pp}) / (2 + \rho_{pp})$ where ρ_{pp} is the genetic correlation between parents on the trait, and gave Li (1955, Chap. 13) as reference for the formula. It does not appear there and has no genetic basis. Clearly, circularity remains. One now needs an estimate of h^2 to obtain ρ_{pp}

from r_{pp} . When Jensen's formula was used to calculate heritability in England's largest study of twins, Vetta (1977) was able to inform the researchers that this formula is not correct. Their response was 'It follows that the formula ought not to be perpetuated' (Adams et al. 1977).

The quest for a formula for estimating heritability from twin data continues. Vogel and Motlusk (1996, p. 769) said 'Twin data can be utilised as an alternative way to get heritability estimates.' As to how it can be done, they said 'An empirical way ... is to calculate alternative estimates from the same (twin) data to determine how well they coincide.' They made a number of unrealistic assumptions and developed formulas for finding h_1^2 and h_2^2 . They then claimed that H^2 lies between h_1^2 and h_2^2 . We cannot commend their formulas.

4. Correlation and Regression Tell Us Nothing about Cause and Effect

The formulas given in the last section use MZ and DZ correlations. Geneticists borrowed the concepts of correlation and regression from statistics. Confusion regarding their significance permeates the scientific literature. For example, Lawrence and Jinks (1973) said: 'This tendency for like to beget like, or in statistical terms, for the stature of parents and offspring to be positively correlated, is a sure sign of genetic determination.' The concept of regression to the mean originated with Galton's (1869) law of filial regression. He noticed that the children of tall parents were, on the whole, not as tall. He called it regression to the mean. He thought that a trait showing regression of progeny mean on mid parental value must be genetically determined. Jensen (1969) discussing some results on IQ says: 'None of these findings is at all surprising from the standpoint of a genetic hypothesis, of which an intrinsic feature is Galton's "law of filial regression."' Eysenck (1975) wrote 'There is no mention (in Kamin's book, 1974), for instance, of the remarkable regression effects which provide such a striking proof of inheritance, as well as making possible estimates of heritability' (see also Eysenck 1976).

This confusion has now entered the area of familial aggregation. We give two examples. Tsuang et al. (2000) found that 'In the Huntington's disease probands with psychosis, the onset of psychosis correlated with the onset of the neurological symptoms of the Huntington's disease.' From this correlation they concluded: 'Patients with Huntington's disease and psychotic symptoms may have a familial disposition to develop psychosis.' Perusse et al. (2000) used a familial correlation model and concluded 'that familial/genetic factors are more important ...' Behavior geneticists Neale and Carden (1992, p. 99) said: 'In other words, genotypes and environments are not

measured directly but their influence is inferred through their effects on the covariances of relatives.'

It is remarkable that despite Fisher's (1924) efforts, confusion persists. Fisher, discussing the phenomenon that children of tall parents are, on average, less tall said: 'This phenomenon has been called "regression to the mean" a use of the term regression quite different from the *technical meaning of the term* explained previously' (italics added). He suggested some reasons for this (see also Vetta 1975). Statisticians know that correlation or regression, in absence of other evidence, tell us nothing about cause and effect. Fisher put it elegantly (p. 196), 'To begin with, these statistics (regressions) do not prove that daughters inherit their height from their fathers; if they prove that they will equally prove that fathers inherited their height from their daughters, which is absurd. What they prove is that the heights of fathers and daughters are influenced by the same causes and it is generally agreed on *quite other grounds*, that the important cause of this similarity is that daughters have a great deal in common with their fathers in their hereditary constitution' (italics added). Thus, familial correlations, covariances or regressions on a trait should not be used to estimate genetic contribution, that is, h^2 or H^2 unless one is absolutely certain, on quite other grounds, that the trait has a genetic component.

Some behavior geneticists use path coefficients (Wright 1921). A path coefficient is a standardized regression coefficient. We do not discuss path analysis. We do not believe that a regression coefficient with or without the help of a diagram can assist us in resolving the problem of cause and effect.

5. Development of Modern Genetics

Modern genetics was developed by Fisher, Wright, Haldane, and Dobzhansky, among others, in the early part of the twentieth century. Fisher was a brilliant geneticist, as well as a brilliant statistician. His 1918 paper *Correlation between relatives on the supposition of Mendelian inheritance* is a comprehensive paper on quantitative genetics. It is difficult to read and almost impossible to understand. It is the basis of most of the model-fitting work in behavior genetics. It is, therefore, important that one understands the assumptions he made and the consequences of their failure.

6. Fisher's (1918) Correlation Formulas are Incorrect

Vetta (1976) showed that Fisher's (1918) correlation formulas are incorrect. He gave two reasons that may have led to the error. A third reason is that Fisher's basic assumption for his model was that the contribution of terms of third and higher degree of smallness are negligible as compared with the con-

tribution of terms of second degree of smallness (which are associated with variances). This assumption is valid only under random mating and not under assortative mating. The terms of third degree of smallness do make a contribution to covariances but the contribution of the terms of fourth and higher degree of smallness is, indeed, negligible. P-c and sib correlation formulas obtained by taking the contribution of terms of third degree of smallness into account meet the criteria discussed above.

7. Fisher's Model in Behavior Genetics

Behavior genetics is a relatively young science (Hirsch 1967). One of the major problems was the existence of genotype-environment interaction for most behavioral traits. There was no method of estimating it and Jinks and Fulker (1970) devised one. Their method was hailed as a breakthrough. It was used by Jensen (1970) and was extended to the multivariate case by Eaves (1972). In 1975 Vetta found that there was an algebraic error in their paper. When this error is corrected their method is useless. This was made known immediately to Professor Jinks and his colleagues. Despite this, Fulker and Eysenck (1979) said 'We can test directly for some form of genotype-environment interaction.' This is not true. In the rest of their paper Jinks and Fulker used Fisher's (1918) *incorrect* formulas to analyze data on some behavioral traits. Their paper is probably the most cited paper in behavior genetics. Eysenck (1979) says 'This book is the first to base itself entirely on these new methods.' Some behavior geneticists regard it as 'seminal' (Martin et al. 1989, p. 5). We disagree.

Most behavioral traits involve assortative mating. Behavior geneticists use incorrect formulas when they fit assortative mating models. Devlin et al. (1997) supplement Bouchard and McGue's (1981) correlations with correlations from twin studies published after 1981. They said: 'IQ heritability is not well characterized.' On the basis of past analyses Herrnstein and Murray's (1994) believed 'that IQ heritability is at least 60 percent, and is probably closer to 80 percent ...' (p. 470). They predicted that this will give rise to dysgenic trends and a cognitive caste will emerge. Devlin et al. (1997) aimed to show that a cognitive caste will not emerge. Using incorrect formulas, they obtained estimates of 0.34 and 0.48 for narrow and broad heritability respectively. They argued that, in view of these smaller estimates, a cognitive caste will not emerge. Actually, a cognitive caste will still emerge but over a longer period of generations, if IQ is a genetic trait. McGue (1997, p. 417) said that Devlin et al.'s analysis follows 'a long line' of analyses. Behavior geneticists keep on analyzing essentially the same data and obtain estimates of heritability of IQ to suit the occasion.

Devlin et al. (1997) wrote expected covariance formulas for different kinships. Each formula had a

genetic and an environmental component. The genetic part of their covariance formulas for dizygotic twins and sibs is $1/2 (1 + r\sigma_A^*) \sigma_A^2$. The genetic part of their parent-child covariance formula is $1/2 (1 + r) \sigma_A^2$. In these formulas, σ_A^2 is the additive variance of the trait, σ_A^* is its standardized additive variance, and r is the phenotypic coefficient of assortative mating. As stated earlier, in the absence of dominance these genetic parts should be equal. Therefore

$$(1 + r\sigma_A^*) \sigma_A^2 = (1 + r) \sigma_A^2 \quad (1)$$

Simplifying, $r\sigma_A^* = r$. This gives, $\sigma_A^* = 1$. But σ_A^* is always < 1 as it is the additive part of the phenotypic variance (except for the trivial case when only additive variance is present). These formulas are not correct. Similarly, their formulas expected covariances of dizygotic twins, siblings reared together, and siblings reared apart are not correct (Table 1, p. 469). Their model is, therefore, useless.

8. Statistics and Behavior Genetics

A recently borrowed concept from statistics is 'shared environment.' McGuffin et al. (2001) said that the second striking finding of behavior genetics is that the contribution of environment 'tends to be of the nonshared type, that is, environmental factors make people different from, rather than similar to, their relatives.' What is 'nonshared' environment? Neale and Cardon (1992, p. 15) said 'The environment between families is sometimes called the *shared environment*.' Thus, within family environment it is the 'nonshared environment.' Behavior geneticists in their analyses of variance of observed data may have found that, for some traits, within-family variance $>$ between-family variance. We do not believe that this justifies their claim.

A statistician working on a designed two-factor component of variance model will have three types of data; (a) observed values, (b) levels of factor I, and (c) levels of factor II. A behavior geneticist only has observed values from an observational study. He makes up for the lack of data (b) and (c) by calculating kinships covariances. He then equates calculated covariances to expected covariances, under assortative mating. His expected covariance formulas are invariably wrong. He then makes various claims which, in our view, cannot be justified.

9. High Heritability and Change in the Population Mean of a Trait

The concept of heritability is surrounded with many misconceptions. One such misconception is: if the heritability of a trait is high then environmental changes have little or no effect on its mean. When with high H^2 , the mean value of a trait changes appreciably,

some geneticists are at a loss to explain the change. For example, see Vogel and Motulsky's (1996, p. 239) discussion of the 'recent increase in stature.' There has also been a secular trend in IQ during since the 1970s. A secular trend in presence of high heritability deserves serious consideration. Fisher (1918) showed that a polygenic trait is normally distributed. The mean and variance of a normal distribution are, however, independent, that is, the factors which affect variance do not affect mean. Therefore, H^2 , the ratio of two variances, does not affect the mean. This is true in industrial production. We do not believe it is true for a genetic trait. The factors that contribute to H^2 , that is, genes, must also affect the mean of the trait. Genes, however, do not determine secular trend. We should, therefore, consider alternate explanations, for example, genes only provide possibilities; the trait has no genetic component; all genes for the trait are 'fixed,' therefore, the variance is due to 'environmental' factors, etc.

10. Can Genetic and Environmental Effects be Separated?

A designed statistical experiment could separate effects of factors on a disorder. Designed experiments are not possible on human beings. Moreover, we shall not be able to sample from all genotypes and all environments. In the absence of designed experiments, the genetic and environmental variances *cannot* be separated. Fisher (1918) did separate the genetic and environmental variances relating to height by assuming random environment. We do not believe that this assumption is viable. It is noteworthy that Fisher never really returned to the subject of this paper. Actually, the concept of 'evolution by adaptation' implies complex relationships between genotypes and environments. The history of human evolution may be divided into two periods: the first period when Homo Sapiens tried furiously to adapt to the then existing environment and the second, when they tried to control it. We are now far into the second period. We do not know the bygone environment to which our bodies adapted. We agree with Fisher (1930) that a species is defined only *in* an environment. Change this environment a little, the species will try to adapt. Change it too much and the species may not be defined. Therefore, research aimed at finding the genetic and environmental components of human behavior has no scientific content.

11. The Future?

McGuffin et al. (2001) envisage a new science of behavior genomics. We suspect that as the unscientific nature of behavior genetic analysis becomes known,

researchers will eschew heritability analysis. HGP has made the identification of a genetic disorder easier. If, for example, a large number of individuals suffering from a disorder have mutation at a locus as compared with the normal type, this provides some evidence of the genetic nature of the disorder. Heritability analysis is useless as it relates to a population and not an individual. To find remedies for genetic disorders, type I models are useful. Venter (2001, *The Independent*, February 12) succinctly summarizes our view when he says, HGP indicates 'to me that we are not hard wired. The idea that there is a simple deterministic explanation—that is: we are the sum total of our genes—makes me as a scientist, want to laugh and cry.'

See also: Behavioral Genetics; Psychological Perspectives; Depression; Genetic Factors in Cognition/Intelligence; Genetic Studies of Behavior; Methodology; Genetic Studies of Personality; Genetics of Complex Traits Through the Life Cycle; Intelligence, Genetics of; Cognitive Abilities; Intelligence, Genetics of; Heritability and Causation; Mental Illness, Genetics of; Schizophrenia

Bibliography

- Adams R, Ghodsian M, Richardson K 1977 *Nature* **266**: 279
- Bouchard T J Jr, McGue M G 1981 Familial studies of intelligence: A review. *Science* **212**: 1055–9
- Capron C, Vetta A R, Duyme M, Vetta A 1999 Misconceptions of biometrical IQists. *Current Psychology of Cognition, Cahiers de Psychologie Cognitive* **18**: 115–60
- Carroll S B 2001 Communication breakdown? Review of *The Century of the Gene*. *Science* **291**: 1264
- Cullen M R 1983 *Mathematics for the Biosciences*. Prindle, Webber and Schmidt, Boston
- Devlin B, Daniels M, Roedre K 1997 The heritability of IQ. *Nature* **388**: 468–70
- Eaves L J 1972 Computer simulation of sample size and experimental design in human psychogenetics. *Psychological Bulletin* **77**: 144–52
- Eysenck H J 1975 Review: The science and politics of IQ. *Bulletin of the British Psychological Society* **28**: 352
- Eysenck H J 1976 Genetic determination and regression effects. *Bulletin of the British Psychological Society* **29**: 406
- Eysenck H J 1979 *The Structure and Measurement of Intelligence*. Springer-Verlag, Berlin
- Feldman M W, Lewontin R C 1975 The heritability hang up. *Science* **190**: 1163–8
- Fisher R A 1918 The correlation between relatives on the supposition of Mendelian inheritance. *Translations of the Royal Society Edinburgh* **52**: 399–433
- Fisher R A 1924 The biometrical study of heredity. *Eugenics Review* **16**: 189–210
- Fisher R A 1930 *The Genetical Theory of Natural Selection*, 2nd edn. Clarendon Press, Oxford, UK (rev. edn. 1958 Dover Publications, Broadway, UK)
- Fisher R A 1951 Limits to intensive production in animals. *British Agricultural Bulletin* **4**: 217–8

- Fulker D W, Eysenck H J 1979 Nature and nurture: Heredity. In: Eysenck H J (ed.) *The Structure and Measurement of Intelligence*. Springer, Berlin, pp. 102–32
- Galton F 1869 *Hereditary Genius: An Inquiry into its Laws and Consequences*. Macmillan, London
- Hartl D L, Clark A G 1997 *Principles of Population Genetics*. Sinauer Associates, Canada
- Herrnstein R J, Murray C 1994 *The Bell Curve*. Free Press, London
- Hirsch J 1967 *Behavior-genetic Analysis*. McGraw-Hill, New York
- Holzinger K J 1929 The relative effect of nature and nurture on twin differences. *Journal of Educational Psychology* **20**: 241–8
- Jensen A 1967 Estimation of the limits of heritability of traits by comparison of monozygotic and dizygotic twins. *Proceedings of the National Academy of Science of the United States* **58**: 115–74
- Jensen A 1969 How much can we boost IQ and scholastic achievement? *Harvard Educational Review* **39**: 1–123
- Jensen A 1970 IQs of identical twins reared apart. *Behavior Genetics* **1**: 133–48
- Jinks J L, Fulker D W 1970 Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behavior. *Psychological Bulletin* **73**: 311–49
- Johannsen W 1903 *Über Erblichkeit in Populationen und in Reinen Linien*. Gustav Fischer, Jena (summary and conclusions trans. in Peters J A 1959 *Classic Papers in Genetics*. Prentice-Hall, Englewood Cliffs, NJ)
- Kamin L 1974 *The Science and Politics of IQ*. Erlbaum, Hillsdale, NJ
- Kempthorne O 1969 *An Introduction to Genetic Statistics*. Wiley, New York
- Lawrence M J, Jinks J L 1973 Quantitative genetics. In: Sheppard P M (ed.) *Practical Genetics*. Blackwell Scientific, Oxford, UK
- Leviton M 1988 *Textbook of Human Genetics*. Oxford University Press, Oxford, UK
- Li C C 1955 *Human Genetics*. McGraw-Hill, New York
- Martin N G, Boomsman D I, Neale M C 1989 Foreword. *Behavior Genetics* **1**: 5–7
- Mather K 1949 *Biometrical Genetics: The Study of Continuous Variation*. Dover Publications, New York
- McGue M 1997 The democracy of the genes. *Nature* **388**: 417–8
- McGuffin P, Brien R, Plomin R 2001 Toward behavioral genomics. *Science* **291**: 1232
- Neale M C, Cardon L R 1992 *Methodology for Genetic Studies of Twins and Families*. Kluwer, Dordrecht, The Netherlands
- Nichols R C 1965 The National Merit twin study. In: Vandenberg G S (ed.) *Methods and Goals in Human Behavior Genetics*. Academic Press, New York
- Perusse L, Rice T, Province M A, Gagnon J, Leon A S, Skinner J S, Wilmore J H, Rao D C, Bouchard C 2000 Familial aggregation of amount and distribution fat and their responses to exercise training in the HERITAGE family study. *Obesity Research* **8**: 140–50
- Roach G F (ed.) 1984 *Mathematics in Medicine and Biomechanics*. Shive Publications, Cheshire
- Snustad D P, Simmons M J, Jenkins J B 1997 *Principles of Genetics*. Wiley, New York
- Thoday J M 1973 Review of educability and group differences. *Nature* **245**: 218–20
- Tsuang D, Almquist E W, Lipe H, Strgar F, DiGiacomo L, Hoff D, Eugenio C, Hayden M R, Bird T D 2000 Familial aggregation of psychotic symptoms in Huntington's disease. *American Journal of Psychiatry* **157**: 1955–9

- Vetta A 1975 A note on regression to the mean. *Social Biology* **22**: 86–8
- Vetta A 1976 Correction to Fisher's correlations between relatives and environmental effects. *Nature* **263**: 316–7
- Vetta A 1977 Estimation of heritability from IQ data on twins. *Nature* **266**: 279
- Vetta A, Capron C 1999 Review of *How the Mind Works* by Steven Pinker. *Cahiers de Psychologie Cognitive* **18**: 105–11
- Vogel F, Motulsky A G 1996 *Human Genetics*. Springer, Berlin
- Wright S 1921 Systems of mating. I–V. *Genetics* **6**: 111–23

A. Vetta and C. Capron

Families and Households, Behavioral Demography of

The last half of the twentieth century has seen the study of household and family develop into a major subspeciality in demography, rivaling in scope and interest more traditional subfields, such as mortality–morbidity, fertility, migration, and urbanization. The study of households and families introduces a new order of complexity into demography since (a) the focus is now on a set of interrelated individuals rather than on the isolated individual or on a population aggregate described by summary measures, and (b) the balancing equations for these residential and kinship units involve several events other than those in the fundamental demographic equation (population change = births – deaths + in-migration – out-migration). Changes in households and families must be related also to marriage, cohabitation, divorce, separation, home-leaving of young adults, as well as other forms of exit and entrance to the group (e.g., the entry of an elderly parent or some other relative). The household may be viewed as a resultant of many individual behaviors.

There are more fundamental complications relating to household and family identity: it is clear when a particular individual begins (birth) and ceases to exist (death), and there is individual continuity over the lifetime. By contrast, it often is not at all clear when a particular household or family begins or ends. Households, unlike individuals, can split apart into two or more smaller households (fission), or can merge to form larger and more complex units (fusion). This fundamental problem has been not solved so much as finessed by extensive data-collection and analysis of *individuals in terms of their family and household statuses and the events changing their statuses*.

But progress has been considerable, particularly with the advent of large-scale social surveys containing event and status history data on marriage, family, and household (individual event-histories are much more common than histories of households or families).

These data have allowed the description and analysis of individual family life-histories and changing household statuses, based on real cohort data. These newer studies stand in contrast to the synthetic approach of the original 'family life-cycle' studies, based on census data (cross-sectional) and sometimes misleading. Recent work has documented a proliferation of family and household behavioral sequences, overturning the stereotypical view of the 'family life-cycle,' based on averages or medians (a development anticipated by Höhn 1987).

Another important breakthrough is the work of Zeng et al. (1997), who have developed accounting equations for a number of specific family household types, equations that can be applied using routine census and vital registration data. The model produces detailed household projections by size and type of household that are consistent with population projections by age, sex, and marital status.

Review articles and summary volumes on the demographic approach to households include: Burch (1979), Bongaarts (1983), Bongaarts et al. (1987), Keilman et al. (1988) and van Imhoff et al. (1995). Work in behavioral household demography, of course, has not occurred in disciplinary isolation, and should be viewed in the context of parallel or overlapping efforts in family sociology, household economics, anthropology, and demographic history.

1. The Definition of 'Household'

The central idea of the household concept is that of *coresidence*: a household is a group of persons who 'live-together' day-to-day, or an individual 'living alone.' In demographic usage, the operational definition of this concept has tended to follow census practice, with households defined in terms of persons occupying a separate housing or dwelling unit (house or apartment), or, less often, the persons who participate in a common domestic economy. The housing unit approach provided a convenient basis for census enumeration (first identify housing units and then enumerate the people usually residing in each), and later, for modern large-scale sample surveys.

1.1 Households and Families

Whereas the *household* is defined in terms of *coresidence*, the *family* is defined in terms of *kinship*, which in turn is based on marriage, descent from a common ancestor, or social fiction (adoption or other forms of 'fictive kinship'). Ryder (1987) has aptly termed these, respectively, the residential, the conjugal, and the consanguineal dimensions of intimate relationships among individuals (see also van Imhoff

et al. 1995). This distinction is not always made in everyday language, however, and the concepts and reality overlap: a household consisting of persons related by kinship is also a family, but such a household seldom contains all of an individual's close kin, the so-called 'extended family.' Modern census practice often used the term *family* to refer to a group of related persons living together (more properly called a *census family*), a usage encouraged by and reinforcing a Western cultural tendency to view the small, nuclear family group as 'the family'—other kin are 'relatives.' Historical and comparative studies of household and family have forced a clarification and refinement of census concepts to take into account the greater complexity of household and family patterns, as have more recent developments such as cohabitation or the joint custody (and split coresidence) of children following divorce. In general, the single word *family* is ambiguous, and should always be related to context, and qualified as necessary with such terms as *nuclear*, *extended*, *census*, *economic*, etc.

1.2 Large Households

The majority of households are relatively small, consisting of groups such as a couple and their children, a small group of students sharing an apartment, or an individual living alone. Such households often are referred to as *private households* in contrast to *institutional households* such as boarding schools, prisons, chronic-care hospitals, mental institutions, convents, or seminaries. There is no inherent limit to the size of private households, but their average size seldom exceeds six, and in urban-industrial societies is closer to three.

In modern census practice, private households can contain boarders, lodgers, or servants, but an arbitrary numerical limit on such members is set, beyond which the household is considered institutional.

1.3 Limitations and Criticisms

The modern household concept does not deal easily with social realities such as residential *compounds*, clusterings of separate but closely related (in terms of kinship or economic activity) households, often physically distinct, occupying a separate farm or walled or fenced space. An example from a polygamous society would be the compound containing a man and his several wives and children, the latter occupying separate huts and eating separately. Laslett (1982) has suggested the term *houseful* for European cases where several smaller nuclear families (some related, some agricultural workers or servants) occupied a well-defined farmstead.

In contemporary urban societies, census practice reveals virtually nothing about relationships among separate households even when such relationships are frequent and close. For example, an older couple and two grown children, all occupying separate apartments in the same neighborhood, or even the same building, would be reckoned simply as three households, with no census item even hinting at their close ties. Relationships are counted within but not across households. It is of interest that this census practice was codified around the middle of the twentieth century, at a time when social scientists, especially family sociologists, were emphasizing (in retrospect exaggerating) the isolation of the nuclear family in modern urban societies. These limitations of the census have been rectified in part by large social surveys including information on numbers and types of kin, on kin location, and on kin interaction.

A related problem has been noted with use of the phrase *living alone* to refer to one-person households. Accurate enough in terms of census housing-unit concepts, its connotations can be misleading with respect to the degree of social integration or isolation of the occupants.

These and other concerns have led some to more or less reject the household concept as ethnocentric or culture-bound, and too limited, artificial, and abstract. But an increasingly popular view of scientific methodology (see, e.g., Giere 1999) reminds us that all scientific concepts (models, theories) are abstract—therefore, limited and artificial—and are better judged in terms of their analytic usefulness for the study of particular problems. The household clearly is the relevant unit for much consumer behavior (from buying a house or renting an apartment to routine purchases of food or clothing), demographic decision making (e.g., migration), and much of our day-to-day intimate social interaction and support, including the intimacy of couples and the continuing care of infants and young children. Thus, housing analysts and planners still are major users of household projections, as opposed to ordinary population projections. And the household remains a major focus for family sociology and for microeconomics (notably the Chicago school of the ‘new household economics’), even as they have broadened their view to encompass such economic realities as intergenerational transfers and such psychological realities as ‘altruism.’

Even so, the household concept needs refinement and extension to take more account of the fluidity of household membership and its ‘fuzziness,’ as more persons can realistically claim membership in two households: children in joint custody arrangements, ‘commuter marriages,’ Caribbean-style ‘visiting relationships,’ and ownership of second homes. Residence will increasingly be seen as a multidimensional concept, with different households fulfilling different household functions for the same individual. The census convention of a one-to-one correspondence

between individuals and households will not be abandoned entirely, but formal and behavioral demographic analysis will of necessity go beyond it.

2. Late Twentieth-century Trends and their Demographic Determinants

The generalization is sometimes advanced that twentieth-century households have tended to become smaller and less complex. The generalization regarding size almost certainly is valid, that regarding complexity less clearly so. Households have tended to become smaller in large measure because on average they contain fewer minor children, consequent to widespread fertility decline. If the total fertility rate is two or below, the typical nuclear family household (husband, wife and own children) will contain three or four members or fewer. Only populations that retain traditional high fertility levels, and perhaps other features of traditional family structure, still report average household sizes of five or six persons and above.

Other factors are at work. In highly urban-industrial populations, smaller average households also have resulted from such factors as the rise of independent living of women (lone mothers, earlier home-leaving of women) and the elderly (widows and widowers), and the splitting of households following divorce. Also relevant in some societies has been the decline of live-in servants and of boarders or roomers. But these latter trends, unlike fertility decline, differ considerably across societies, depending on cultural traditions relating to coresidence with nonnuclear kin. In Japan, for example, the norm requiring taking in elderly parents remains strong. In Italy, adult children tend to remain in the parental home until marriage. Paradoxically, these familistic patterns are facilitated by low fertility, with a small nuclear family leaving space for young adults and elderly who might otherwise be crowded out. Mitchell et al. (1989), for example, show that in Canada number of siblings is an important determinant of age at leaving the parental home. These examples underline the point that household-formation behaviors feed back on one another.

In many western societies, rising rates of divorce and of nonmarital cohabitation have led to considerable fluidity in household formation, particularly among young adults and among older persons following divorce or widow(er)hood. Looking to the future, one might anticipate even greater flexibility in household arrangements, including among the elderly, as individuals develop broader repertoires of behavior based on early experience as children and young adults.

The older stereotype identifying large households with complex households of extended kin must give way to the view that smaller households can also be more rather than less complex, given greater degrees

of freedom to add members due to reduced domestic demands of smaller numbers of minor children. Like many other demographic behaviors, household formation and dissolution now show more heterogeneity in most dimensions except size.

In most societies, coresidence of kin has been lessened by high rates of migration, and the consequent location of kin in different, sometimes distant, communities. In urban-industrial societies, this includes urban-to-urban labor migration as well as retirement-related migration of the elderly. In developing societies, a major factor is rural-urban migration and, in some cases, high rates of overseas migration. Kinship contacts persist, but they are attenuated by distance.

3. Behavioral Theories of Household Dynamics

As with other demographic behaviors, explanation of household dynamics must be partly formal, partly behavioral. But unlike other demographic behaviors, accession to or relinquishment of a household status is not a unitary behavior. In traditional terms, an individual may become a household 'head' by setting up his or her own household, by being widowed or divorced, by the departure of one's last child (if already widowed or divorced), and so forth. It is this fact that has made the fitting of so-called 'household headship' curves relatively intractable: they are the resultant of too many different behaviours (Corner 1987). A theory of household formation must, therefore, in some sense be a theory of marriage, divorce, fertility, mortality, home-leaving, etc.—all combined in some meaningful way.

3.1 Microeconomic Theories of Household Formation

In this field, as in demography generally, economics has had considerable theoretical influence, notably from the works of the 'Chicago School' and from empirical studies of household formation, especially the rise of one-person households. Becker (1981) viewed marriage and household formation in terms of a traditional sexual division of labor, with husband and wife coming together for practical economic advantage in a 'household economy.' Women tended to remain in the home performing domestic tasks, while men entered the labor market for wages, both pursuing comparative advantages associated with differential earning potential. Davis (1984) called this arrangement the 'breadwinner' system, and associates it with the Victorian era. Prior to that, on family farms or in small enterprises, both husband and wife worked to support the family economically.

Clearly, such a system no longer prevails in Western urban-industrial societies, as women have entered the

labor market in force, pursuing careers throughout their adult lifetimes. In some contexts and especially among young adults, women often appear to have the advantage in terms of employment and income, as work requiring physical strength has declined in importance. Women are not as dependent as formerly on men for their economic well-being, and are more able to establish their own households. At the same time, the costs of maintaining desired levels of consumption have risen, so that having more than one wage-earner per household is advantageous. Thus, both members of a couple are apt to work, bringing in more income but foregoing the real economic advantages of Becker's household division of labor, even if not along traditional lines. There still are economic rationales for marriage and coresidence, but they are somewhat different than in earlier times.

A central notion in economic analysis of the household is that there are economies of scale to be gained by coresidence—in the provision of housing, utilities (heat, lighting, etc.), food and meal preparation, and consumer durables such as stoves, refrigerators, or automobiles. A *leitmotiv* in the economic literature on households is that rising levels of living have led to the purchase of greater privacy and independence, with a sacrifice of economies of scale associated with the sharing of living quarters.

In their influential article (*Privacy, Poverty and Old Age*, Beresford & Rivlin (1966)) began a research tradition of empirical research on 'living alone' in one-person households. The aim was to explain the rise in the 'propensity' to live alone, that is, the rise in age-sex-marital status specific rates of living alone, notably among older women and young adults of both sexes. The central explanatory idea was that high real income was being used to purchase more household 'privacy,' viewed as a luxury in the technical economic sense, but a luxury that had always been desired. In keeping with the tenor of economic thinking of the times, any possible shift in 'tastes' or values relating to household arrangements was played down.

Further research tended to qualify a pure 'income' explanation. Clearly, individuals would not live alone if they literally couldn't afford it. But older women, for example, lived alone in greater numbers partly due to new social programs (social security, government pensions, old-age assistance), created in response to demand or felt need, based in turn on new ideas about individual independence. The larger explanation was social and political as well as economic. Similarly, among young adults, the pursuit of independence from parents often has been intense, with youth leaving affluent homes to live in relative squalor, a pattern documented in path-breaking research based on the Survey of Labor and Income Dynamics (SLID), one of the first major longitudinal surveys (Hill & Hill 1976).

Many of these household dynamics can be related to a general breakdown of traditional patterns of age-sex

stratification (the 'decline of the patriarchy'), in the family as well as society at large. Women, older persons, children, and young adults have sought and been accorded rights and perquisites formerly reserved to the middle-aged male. This has led to greater competition *within* households for scarce household goods such as personal space, quiet, privacy, and freedom of movement (Burch 1985).

In the West, these developments clearly are bound up with a strong cultural strain towards individualism and with its full expression in situations of affluence. Whether others societies will follow is unclear (McDonald 1992). Similarly, in the absence of affluence individualism may be sacrificed, as coresidence is sought for economic reasons—the *estrategias de sobrevivencia* emphasised by the Latin American social demographic school.

In addition to economic and cultural factors, early research by Sweet (1977) and by Kobrin (1976) focused attention on a demographic supply factor in the choice of household statuses: to the extent that coresidence continued to be seen as a family matter, even an act of relative intimacy, one's options for coresidence were limited by the number and availability of suitable kin. Sweet succinctly noted that an older widow cannot live with a child if she has no living children.

3.2 Demographic Theories of Household Formation

Ermisch (1988) has developed and empirically tested what might be called a two-part model of household formation, one in which 'demographic influences on household formation and composition can be separated from their economic influences ...' (p. 24). In this approach, demographic events of marriage, birth, death, and home-leaving lead to the formation of 'minimal household units' (childless unmarried adults, lone parents with children, childless married couples, married couples with children). These 'minimal household units' then distribute themselves into distinct household arrangements in response to socioeconomic factors. The model is not meant to imply that household formation literally occurs in two temporally ordered stages, but is seen only as an analytic convenience.

Burch & Matthews (1987) have proposed a decision-making model that is nearly the opposite of Ermisch's—with couples or individuals giving priority to a choice of household living arrangement, and then deciding on demographic behaviors (marriage, fertility, etc.) accordingly.

Only detailed descriptions of decision-making processes over individual life courses would clarify the order of demographic and household decision-making in reality. Until now, adequate survey data are limited to events, not the decision processes that underlie them. And, of course, some changes in household status are not chosen by the individual: spouses or

partners die or separate; grown-up children leave home; circumstances dictate taking in an elderly parent. The question then is one of inertia v. choice of a different arrangement.

Whether an integrated theory of household dynamics is possible is a moot question. But some contemporary philosophers of science (Giere 1999) would urge us not to worry on that score. Theory is seen as consisting of a large collection of theoretical models pertaining to the household, some useful for some analyses, others useful for others—theory as a toolkit, rather than as 'truth.' And the household dynamics literature is now rich with such tools.

3.3 Household Formation and 'Social Capital'

Since the 1960s, theories of demographic behavior have been subject to strong—some would say excessive—influences from neoclassical microeconomics, with various demographic choices viewed as resulting from rational and forward-looking individual assessments of costs and benefits ('lifetime utilities'). Sociologists and psychologists, among others, have objected that the economic perspective provides a narrow and unbalanced view of human behavior and social structure, and have tried to extend the core microeconomic model to deal with human realities such as emotions, social norms, and social and cultural systems (leading economists of demographic behavior also attempted such broadening, as in Becker's concern with altruism, or Easterlin's with 'psychic costs' of fertility control). A recent attempt has been the introduction into the study of fertility of concepts of 'social interaction,' 'social influence,' and 'diffusion' (Bongaarts & Watkins 1996, Montgomery & Casterline 1996).

But there have been few efforts (and no successes) at constructing a theoretical scheme that does not retain microeconomic ideas at its core—perhaps a reflection that when all is said and done, a good deal of human behavior and social structure is at base purposive and rational. Clearly this is the case with households, which serve a number of purely utilitarian functions: provision of shelter, heat, physical security, sleep; storage of personal property; and the purchase, preparation, and consumption of food. But households also meet nonmaterial needs, as the locus of various forms of companionship and intimacy, social support, and the formation of 'social capital.'

A programmatic article on this concept by Astone et al. (1999) seems destined to shape behavioral research in household demography for some time to come. These authors call for a reorientation of family and household demography around the concept of *social capital*—defined as resources of an individual that derive from his or her social ties (p. 2)—with that concept reinserted into the context of social exchange theory, from which it emerged. 'The formation of sexual partnerships, the birth and rearing of children, and both intra-generational and inter-generational

transfers constitute major forms of investment in social capital in virtually all societies' (Astone et al. 1999, p. 203). Clearly, their concern is with both households and families (including extended kinship networks) as these have been defined earlier.

This perspective raises a number of interesting questions regarding contemporary household patterns in urban-industrial societies' in general whether they have moved too far in the direction of solitary living—what Kuznets (1978) referred to as 'the apartness of adults,' with associated sacrifices of both economic and social-psychological benefits of coresidence. More specific questions might include the following: Do persons in one-person households suffer from a lack of enduring social ties, or are the four walls of a house or apartment relatively unimportant as social boundaries? Do children suffer from growing up in households with only two adults, as compared with those containing aunts or uncles, or grandparents—do they grow up with a more limited repertoire of behaviors; or, are the limitations of the household offset by increased social interaction with adults at 'daycare.' Are households in which one member of a couple (whether a woman or a man) remains at home performing 'domestic' tasks better off in terms of 'social capital' formation even if household income is lower?

The study of these and other questions pertaining to household and family structure might be strengthened by focus on issues widely viewed as 'social problems.' That is, the development of household and family demography might benefit not just from theoretical reorientation as suggested by Astone et al. (1999) but also by greater attention to real-world social problems.

See also: Family and Kinship, History of; Family Size Preferences; Family Systems and the Preferred Sex of Children; Family Theory and the Realities of Childbearing Behavior; Family Theory: Competing Perspectives in Social Demography; Family Theory: Complementarity of Economic and Social Explanations; Family Theory: Economics of Intergenerational Relations; Family Theory: Role of Changing Values; Historical Demography

Bibliography

- Angel R, Tienda M 1982 Determinants of extended household structure. *American Journal of Sociology* **87**: 1360–83
- Astone N M, Nathanson C A, Schoen R, Kim Y J 1999 Family demography, social theory and investment in social capital. *Population and Development Review* **25**: 1–32
- Becker G 1981 *A Treatise on the Family*. University of Chicago Press, Chicago
- Beresford J C, Rivlin A M 1966 Privacy, poverty, and old age. *Demography* **3**: 247–58
- Bongaarts J 1983 The formal demography of families and households: An overview. *IUSSP Newsletter* **17**: 27–42

- Bongaarts J, Burch T K, Wachter K W (eds.) 1987 *Family Demography: Methods and their Application*. Clarendon, Oxford, UK
- Bongaarts J, Watkins S C 1996 Social interactions and contemporary fertility transitions. *Population and Development Review* **22**: 639–82
- Burch T K 1979 Household and family demography: A bibliographic essay. *Population Index* **45**: 173–95
- Burch T K 1985 Changing age-sex roles and household crowding: A theoretical note. In: *Proceedings, International Population Conference, Florence, 1985*, International Union for the Scientific Study of Population, Liège, Vol. 3, pp. 253–62
- Burch T K 1995 Theories of household formation: progress and challenges. In: van Imhoff E, Kuijsten A C, Hooimeijer P, van Wissen L J C (eds.) *Household Demography and Household Modelling*. Plenum Press, New York, pp. 85–108
- Burch T K, Matthews B J 1987 Household formation in developed societies. *Population and Development Review* **13**: 495–511
- Corner I 1987 Household projection methods. *Journal of Forecasting* **6**: 271–84
- Davis K 1984 Wives and work: the sex role revolution and its consequences. *Population and Development Review* **10**: 397–418
- Ermisch J 1988 An economic perspective on household modeling. In: Keilman N, Kuijsten A C, Vossen A (eds.) *Modelling Household Formation and Dissolution*. Clarendon Press, Oxford, pp. 23–40
- Giere R N 1999 *Science Without Laws*. University of Chicago Press, Chicago
- Hill D, Hill M 1976 Older children and splitting off. In: Duncan G J, Morgan J N (eds.) *Five Thousand American Families—Patterns of Economic Progress, Vol. IV, Family Composition Change*. Institute for Social Research, Ann Arbor, pp. 117–54
- Höhn C 1987 The family life cycle: needed extensions of the concept. In: Bongaarts J, Burch T K, Wachter K W (eds.) *Family Demography: Methods and their Application*. Clarendon, Oxford, UK, pp. 65–80
- Keilman N, Kuijsten A C, Vossen A (eds.) 1988 *Modelling Household Formation and Dissolution*. Clarendon Press, Oxford, UK
- Kobrin F 1976 The fall in household size and the rise of the primary individual in the United States. *Demography* **13**: 127–38
- Kuznets S 1978 Size and age structure of family households: exploratory comparisons. *Population and Development Review* **4**: 187–223
- Laslett P 1982 Introduction. In: Laslett P, Wall R (eds.) *Household and Family in Past Time*. Cambridge University Press, Cambridge, UK, pp. 1–90
- McDonald P 1992 Convergence or compromise in historical family change. In: Berquó E, Xenos P (eds.) *Family Systems and Cultural Change*. Clarendon Press, Oxford, pp. 15–30
- Mitchell B, Wister A, Burch T K 1989 The family environment and leaving the family home. *Journal of Marriage and the Family* **51**: 605–13
- Montgomery M R, Casterline J B 1996 Social learning, social influence, and new models of fertility. In: Casterline J B, Lee R D and Foote K A (eds.) *Fertility in the United States: New Patterns, New Theories*. Supplement to *Population and Development Review* **22** (S): 151–75
- Ryder N 1987 Discussion. In: Bongaarts J, Burch T K, Wachter K W (eds.) *Family Demography: Methods and their Application*. Clarendon Press, Oxford, UK, pp. 345–66

- Sweet J A 1977 Demography and the family. *Annual Review of Sociology* 3: 363–405
- van Imhoff E, Kuijsten A C, Hooimeijer P, van Wissen L J C (eds.) 1995 *Household Demography and Household Modelling*. Plenum Press, New York
- Zeng Y, Vaupel J, Wang Z 1997 A multidimensional model for projecting family households—with an illustrative numerical application. *Mathematical Population Studies* 6: 187–216

T. K. Burch

Families and Households, Formal Demography of

‘Family and household demography’ differs from traditional demography in that it explicitly recognizes and studies *relationships between individuals*. In traditional demography, the core processes of birth, death, marriage, divorce, and migration are typically studied as occurring to individuals in isolation. In contrast, family and household demography studies how these processes occur to multiple persons interacting with each other. In doing so, additional relevant processes like entering or leaving co-residence are also studied. In short, relationships between individuals are the central focus of family and household demography. This article outlines the main subfields within the broad area of family and household demography, as well as the most important formal concepts.

1. Classification by Subject Matter

Three types of relationships between individuals can be distinguished (Ryder 1987):

- (a) *Consanguineal* relationships run via the fundamental parent–child relation. Any two persons sharing a common ancestor, i.e., linked by a sequence of birth events, are said to be related by blood.
- (b) *Co-residence* relationships are between persons living together. A foolproof definition of co-residence does not exist (think, for instance, of student housing, lodgers, or part-time households), but ‘sharing roof and manger’ is not bad for a start.
- (c) *Conjugal* relationships are between individuals who are married to each other. In fact, living together in marriage is a special form of co-residence. To many people, the bond of marriage is so special that it is (almost or fully) put on the same footing as the consanguineal bonds mentioned under (a). In modern societies, the traditional marriage is gradually replaced by cohabitation (consensual unions). In these societies, ‘being married’ is more and more replaced by the more neutral ‘living together as a couple,’ which includes both formal marriages and consensual unions, and frequently also includes homosexual cohabitation. The United Nations. Statistical Division (1998) ex-

plicitly recommends: ‘Couples living in consensual unions should be regarded as married couples’ (p. 67).

The broad field of family and household demography studies all these links between individuals, in their own right but also in relation to other demographic processes (e.g., migration of household, mortality patterns in kin groups). Within the broad field, several subfields can be identified, according to the type of relationship on which they concentrate:

- (a) *Demography of kinship* studies links of type (a) and (c).
- (b) *Demography of households* studies links of type (b), but also of type (c) since ‘living together as a couple’ is a special form of ‘living together’ (see also *Families and Households, Behavioral Demography of*).
- (c) *Demography of the family* studies the simultaneous links of type (a) and (c), on the one hand, and of type (b), on the other: a family is a group of co-resident kin.
- (d) *Demography of marriage* studies links of type (c).

2. Data Issues

The data sources for household and family demography are the same as for demography in general: censuses, population registers, surveys (in kinship studies, genealogies are also sometimes used). However, *definitional problems* are particularly strong in household and family data.

There are two sets of definitional problems. The first relates to criteria for establishing *whether or not a certain relationship exists* between two persons. Examples of such problems include:

- (a) Two persons (e.g., a mother and a student child) live together during weekends only. Do they belong to the same household?
- (b) Two persons form a household. How to judge whether they ‘live together as a couple’?
- (c) Two students live in an apartment, occupying separate rooms but sharing kitchen and bathroom. Do they count as one or as two households?

The second group of definitional problems relates to criteria for *classifying households and/or family into categories*, like those used in published household statistics. Examples include:

- (a) A woman lives with her 30-year old son. Should this household be termed a lone-parent household?
- (b) A lone mother cohabits with a new partner. Is this a ‘couple with children’ household, or rather a ‘one-parent family plus other(s)’?

Such definitional degrees of freedom result in actual definitional differences, the consequences of which are potentially very large (Keilman 1995). Therefore, demographers must be well aware of the precise definitions underlying their household and family data. In descriptive studies, definitional differences can seriously distort international or intertemporal

comparisons. In modeling exercises, it is crucial that model parameters are derived from data that are definitionally consistent with the variables in the model. In order to stimulate international comparability of household and family data, the United Nations has specified detailed, extremely useful recommendations for data collection, definition, and tabulation in population and household censuses (United Nations, Statistics Division 1998).

3. Unit of Analysis

A major problem in the formal demography of families and households is whether the group or the individual should be taken as the primary unit of analysis. Almost by definition, the behavior of the individuals within the group are highly interdependent. Therefore, analyzing the behavior of the group cannot be done without proper account of the characteristics of its members; conversely, analyzing the behavior of the individual cannot be done without proper account of the characteristics of the group (including those of the other individuals in the group).

Today, most family and household demographers agree that it is most convenient to work at the level of *individuals*. The main reason for this is that, unlike individuals, groups do not have a clearly defined origin and destination in time since members can enter and leave all the time. Thus, modern practice is to define a group in terms of its members and the relationships between them, rather than defining individuals in terms of the group to which they belong.

Because of strong interdependence of within-group behavior, proper modeling of individual family or household members' behavior requires sufficient information on the characteristics of the other individuals involved. In microlevel data, this implies the inclusion of *pointers* or reference variables to other individuals in the dataset, with sufficient detail on the type of the relationship to that other individual. A main advantage of such data is that it can provide very rich descriptive information on different family and household types, as opposed to standard census breakdowns by just a few main categories.

Naturally, when using sample survey data, proper account should be given to the representativeness of the data at different units of analysis. For example, a random sample of individuals is *not* representative at the household level, since large households will be over-represented. Weighting schemes must be used to correct this and other potential distortions.

4. Models in Family and Household Demography

A model is a simplified, quantitative description of reality. In family and household demography, models provide a formal description of the quantitative relationships between family or household charac-

teristics and their demographic determinants and consequences (Bongaarts 1983).

At the heart of any modeling exercise lies the specification of the *state space*: the representation of the components of the system of interest. At the level of the *individual*, the state space consists of a number of characteristics or attributes (e.g., age, sex, position in the household, age of spouse, number of children), each of which can take a certain value. At the population level, the state space consists of all possible combinations of attribute values: it is a breakdown of all individuals comprising the population by relevant characteristics.

The ultimate purpose of model building is to obtain an understanding about this multidimensional breakdown of the total population. A classical demographer is satisfied with a breakdown by age and sex. A marriage demographer would want to add at least 'marital status' as an additional variable. A kinship demographer would want to add a complete breakdown by number of living children, grandchildren, siblings, parents, and so on. A household demographer would want to have, next to age and sex, variables like household size and household type, in a classification that is sufficiently detailed for the research or policy problem at hand. All these demographers have in common that they want to understand why an observed population breakdown is how it is; how particular changes in behavior (for example, total fertility level) would change this breakdown; or how the currently observed breakdown will develop over time (forecasting).

The particular problem in household and family models is that the relevant state space is potentially very large and complicated. This has forced family and household demographers to sacrifice all kinds of descriptive, behavioral, and/or conceptual detail. Evidently, the result is that existing household and family models reflect a wide range of different modeling strategies. Three important classes of these strategies are discussed in the sections below.

4.1 Severing the Link

As said before, family and household demography is all about relationships between individuals. A particularly important type of 'sacrifice' frequently made by family and household demographers is to discard most of the information on the characteristics of the individuals *with whom* a particular relationship exists. For example, in the case of a marriage model, a woman aged 40 and married to a man aged 42 is classified as being a married woman aged 40: the information about the age of the husband is discarded. Without discarding, the total population would require a table of $2 \times 100 \times (1 + 100) = 20,200$ cells (2 sexes, 100 age groups, 1 for nonmarried, 100 for married by age of spouse); with discarding, a table of $2 \times 2 \times 100 = 400$ cells is sufficient (2 sexes, 2 for

nonmarried vs. married, 100 age groups). In other words, a frequently used modeling strategy is to collapse all information on all individuals to whom person x has relevant relationships into just a few attributes of person x , thus in fact ‘severing’ the explicit link to those related individuals.

An example of this ‘severing the link’ strategy is provided by marital-status models of the multistate type (see *Multistate Transition Models in Demography*). Here, individuals are classified by sex s (m for men, f for women), age x , and marital status i , where i can take the following values: N for never married, M for married, D for divorced, W for widowed. The development over time of the population is described in terms of *events*, which are changes in the attributes of the individual members of the population. For example, next year’s 31-year old married female population is given by (ignoring migration):

$$\begin{aligned} P(f, 31, M, t+1) = & P(f, 30, M, t) \\ & + E(f, 30, N, M, t) + E(f, 30, D, M, t) \\ & + E(f, 30, W, M, t) - E(f, 30, M, D, t) \\ & - E(f, 30, M, W, t) - E(f, 30, M, \text{dead}, t) \end{aligned}$$

where $P(s, x, i, t)$ denotes the population at the start of year t of sex s , age x , in marital status i , and $E(s, x, i, j, t)$ denotes the number of individuals of sex s and age x who during year t experience an event from position i to position j . The various numbers of events depend on the size of the relevant population at risk of experiencing the event, and on the rate at which individuals in the population at risk experience the event. For example, the number of women who marry for the first time at age 30 in year t , i.e., $E(f, 30, N, M, t)$, equals the product of two factors: the number of never-married 30-year old women $P(f, 30, N, t)$, and the first-marriage rate for 30-year old women $M(f, 30, N, M, t)$.

Given a full set of age-, sex-, origin-, and destination-specific rates, the population by age, sex, and marital status can be projected forward in age and time. Also, a marital-status life table can be calculated. However, a crucial characteristic of this model is that marital-status events for one sex are essentially independent of the demographic situation for the other sex. As a result, in this model it is not guaranteed that the resulting population flows for men and women will be mutually consistent. For example, the number of married men having a divorce should equal the number of married women having a divorce, but there is no mechanism in the model to ensure this equality. This problem is known as the *two-sex problem* in nuptiality models (Keilman 1985a). It is a direct consequence of the fact that the explicit link between marriage partners is missing. There are several solutions to the two-sex problem, each of which introduces fresh problems:

(a) To restore the link between marriage partners. This greatly expands the state space, since married

persons must be simultaneously classified by own age *and* by the age of their spouse. It also greatly complicates the modeling of numbers of events, since rates must be defined for all combinations of male *and* female ages and marital statuses (see also *Population Dynamics: Two-sex Demographic Models*).

(b) A slightly different way of restoring the link between marriage partners is to recast the model in terms of *microsimulation*. See Sect. 4.2.

(c) To concentrate on one sex (usually: women) only. This is convenient, but obviously unrealistic.

(d) To introduce an adjustment mechanism, by which aggregate numbers of events (i.e., aggregated over all ages) are forced to satisfy certain requirements. For example, if the model initially produces 1,000 male divorces and 800 female divorces, male divorces are adjusted downward and female divorces upward to result in, say, 900 divorces for each sex. Such procedures are called *consistency algorithms* (Keilman 1985b, Van Imhoff 1992). Although such algorithms can be given a satisfactory behavioral interpretation, they nevertheless have an artificial ring and are certainly rough approximations only.

(e) A combination of the previous two ‘solutions’ leads to consistency algorithms that are one-sex dominant. For example, in a female-dominant marital status model, the total number of male divorces is determined by the female married population and female divorce rates.

Marital status models are by no means the only models in which ‘severing the link’ leads to consistency problems. All household and family models which are of the multistate type suffer from essentially the same problem.

4.2 Macro- vs. Micromodels

In principle, an algebraic representation of a model can be manipulated to study the model’s properties. If the model is not too complicated, it could be analyzed in this way. Because family and household models easily get quite complex, analytic models are rare in family and household demography. A good example of an analytic model (although valid under a number of quite restrictive assumptions only) is the kinship model of Pullum (1982).

For more complex models, analytic manipulation becomes infeasible, so that *numerical simulation* methods have to be used to study the implications of the model. Here, two fundamentally different strategies can be distinguished: *macrosimulation* vs. *microsimulation* (Van Imhoff and Post 1998; see also *Microsimulation in Demographic Research*).

In macromodels, numbers of events (e.g., ‘first marriages to women aged 25’) are obtained by applying rates (e.g., ‘first-marriage rate for 25-year old women’) to a *group* of individuals of a certain size (e.g., ‘number of never-married women aged 25’). In

micromodels, events are obtained by applying probabilistic decision rules (Monte Carlo experiments) to *individual persons*, separately for a number of individual persons. To follow the example of first marriages: for each individual never-married woman aged 25 in the database, draw a random number between 0 and 1; if this random number is less than the first-marriage probability, the woman is deemed to marry, otherwise she remains unmarried.

In macrosimulation, the calculations are carried out in terms of the cells in the aggregate cross-classification table: for each cell, the model should evaluate how the number it contains will change over time. Microsimulation, on the other hand, does its calculations in terms of the individual records: for each individual, the attribute vector is updated according to the specifications of the model and the results of the Monte Carlo experiments. The storage and retention of information in microsimulation occurs via a *list* of individuals and their attributes; in macrosimulation this is done via the aggregate cross-classification *table*. If there are K attributes and M_i categories for attribute $i = 1 \dots K$, the table at the macrolevel consists of $M_1 \times M_2 \times \dots \times M_K$ cells; in contrast, at the microlevel a total population of N individuals can be described by a matrix with $N \times K$ cells. For most applications where a relatively large number of attributes is considered, the size of the aggregate table is much larger than the size of the list. This is particularly the case in household and family models, where characteristics of the individuals to whom relationships exist need to be included. In macromodels this would lead to incredibly large tables.

In micromodels, contrary to macromodels, it is quite easy to keep links between individuals intact, simply by including in the individual records of the database some reference numbers (pointers) to other records (persons) in the database. As a result, the consequences of an event that is simulated for one person can be easily determined and updated for the other persons involved. Thus, the two-sex problem or more generally the consistency problem discussed in the previous section is more or less 'automatically' solved in micromodels. The other side of the coin is that, for events that involve creating a *new* link between individuals, a matching problem has to be solved (e.g., if a woman is simulated to marry, a concrete husband must be identified).

Another main advantage of microsimulation is that it can provide very *rich output*. The output of a microsimulation model consists of a database with individual data that can be aggregated in an almost infinite number of ways. Compare this with the situation in macromodels, where the aggregation scheme is fixed once the model (i.e., the state space) has been specified. Apart from detailed cross-sectional counts, the microsimulation database can also be used to construct longitudinal information, e.g., in the form of individual biographies.

In short, microsimulation offers three main advantages that are of particular relevance in family and household demography: it can handle a large state space; relationships between individuals can easily and explicitly be retained; and it provides richer output.

This does not automatically imply that microsimulation is the ideal methodology for family and household modeling. As always, these advantages are bought at a cost, and the cost can be heavy. The most important drawbacks of microsimulation are twofold: first, the model results are subject to random variation, and the degree of this randomness gets larger as more explanatory variables are included; second, it takes a lot of time and effort to construct a concrete microsimulation application.

4.3 Dynamic vs. Static Models

Populations change over time, and so do household and family structures. Thus, if a model is to describe this change, it must always contain the *time* element in one way or another. In this sense, all demographic models are dynamic by definition. However, there are many ways in which time can be included into a model. Merely adding an index t to all model variables and parameters hardly warrants the term 'dynamic.' A truly dynamic model should not only specify what the system looks like 10 years from now, but also how the system is supposed to get from here to there. In other words, the *processes* that underlie the changes in the system variables should be explicitly included in the model. In a truly dynamic model, the focus is on the *events* of formation and dissolution of relationships between individuals. The LIPRO household model belongs to this category (Van Imhoff and Keilman 1991, Van Imhoff 1995).

In contrast, *static* models do little more than compare household and family structures at different points in time, thus essentially treating the underlying processes of household and family formation and dissolution as a black box. The classical headship ratio models belong to this category. The headship ratio is the proportion of household heads in a population category, e.g., '80 percent of all 40-year old males are head of a household.' If from a population forecast (see *Population Forecasts*) we know that the number of 40-year old males will increase from 1,000 to 1,500, and we extrapolate that the headship ratio for 40-year old males will increase from 80 percent to 90 percent, then it follows that the number of households headed by a 40-year old male will increase from 800 to 1,350. With headship ratios for all ages and both sexes, a full population forecast can in this way be easily converted into a forecast of the number of (private) households. Several generalizations of the classical headship ratio model exist (Kono 1987). However, they all suffer from the disadvantage that the headship ratio is a static concept: the model cannot tell us why the current ratio of 80 percent will increase to 90 percent.

In a way, the dynamic vs. static issue is gradual rather than fundamental. For instance, if a demographic forecaster hypothesizes that marriage rates will fall by 10 percent between now and 10 years into the future, one could argue that the marriage component of the model is static since the demographic forecaster does not specify *how* the marriage rates are going to fall. Similarly, the headship ratio method is generally termed static since the changes in age-specific headship are not explicitly specified; however, the changes in the age-specific population size to which these headship ratios are applied *are* explicitly modeled, so the headship ratio model is dynamic at least to some extent.

In fact, leaving some model variables static rather than 'truly' dynamic is an effective strategy to reduce the complexity of the model. The decision of where to stop making variables dynamic is governed by the fundamental tradeoff between information intensity on the one hand, and the capacity to make meaningful predictions on the other. This tradeoff must be faced in every effort of modeling human behavior.

The household model used by Statistics Netherlands for its official household forecasts for The Netherlands is a nice example of a carefully constructed mixed dynamic-static household model (De Beer 1995). This model consists of two layers. The first layer is a multistate marital status model, producing forecasts of the population by age, sex, and marital status. In the second layer, each combination of age, sex, and marital status is distributed over various *household positions* (e.g., living alone, with partner, as lone parent, as dependent child, etc.), using household position proportions which are extrapolated into the future. The rationale for this strategy is that the available statistical data are sufficiently reliable to allow a dynamic approach to nuptiality, but do not warrant a dynamic treatment of household formation and dissolution. The resulting model is dynamic to the extent that household events are related to nuptiality behavior.

See also: Family Theory: Economics of Marriage and Divorce; Household Archaeology; Household in Anthropology; Kinship in Anthropology

Bibliography

- Bongaarts J 1983 The formal demography of families and households: An overview. *IUSSP Newsletter* 17: 27–42
- De Beer J 1995 National household forecasts for The Netherlands. In: Van Imhoff E, Kuijsten A, Hooimeijer P, Van Wissen L (eds.) *Household Demography and Household Modeling*. Plenum, New York/London
- Grebenik E, Höhn C, Mackensen R (eds.) 1989 *Later Phases of the Family Cycle*. Clarendon Press, Oxford, UK
- Keilman N W 1985a Nuptiality models and the two-sex problem in national population forecasts. *European Journal of Population* 1: 207–35
- Keilman N W 1985b Internal and external consistency in multidimensional population projection models. *Environment and Planning A* 17: 1473–98
- Keilman N W 1995 Household concepts and household definitions in Western Europe: Different levels but similar trends in household developments. In: Van Imhoff E, Kuijsten A, Hooimeijer P, Van Wissen L (eds.) *Household Demography and Household Modeling*. Plenum, New York/London
- Keilman N W, Kuijsten A, Vossen A (eds.) 1987 *Modelling Household Formation and Dissolution*. Clarendon Press, Oxford, UK
- Kono S 1987 The headship rate method for projecting households. In: Bongaarts J, Burch T K, Wachter K W (eds.) *Family Demography*. Clarendon Press, Oxford, UK
- Pullum T W 1982 The eventual frequencies of kin in a stable populations. *Demography* 19: 549–65
- Ryder N 1987 Discussion. In: Bongaarts J, Burch T K, Wachter K W (eds.) *Family Demography*. Clarendon Press, Oxford, UK
- United Nations. Statistical Division 1998 *Principles and Recommendations for Population and Housing Censuses, Revision 1*. United Nations, Department of Economic and Social Affairs, Statistics Division, Statistical Papers Series M No. 67/Rev.1, New York
- Van Imhoff E 1992 A general characterization of consistency algorithms in multidimensional demographic projection models. *Population Studies* 52: 159–69
- Van Imhoff E 1995 LIPRO: A multistate household projection model. In: Van Imhoff E, Kuijsten A, Hooimeijer P, Van Wissen L (eds.) *Household Demography and Household Modeling*. Plenum, New York/London
- Van Imhoff E, Keilman N 1991 *LIPRO 2.0: An Application of a Dynamic Demographic Projection Model to Household Structure in The Netherlands*. Swets & Zeitlinger, Amsterdam/Lisse
- Van Imhoff E, Post W 1998 Microsimulation methods for population projection. *Population* 52: 889–932

E. van Imhoff

Families as Educational Settings

1. Introduction

Children's education is viewed here in broad terms, reflecting not only academic standing in the formal setting of school, but also general cognitive development. In that vein, family influences on children's educational progress are myriad, as family provides the primary context for children's development and also reflects the broader societal and cultural contexts within which the family is embedded. Both change and continuity in family influences are noted during infancy, early childhood, and the school years.

2. Early Research

Much of the early research in family influences on children's educational progress examined ways in which academic outcomes were related to demo-

graphic characteristics such as parent marital status and education, ethnicity, birth order, and socioeconomic status. This early work revealed consistent, moderate relationships between most of these factors and children's cognitive development. In particular, more positive educational outcomes were found in families where the parents' marriage remained intact, where parents were better educated and had higher income, and for children who were first- rather than later-born. Psychologists and educators have noted that such group comparisons, which have been labeled 'social address' (Bronfenbrenner 1979), are less informative than research which identifies the processes underlying these group differences. Four categories of process variables are summarized here: parental beliefs, instruction from parents and siblings, structuring of the home environment and family life, and parents' behaviors outside the home.

3. Parental Beliefs

Parents' beliefs affect children's educational outcomes both through their influences on parental behaviors and on children's developing beliefs. For instance, a parent will speak to her 6-month-old daughter more often if she believes the infant can benefit from 'conversations' with an adult partner. Parents' beliefs also shape the developing beliefs of their children, many of which serve a motivational role for the child. For instance, a child whose father conveys that a failure was due to lack of effort rather than lack of ability will be more likely to persevere on a future task. Three types of parental beliefs are described here: parents' educational aspirations for their children, attributional beliefs about the reasons underlying academic outcomes, and beliefs about child rearing.

3.1 Parents' Educational Aspirations

Most parents have expectations for their children's performance in school and eventual educational attainment. Both aspirations regarding the child's future education level and projections of what specific grades a child will obtain are frequently highly optimistic. Thus, for instance, many parents who themselves have failed to complete high school expect that their own children will be college graduates.

It is clear that children's competencies inspire parents' aspirations (i.e., parents have higher expectations of more competent children). In general, the research literature also indicates an independent effect of parent expectations on children's achievement: Parents with more ambitious aspirations have, on average, children who excel academically and who are more likely than other children to pursue higher education (e.g., Okagaki and Frensch 1998). Parents who value education and who have high aspirations

for their children are likely to foster achievement striving in their children and to provide a supportive learning environment in the home. Another way in which high parental expectations foster children's educational development is through the degree of challenge parents provide in learning situations.

3.2 Parents' Attributional Beliefs

The classic work of Rotter (1966) and Weiner (1979) demonstrated that individuals vary in their beliefs about the reasons underlying success and failure, and that future performance is related to these beliefs. When failure is attributed to an internal controllable factor such as effort, the individual is more likely to persist than if the failure is attributed to lack of ability or an external factor such as teacher preference or luck. Conversely, a belief that success is due to the internal factors of high ability and hard work is likely to foster persistence. Attributional beliefs are thus posited to influence children's outcomes largely through their motivational effect on children's persistence in the face of failure and/or challenging tasks.

Much of the research in parents' attributions has entailed cross-cultural investigations of ethnic or national differences in children's academic achievement. This work shows that Asian parents are more likely to attribute their children's academic outcomes to effort than are parents of European cultural heritage; these beliefs are then presumed to lead to academic superiority of Asian students over European American students, particularly in math. Although considerable evidence has supported the notion that children's beliefs influence their achievement striving, particularly in the face of failure, it is unclear that beliefs explain cross-national differences in achievement. European parents are more likely than European Americans to espouse ability attributions, yet children's math achievement in a number of European countries is higher than in the USA. Cultural variations in beliefs are undoubtedly confounded with other cultural and national differences such as schooling characteristics.

3.3 Parents' Beliefs about Child-rearing

Parents' beliefs about child-rearing also contribute to the educational qualities of the home environment. These beliefs include parents' views of optimal ways to socialize children (e.g., discipline practices, how and to what extent autonomy is fostered) and age-related expectations of children's abilities. A substantial body of literature indicates more positive cognitive and social outcomes for children whose parents show warmth while exhibiting firm yet flexible control, a style called democratic parenting by Baldwin (1955), and later described by Baumrind (1973) as authoritat-

ive parenting. Authoritative parents rely primarily on reasoning as a discipline tool rather than punishment and view their children's good behavior as the primary goal of their parenting rather than obedience. Thus, parents who value their children's autonomy, and whose knowledge and beliefs are consistent with an authoritative parenting style are more likely to foster educational progress in their children, as marked by school grades, low incidence of discipline problems, and healthy self-esteem. However, some recent research indicates cultural variation in the relationship between parenting style and children's outcomes, both in the importance of autonomy encouragement and the possible detrimental effects of harsh discipline.

Parents vary tremendously in their beliefs about children at different ages. For instance, some parents expect that 18-month-olds can be potty-trained, whereas other parents are nonplussed when the three-year-old continues to have 'accidents.' By the same token, some parents expect their three-year-olds to master rudimentary concepts of number and quantity that other parents would not expect before the age of seven. Although parents may not be conscious of many of these beliefs, they nevertheless shape everyday activities in the household, family rules, and the messages children are given—both blatant and subtle—regarding their abilities. For example, parents who believe that an eight-year-old is capable of planning are more likely than other parents to expect their child to pack his/her school lunch, select weather-appropriate clothes, and budget time to include homework, soccer practice, and leisure in a given evening.

Research in this area has indicated that a U-shaped function operates; expectations that are either too low or too high are not optimal for the child. According to Vygotsky's concept of scaffolding (1978), parents foster children's cognitive progress when they provide graduated cues that enable the child to function at a level slightly beyond what the child could perform independently. A parent who persistently has low expectations is unlikely to challenge the child to progress further. On the other hand, expectations that are too high (e.g., attempting to teach algebra to an average kindergartner) are equally unlikely to enhance the child's development.

4. *Instruction by Parents and Siblings*

Parent beliefs affect child outcomes primarily through their impact on parent behaviors. One type of parent behavior that shapes the educational quality of the home is parent instruction. In addition, siblings are an important source of information for young children. Instruction from parents and siblings are discussed here together. From birth onward, parents and siblings are a very important source of children's knowledge. The parent's early language use guides the

development of the infant's ability to distinguish among phonemes. The toddler observes an older sibling's interactions with friends and imitates that behavior later in other social situations. A young adolescent acquires concepts of social stereotypes through his parents' conversation at the dinner table.

Family influences are readily evident in a child's acquisition of specific skills and knowledge about the world. Thus, for example, a child whose parents are swimmers will probably learn to swim, and the child of musicians is likely to know more about music than peers from nonmusical families. Family also provides the primary context for children's cognitive development in less obvious ways. Through everyday conversations and activities with family members, children acquire concepts of schemas (e.g., what a visit to the doctor's office entails), grammar, and mathematical concepts (e.g., the early distinction between 'one' and 'more;' how to fairly divide the last piece of cake). Children learn memory strategies (e.g., where can I put my show-and-tell item so I won't forget it tomorrow morning), meta-cognitive knowledge (I can't concentrate on my homework while big brother is listening to that music), and problem solving (How should I clean up this spilled sugar?).

Some instruction from family members is verbal; other 'instruction' is nonverbal, communicated through body language and mood, or is modeled through behavior that the child later imitates. Instruction that takes place in the home is markedly different from that found in the school setting, in that it is largely unplanned, informal, and context-grounded, rather than the planned, formal, and context-independent instruction that occurs in school. Most instruction in the home occurs spontaneously based on the specific events, feelings, and needs of a given moment in the family life. Because it is so highly context-dependent, this instruction from siblings and parents provides the basis for most of the child's knowledge and skills of how to survive in the day-to-day world.

5. *Structuring of the Home Environment*

In addition to their beliefs and instruction, parents shape the educational qualities of the home environment in important ways through their choices regarding the physical qualities of the home and the child's activities. The well-established relationship between socioeconomic status and children's academic success has been attributed in part to substantial material differences in the home environments of children. Space, availability of developmentally appropriate games and toys, warm and responsive caregivers, and exposure to cognitively stimulating adults and new experiences have all been found to be related to later IQ or school achievement. A child's needs change with age. For instance, a supportive

home environment for infants includes a responsive mother and appropriate play materials, whereas in mid-childhood emotional climate of the home and child's participation in the family are associated with higher school achievement (Bradley et al. 1988).

In addition to the physical and social characteristics of the home environment, the lifestyle choices that parents make for their children determine on a day-to-day basis the educational experiences of the children. These lifestyle choices include how children spend their time outside of school; parental supervision of homework; the amount of television the child is permitted to watch; the extent to which play with peers is monitored; and so forth. All of these factors are related to children's school achievement. Differences in the educational outcomes of children of divorce versus intact marriages is now attributed in part to reliable differences in the home environments and daily activities of these groups of children. Single parents, who on average must survive on a lower income than couples, have fewer material and emotional resources to devote to their children. Thus children in single parent households are more likely than peers in two-parent households to have little adult supervision during after-school hours, to interact with parents who are stressed, and to shoulder heavy domestic responsibilities at an early age. These qualities of the home environment rather than parental marital status have the strongest influence on children's cognitive and emotional outcomes.

6. Parents Outside the Home

The child's development occurs within a multilevel ecological framework (Bronfenbrenner 1979). In addition to the microsystem level, where the child is involved in all interactions (e.g., family, school, church), other settings influence the child's development. At the mesosystem level, microsystem interactions occur without the child's presence (e.g., a parent attending a school open house). The exosystem, in turn, is comprised of those settings which never include the child, but which nonetheless influence the child. Two of these broader settings are discussed here: parents' employment (an exosystem influence) and parents' involvement in the child's school (a mesosystem influence).

6.1 Parental Employment and Children's Development

Early investigations in parent employment and children's outcomes were mostly group comparisons aimed at determining whether or not maternal employment had a detrimental effect on children. These investigations indicated no differences in quantity or quality of parenting systematically related to maternal

employment status. More recent research has identified process variables related to parent employment that have both positive and negative influences on children.

According to scarcity or stress theory, parents have a limited amount of time and energy, and parental employment is deleterious to children's developmental outcomes to the extent to which it decreases parents' time, energy, and patience. Stress theory reflects a strong trend in the USA and many other industrialized nations: Namely, in spite of the impressive technological advances of the past century, most parents work harder than ever (Schor 1991). Particularly as maternal employment rates have increased, the everyday lives of children have become harried. Thus one consequence of parental employment for children is the resulting stress associated with parental employment-related absence, whether that absence is actual or psychological, as stressed parents are preoccupied, tired, and impatient. A second consequence of demographic changes in employment patterns is that children spend a greater amount of time with other caretakers. Therefore, parents' selection of day care for infants and preschoolers, and their choices of after-school activities for school age children have real impact on children's developmental outcomes.

Parental employment can also benefit children's educational progress. On the one hand, parents gain in self-esteem, competency, and social support (as well as income) through their employment; these benefits for parents may translate into benefits for children. Mothers who are employed in professional positions serve as positive role models for daughters. In addition, parents' work characteristics are related to the quality of their parenting. Parents who enjoy their work, who are socially rewarded for it, and who are intellectually challenged in the work setting are more likely to show warmth and greater flexible but firm control with their children than parents who have less satisfying employment (Greenberger et al. 1994).

6.2 Parents' Involvement in Schooling

Formal education is an important part of the everyday lives of children and—along with family—is the most important source of children's educational progress in most countries. Therefore, parents' attitudes toward school and involvement in school activities have long been studied as an important factor related to children's educational progress. Parent involvement in children's formal education is positively related to their children's achievement: Parents of higher-achieving children tend to have more positive attitudes toward school, attend school functions, volunteer at their children's schools, and know teachers better than parents of lower-achieving children. Some theorists have pointed out that families of higher socioeconomic status are more comfortable with the cultural milieu of

schools, and thus find it easier to approach teachers and school administrators than families of lower socioeconomic status. These barriers to parent involvement are even greater when parents represent an underprivileged minority group and have had previous contacts with school officials that were negative (Clark 1983).

7. Current Research Directions

The home is one setting embedded within broader societal, cultural, and generation settings. Much current research examines how the family is changing due to changes in demographic factors such as parent employment, poverty, divorce and remarriage, and migration. Current research also emphasizes cultural variations in family life, both in terms of similarities and differences across groups, and in understanding what is optimal for children's healthy development (e.g., Okagaki and Frensch 1998). Finally, current research seeks to better understand family as an educational setting within a developmental framework, studying both the ways in which family influences change across childhood and the long term effects of early enhancement or insult.

See also: Education and Employment; Family and Schooling; Family as Institution; Family Processes; Nontraditional Families and Child Development; Parenting in Ethnic Minority Families: United States; Parents and Teachers as Partners in Education; Socialization in Adolescence; Socialization in Infancy and Childhood

Bibliography

- Baldwin A L 1955 *Behavior and Development in Childhood*. Dryden Press, New York
- Baumrind D 1973 The development of instrumental competence through socialization. In: Pick A D (ed.) *Minnesota Symposium on Child Psychology*. University of Minnesota Press, Minneapolis, MN, Vol. 7, pp. 3–46
- Bradley R H, Caldwell B M, Rock S L 1988 Home environment and school performance: A ten-year follow-up and examination of three models of environmental action. *Child Development* **59**: 852–67
- Bronfenbrenner U 1979 *The Ecology of Human Development: Experiments by Nature and Design*. Harvard University Press, Cambridge, MA
- Bronfenbrenner U 1986 Ecology of the family as a context for human development: Research perspectives. *Developmental Psychology* **22**: 723–42
- Clark R 1983 *Family Life and School Achievement: Why Poor Black Children Succeed or Fail*. University of Chicago Press, Chicago
- Connors L J, Epstein J L 1995 Parent and school partnerships. In: Bornstein M H (ed.) *Handbook of Parenting. Applied and Practical Parenting*. Erlbaum, Mahwah, NJ, Vol. 4, pp. 437–58

- Greenberger E, O'Neill R, Nagel S K 1994 Linking workplace and homeplace: Relations between the nature of adults' work and their parenting behaviors. *Developmental Psychology* **30**: 990–1002
- Hart B, Risley T R 1995 *Meaningful Differences in the Everyday Experience of Young American Children*. Brooks, Baltimore, MD
- Okagaki L, Frensch P A 1998 Parenting and children's school achievement: A multiethnic perspective. *American Educational Research Journal* **35**: 123–44
- Rotter J B 1966 Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs* **80**: No. 609
- Schor J B 1991 *The Overworked American*. Basic Books, New York
- Scott-Jones D 1984 Family influences on cognitive development and school achievement. *Review of Research in Education* **11**: 259–304
- Vygotsky L S 1978 *Mind in Society*. Harvard University Press, Cambridge, MA
- Weiner B 1979 A theory of motivation for some classroom experiences. *Journal of Educational Psychology* **71**: 3–25

B. Kurtz-Costes

Family and Consumer Sciences: United States

Family and Consumer Sciences is the study of relationships among individuals, families, and communities and the social, economic, political, biological, physical, and aesthetic environments in which they function. An early widely accepted definition stated:

... in its most comprehensive sense (it) is the study of the laws, conditions, principles and ideals which are concerned on the one hand with man's immediate physical environment and on the other hand with his nature as a social being, and is the study specifically of the relations between those two factors (Lake Placid Conference 1902).

Themes which characterize the field include (a) holistic, systems approach to the study of individuals, families, and communities, (b) interdisciplinarity as the field draws knowledge from the social, behavioral, and life sciences, from technology and engineering, and from the arts and humanities, and (c) integration and synthesis of concepts as they are applied to well-being. A continuum from discovery and integration of knowledge through its application in practical settings also is a defining characteristic of the field.

Most of the work of the profession fits this characterization; however, degrees of emphasis on particular elements and the desire to communicate that emphasis through nomenclature has led to the use of a variety of names. This is most evident in the naming of colleges in higher education institutions in the United States. Human ecology, human sciences, human environ-

mental sciences, human resources, as well as family and consumer sciences are used to designate this field of study in the academy.

Family and consumer sciences is mission-oriented. As a profession, field of study, and critical science, the mission of family and consumer sciences is 'the empowerment of families to function interdependently and the empowerment of individuals to perform family functions' (Green 1996). The mission-oriented nature of family and consumer sciences is reflected in statements about areas of leadership for the profession, to wit:

- (a) improving individual, family, and community well-being;
- (b) impacting the development, delivery and evaluation of consumer goods and services;
- (c) influencing the development of policy; and
- (d) shaping societal change; thereby enhancing the human condition ('The Conceptual Framework for the 21st Century' 1994).

1. Definition and Body of Knowledge

Family and consumer sciences is the study of relationships among individuals, families, and communities and the reciprocal relations of family to its human-created and natural environments. Specific elements in the body of knowledge have varied somewhat over time. A task force of individuals from many of the organizations linked to family and consumer sciences began updating the specification of the body of

knowledge in 2000. The work of this group is incorporated in Fig. 1, which depicts the core concepts in family and consumer sciences (Baugher et al. 2000).

1.1 Basic Human Needs

A primary expectation of families in most societies is that they work toward meeting basic human needs. Abraham Maslow's conceptualization summarizes these needs as physiological needs (food, shelter, clothing), safety (security and the absence of fear), love and belongingness (human emotions connecting individuals in human systems), self-esteem (sense of competence and respect for self), self-actualization (achieving one's human potential) (see Paolucci et al. 1977).

Starting from the center of this model and moving outward, three interdependent concepts are represented: individual well-being, family well-being, and community vitality. Subsumed within these concepts, but not delineated, are multiple specific areas of research within family and consumer sciences. These concepts also represent the mission-oriented nature of family and consumer sciences. Together, the three areas reflect the ultimate concern of family and consumer sciences to enhance the quality of life.

1.2 Individual Well-being

Individual well-being refers to the promotion of a person's physical and mental health, knowledge and

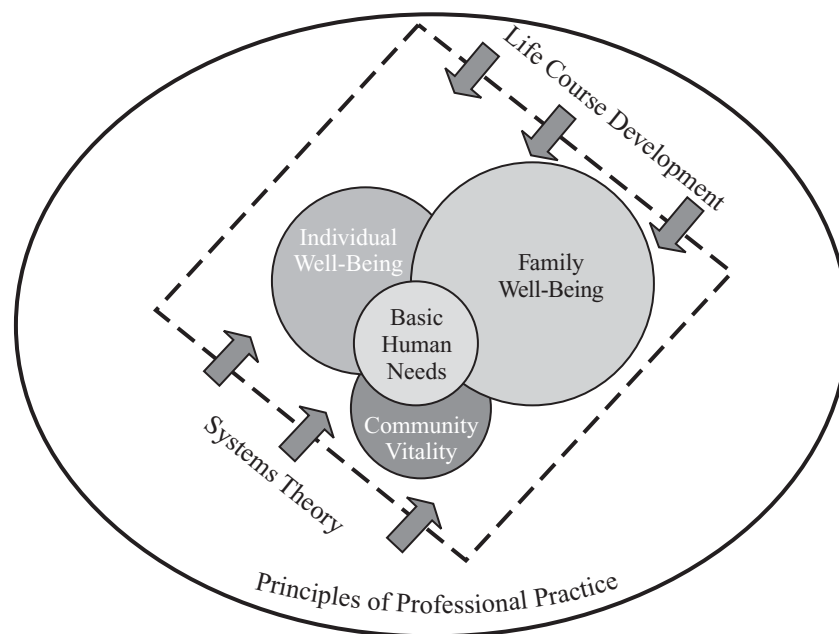


Figure 1
Core concepts in family and consumer sciences

skills, values, and civic responsibilities. Research in this domain may be as basic as studies to understand human metabolic processes and human genetic make-up, personality traits, or self-esteem as reflected in the selection of apparel. Application of this concept is reflected in numerous family and consumer sciences programs to increase knowledge and skills and foster more effective personal functioning.

1.3 Family Well-being

Family well-being is concerned with the development and dissemination of knowledge designed to empower families to meet the basic human needs of their members and engage in processes that build sustainable viable communities. Roles, responsibilities, and functions of the family have predominated the concerns of family and consumer sciences rather than structure (Green 1996). This is reflected in the definition of family adopted by the American Association of Family and Consumer Sciences. Family is defined as 'two or more persons who share resources, share responsibility for decisions, share values and goals, and have a commitment to one another over time' (Bivens et al. 1975, p. 26).

In a statement on the 'intellectual ecology' of family and consumer sciences Green (1996) identifies the substantive core of family and consumer sciences as family functions. The family functions framework addresses the role of families in meeting the needs of its members (i.e., basic human needs). In the developed economy and contemporary society of the United States, there is great interdependence of families and other systems in producing the goods and services to meet these needs. Accreditation documents of the American Association of Family and Consumer Sciences have delineated these aspects of human needs in relation to the near environment that are of concern to family and consumer sciences (see Council for Accreditation 1995). The following concepts would generally be found in such a listing:

- (a) Family systems, including economic, social, cultural, political, and demographic factors; family dynamics and roles;
- (b) Interdependence of families and communities;
- (c) Human development across the life course;
- (d) Resource development and management, including production, consumption, disposition, management and decision-making;
- (e) Human nutrition and food in the development, health and wellness of individuals, families, and communities;
- (f) Design and changing technology in relation to the development of home, community, and work environments that are supportive of consumers and families;
- (g) Role of apparel and textiles in meeting the physical, social, and aesthetic needs of consumers and families.

1.4 Community Vitality

Community vitality fosters interdependency and collaboration designed to enhance the well-being of individuals and families. This concept may apply to those who reside in a specific locality or to groups who perceive themselves as a community based on shared interests and beliefs. Technology makes it possible to sustain community vitality in cyberspace. Community vitality can encompass the global community.

1.5 Systems Theory

A system is defined as interrelated elements having functional unity within a larger system. 'Ecosystem' is the regular interdependency of organisms within their environment. The basic concepts found in systems theory are development, relationships, interdependence, and reciprocity. The fundamental theoretical framework that gives family and consumer sciences its dynamic and integrative nature is family ecosystems. Paolucci et al. (1977) define family ecosystems as follows:

Family members, their external environments as perceived by them, and the web of human transactions carried out through the family organization constitute the basic elements of the family ecosystem. One fundamental characteristic of the family ecosystem is that it is made up of a collectivity of interdependent but independent parts working together to achieve a common purpose. Each element (organism and environment) is interrelated (p. 15).

Human ecology theory is congruent with family ecosystems, and human ecology theory is often identified as the theoretical foundation of family and consumer sciences. Bubolz and Sontag (1993) identify the uniqueness of human ecology theory as its 'focus on humans as both biological organisms and social beings in interaction with their environment' (p. 419). The use of the human ecological perspective is found in many social science disciplines and applied fields.

1.6 Life-course Development

The concept of life course refers to a sequence of socially defined, age-graded events and roles that individuals enact over time (Elder 1998). Life-course thinking appears in most of the behavioral sciences scholarly literature after the 1960s.

As understood in life-course theory, human development is a co-active process in which social, cultural, psychological, and biological influences interact over time in the context of culture and social structures (Elder 1998). The concept of life-course development for individuals has been extended in family and consumer sciences to families, which also are observed to have experiences and characteristics described as life course events. An understanding of

life-course development is important in family and consumer sciences because environments in which families function, and families themselves, change over time.

1.7 The Professional Context

As shown in the model, principles of professional practice are a part of the core concepts of family and consumer sciences. Brown and Paolucci (1979) identify three criteria for identity as a profession:

(a) A profession is oriented toward providing a service;

(b) The service involves intellectual activity, including practical judgment, which requires that the professional master theoretical knowledge related to the work;

(c) Members in the profession seek to assure that work within the profession is morally defensible both in nature and in the quality of performance.

Family and consumer sciences includes courses and experiences to facilitate a professional orientation with regard to ethics, valuing diversity, critical thinking skills, working in partnerships, and advancing policies that support the well-being of individuals, families, consumers, and communities.

1.8 Synergistic/Integrative Approach

Family and consumer sciences claims as one of its distinguishing characteristics a 'synergistic, integrative nature' (Council for Accreditation 1995). The overlapping circles, dotted lines, and arrows in Fig. 1 depict this interplay among concepts. Addressing issues of wellness, for example, in the context of multiple factors rather than as isolated elements is the essence of the synergistic concept. For example, a dietitian needs to understand that obesity and impending cardiovascular disease are not just a matter of nutrition. Obesity is probably related to cultural food patterns, economic resources, and family relationships.

The integrative focus entails recognizing the interrelatedness of individuals, families, and consumers with their environments. In the example, the absence of sidewalks and parks with walking trails in a community may be a barrier to the state health agency's goal of reducing obesity through fitness programs.

2. Specializations and Family and Consumer Sciences

Like a chameleon, family and consumer sciences is perceived differently by observers depending on the context and the 'colors' highlighted in that context. In

the mid-twentieth century, Lee and Dressell (1963) observed three conceptions of the field, as follows:

(a) a single field with a broad general perspective and a number of specialties;

(b) a unified field with subspecialties embedded in the home and family;

(c) a collection of disciplines with no unifying theme or 'anchor.'

Families may be recognized as the basic social unit of society, but technology and changing social norms increasingly influence how family functions are carried out. In economically developed countries, such as the United States, the household produces only a small portion of the goods and services it consumes. The family has become increasingly a managerial unit.

As family roles changed and new technology, products, and services emerged, specializations developed related directly to the family or derived from the functions of the family (e.g., providing food, shelter, nurturance, clothing, and reproducing children and culture) (Bailey and Firebaugh 1986). Some specializations developed around certain human issues (e.g., nutrition, dependent care of the young and the elderly, housing). Other specializations focus on the application of knowledge in business, industry and agencies (e.g., restaurant management, interior design, marriage and family therapy). Certain specializations may or may not be found in family and consumer sciences academic units. At the turn of the twenty-first century, examples of the three conceptions of family and consumer sciences can still be found in academic units in the United States.

3. History

Family and Consumer Sciences evolved from home economics. Home economics emerged at the turn of the twentieth century in the context of the progressive reform movement. The organizers of the home economics movement were educated women and men who deplored the problems associated with urban crowding, child labor, malnutrition, immigration, and lack of education; but they also believed that the application of scientific knowledge could improve the daily lives of people and the policies of public agencies and employers (Stage 1997, Vincenti 1982). They organized courses to improve nutrition and the management of resources in public schools and neighborhood organizations in Boston and New York City. They launched campaigns to reform corrupt practices of landlords, employers, and businesses. They advocated for responsible 'municipal housekeeping' in which government would improve living conditions and citizens would be involved in governing bodies.

The educational innovations of the 1860s, such as the creation of land-grant universities through the Morrill Act of 1862 (whose express purpose was to educate the children of the common folk), offered

opportunities for higher education where they were not previously available. Home economics became a major point of entry for women into public higher education. The 'second Morrill Act' was passed in 1892, authorizing institutions of higher education for black citizens in the Southern States. The home economists who worked in the Extension Service and in the Southern schools and colleges worked in a segregated system with few resources (Harris 1997). Despite these constraints, their efforts greatly improved the daily lives of their constituents and offered career opportunities to those with formal education.

In 1899 began a series of meetings at the Lake Placid Club at Morningside, New York, for the expressed purpose of uniting efforts in the study of the home under one name and to organize a body of knowledge (see Brown 1985 and Lake Placid Conference Proceedings 1902–8). The proceedings record lively dialogue about nomenclature and content. Home economics was adopted as the name of the discipline. Ellen H. Swallow Richards, a chemist and the first woman graduate and first woman faculty member at the Massachusetts Institute of Technology, emerged as a zealous leader for academic programs, demonstration and outreach projects, research, and professionalization of the field (Hunt 1942, Stage 1997). Ellen H. Swallow Richards is generally identified as the 'founder' of the home economics movement.

The growing strength of home economics in terms of students enrolled, organizational structure, and public policy advocacy is documented in Pundt (1980). She describes the first decade after the founding of the American Home Economics Association in 1909 as 'continuing the struggle of the new discipline to define itself and find a rightful place among the more orthodox disciplines of a college or university ...' (Pundt 1980, p. 25).

Ideas about the modernization of family roles and the application of science in the home took a central place in home economics during the first quarter of the century. Behavioral sciences greatly influenced home economics teaching about child development and parenting, and child rearing was integrated into the home economics curriculum (Grant 1997). Principles from industrial engineering were applied to work simplification and design in the kitchens of America. Home economists were called upon during the 1930s to help families meet the devastation of the Great Depression. An outcome of these efforts was that nutrition and dietetics achieved a secure place in home economics curriculum and outreach programs (Babbitt 1997).

Dramatic changes in social roles, attitudes, and organization at mid-century and beyond prompted a series of conferences—New Directions (1959) and New Directions II (1975), the French Lick Conference (1961), the Eleventh Lake Placid Conference (1973), and Home Economics Defined (1979)—as the field engaged in a 'search for identity' (Brown 1985). The

focus of most of these conferences and 'white papers' was the purpose and subject matter central to the field (Simerly et al. 2000).

In 1975, a definition of home economics was published in *New Directions II*:

The study of the reciprocal relations of family to its natural and man-made environments, the effects of these singly or in unison as they shape the internal functioning of families, and the interplay between the family and other social institutions and the physical environment (Bivens et al. 1975, pp. 26–27).

The Scottsdale Conference held in 1993 examined the mission, breadth, scope, and name of the field (Simerly et al. 2000, Stage and Vincenti 1997). The outcome of the meeting was the recommendation of the name Family and Consumer Sciences for the field, which was subsequently ratified by the sponsoring organizations. A systematic statement of the basic beliefs, assumptions, professional practice, and areas of focus was developed (The Conceptual Framework for the 21st Century 1994).

4. Research Methods

Research in family and consumer sciences covers the entire spectrum from empirical studies using survey research methods, experimental laboratory studies, historical analysis, qualitative research such as case studies and personal narrative, needs assessments, and evaluation studies. Both 'basic research' (i.e., research conducted for the sake of testing theory, or to study phenomena with no expectation of a utilitarian outcome) and 'applied research' (i.e., research directed toward the solution of specific problems) are found in family and consumer sciences (Touliatos and Compton 1988). Intervention projects based on experimental design are used to test the efficacy of programs aimed at behavioral change.

Certain research methods tend to be associated with specializations within family and consumer sciences. For example, nutrition, food science, and textile research tends to be conducted in laboratories using basic or experimental research methods. While animal models have long been used to study diet-related diseases, studies with human subjects are becoming more prevalent. Population studies based on the methods of epidemiology are employed in assessing the extent of dietary diseases.

Areas in textiles and apparel such as fashion merchandising and international trade adopt a variety of research methods ranging from historical analysis, psychometrics, and secondary data analysis. Textile scientists concerned with the environmental impacts of processing conduct field studies and operate pilot plants to investigate their research questions.

Demography is concerned with human population and changes over time, thus it plays a major role in

social sciences and history. Family demography has traditionally been used to study macrolevel family patterns, but as life-course events are increasingly matters of choice, the importance of both macro and microlevel understanding is recognized (Teachman et al. 1999). Demographic methods are used to study housing, family economics, marriage, and fertility.

Research in family studies relies heavily on cross-sectional surveys and quantitative analysis. Data for these studies may be collected through interviews or self-administered questionnaires. Longitudinal analysis based on panel studies of the same subjects provide insights about families and their behavior over time. Acock (1999) observes that advances in computer technology and software facilitate the analysis of quantitative data which was not possible as recently as the 1970s.

Observational studies either in clinical or laboratory settings or the 'naturalistic' setting of the home are used to study family interaction. Qualitative methods are increasingly accepted as legitimate for the study of family and consumer issues. Qualitative family research uses content analysis of in-depth interviews, documents and other artifacts, and case studies (Rosenblatt and Fischer 1993). Qualitative studies are especially useful in the study of meanings, perceptions, and other subjective phenomena.

The mission-oriented nature of family and consumer sciences creates a close affinity for action research. The primary objectives of action research are solving problems and improvement of practice, but increased knowledge is also a potential outcome (Touliatos and Compton 1988). Participatory action research incorporates the indigenous knowledge of the community. Valuing such knowledge is important in facilitating societal change.

5. Outreach and Application

The Hatch Act, passed in 1887, provided funds to the States to establish agricultural experiment stations at the land-grant universities. The goal was solving practical problems. The Smith-Lever Act was passed in 1914, creating the Cooperative Extension Service. This provided the institutional vehicle for disseminating the research results from the universities and encouraging citizens to adopt practices to improve their living standards (see Babbitt 1997, Firebaugh and Redmond 1996, Harris 1997). The Cooperative Extension Service was a unique creation of higher education in the United States. It continues to have the most access to the most people through its network of county-based education programs.

The tradition of taking knowledge to the people has been modified through the years and adopted by family and consumer sciences programs, and many related social and behavioral sciences, in all types of higher education institutions. Community-university

collaboration in identifying and addressing the problems facing families and youth in the United States is creating effective programs and solutions to many of these problems (Chibucos and Lerner 1999) and increasing the value of outreach scholarship within higher education (Lerner and Simon 1998). Limitations of space preclude a comprehensive review of the multitude of models and programs in which family and consumer sciences is in partnership with other disciplines and with constituents in the application of knowledge to address quality of life issues of the twenty-first century.

6. The Future

Family and consumer sciences has evolved with growing sophistication in research methods, proliferation of specializations within the field, use of new technologies for dissemination of knowledge, and advocacy for policies to protect families and consumers. The field, however, has remained focused on its mission. Fidelity to its mission and responsiveness to change are expected in the future.

Application of knowledge to persistent practical problems has characterized family and consumer sciences. The list of accomplishments is vast: the school lunch program, delineation of requirements for nutrients, standardization of clothing sizes, best practices models in dependent care, Truth in Lending and Truth in Saving legislation, to name a few. New issues present themselves in the twenty-first century. Most of these issues are complex and require multiple perspectives. More collaborative projects across the disciplines and with community groups can be expected in the future. New guidelines in federal and foundation funding, designed to foster a closer integration of research and application, will encourage family and consumer sciences researchers and practitioners to engage in more collaborative outreach efforts.

Increased availability of information through the Internet and rising education levels of the population are expected to change the focus of professional practice from an orientation of being an 'expert' to being a 'facilitator.' The holistic family ecosystems theoretical framework and the integrative, synergistic orientation of family and consumer sciences will provide a comparative advantage for practitioners working in the 'information society.'

Deliberations about the nature of family and consumer sciences as a discipline are sure to continue. Conceptualizations about the body of knowledge (Baughner et al. 2000), discourse on the nature of the discipline (Brown and Paolucci 1979), and attempts to clarify and unify the discipline and the profession (Green 1996, Simerly et al. 2000) are echoes of the discussions at the Lake Placid Conferences. Such examination of family and consumer sciences will foster its future vitality and relevance as a field of study and a profession.

See also: Community Health; Community Sociology; Consumer Culture; Consumer Economics; Consumer Psychology; Consumption, Sociology of; Family and Kinship, History of; Family and Schooling; Family, Anthropology of; Family as Institution; Family Bargaining; Family Business; Family Health; Family Law; Family Processes; Life Course: Sociological Aspects; Religion: Family and Kinship; Segregation Indices

Bibliography

- Acocck A C 1999 Quantitative methodology for studying families. In: Sussman M B, Steinmetz S K, Peterson G W (eds.) *Handbook of Marriage and the Family*, 2nd edn. Plenum Press, New York, pp. 263–89
- Babbitt K R 1997 Legitimizing nutrition education: The impact of the Great Depression. In: Stage S, Vincenti V B (eds.) *Rethinking Home Economics: Women and the History of a Profession*. Cornell University Press, Ithaca, NY, pp. 145–62
- Bailey L, Firebaugh F M 1986 *Strengthening Home Economics Programs in Higher Education*. College of Home Economics, The Ohio State University, Columbus, OH
- Baughner S, Anderson C I, Green K B, Nickols S Y, Shane J, Jolly L, Miles, J 2000 Body of knowledge of family and consumer sciences. *Journal of Family and Consumer Sciences* 92(3): 29–32
- Bivens G, Fitch M, Newkirk G, Paolucci B, Riggs E, St. Marie S, Vaughn G 1975 Home economics—New directions II. *Journal of Home Economics* 67(3): 26–7
- Brown M 1985 *Philosophical Studies of Home Economics in the United States: Our Practical-Intellectual Heritage, Vol. I*. Michigan State University, East Lansing, MI
- Brown M, Paolucci B 1979 *Home Economics: A Definition*. American Home Economics Association, Washington, DC
- Bubolz M M, Sontag M S 1993 Human ecology theory. In: Boss P G, Doherty W J, LaRossa R, Schumm W R, Steinmetz S K (eds.) *Sourcebook of Family Theories and Methods: A Contextual Approach*. Plenum Press, New York, pp. 419–48
- Chibucos T R, Lerner R M 1999 *Serving Children and Families Through Community-University Partnerships: Success Stories*. Kluwer Academic, Norwell, MA
- Council for Accreditation 1995 *Accreditation Documents for Undergraduate Programs in Family and Consumer Sciences, 1995 Edition*. American Association of Family and Consumer Sciences, Alexandria, VA
- Elder G H, Jr. 1998 The life course and human development. In: Lerner R M (ed.) *Handbook of Child Psychology*. Wiley, New York, pp. 939–91
- Firebaugh F M, Redmond M B 1996 Flora Rose: A leader, innovator, activist, and administrator. *Kappa Omicron Nu Forum* 9(2): 7–16
- Grant J 1997 Modernizing mothers: Home economics and the parent education movement, 1920–1945. In: Stage S, Vincenti V B (eds.) *Rethinking Home Economics: Women and the History of a Profession*. Cornell University Press, Ithaca, NY, pp. 55–74
- Green K B 1996 Our intellectual ecology: A treatise on home economics. In: Simerly C, Light H, Mitstifer D I (eds.) *A Book of Readings: The Context for Professionals in Human, Family and Consumer Sciences*. American Association of Family and Consumer Sciences, Alexandria, VA, pp. 1–7
- Harris C 1997 Grace under pressure: The Black Home Extension Service in South Carolina, 1919–1966. In: Stage S, Vincenti V B (eds.) *Rethinking Home Economics: Women and the History of a Profession*. Cornell University Press, Ithaca, NY, pp. 203–28
- Hunt C L 1942 *The Life of Ellen H. Richards*. American Home Economics Association, Washington DC
- Lake Placid Conference on Home Economics 1901 *Proceedings of the First, Second, and Third Conferences*. American Home Economics Association, Washington, DC
- Lake Placid Conference on Home Economics 1902 *Proceedings of the Fourth Annual Conference*. American Home Economics Association, Washington, DC
- Lake Placid Conference on Home Economics 1903 *Proceedings of the Fifth Annual Conference*. American Home Economics Association, Washington, DC
- Lake Placid Conference on Home Economics 1904 *Proceedings of the Sixth Annual Conference*. American Home Economics Association, Washington, DC
- Lake Placid Conference on Home Economics 1905 *Proceedings of the Seventh Annual Conference*. American Home Economics Association, Washington, DC
- Lake Placid Conference on Home Economics 1906 *Proceedings of the Eighth Annual Conference*. American Home Economics Association, Washington, DC
- Lake Placid Conference on Home Economics 1907 *Proceedings of the Ninth Annual Conference*. American Home Economics Association, Washington, DC
- Lake Placid Conference on Home Economics 1908 *Proceedings of the Tenth Annual Conference*. American Home Economics Association, Washington, DC
- Lee J A, Dressell P 1963 *Liberal education and home economics*. Bureau of Publications, Teachers College, Columbia University, New York
- Lerner R M, Simon A 1998 *University-Community Collaborations for the Twenty-First Century: Outreach Scholarship for Youth and Families*. Garland, New York
- Paolucci B, Hall O A, Axinn, N 1977 *Family Decision Making: An Ecosystems Approach*. Wiley, New York
- Pundt H 1980 *AHEA: A History of Excellence*. American Home Economics Association, Washington, DC
- Ralston P A 1996 Flemmie P. Kittrell: Her views and practices regarding home economics in higher education. In: Simerly C, Light H, Mitstifer D I (eds.) *A Book of Readings: The Context for Professionals in Human, Family and Consumer Sciences*. American Association of Family and Consumer Sciences, Alexandria, VA, pp. 181–90
- Rosenblatt P C, Fischer L R 1993 Qualitative family research. In: Boss P G, Doherty W J, LaRossa R, Schumm W R, Steinmetz S K (eds.) *Sourcebook of Family Theories and Methods: A Contextual Approach*. Plenum Press, New York, pp. 167–77
- Simerly C B, Ralston, P A, Harriman L, Taylor, B 2000 The Scottsdale initiative: Positioning the profession for the 21st century. *Journal of Family and Consumer Sciences* 92(1): 75–80
- Stage S 1997 Ellen Richards and the social significance of the home economics movement. In: Stage S, Vincenti V B (eds.) *Rethinking Home Economics: Women and the History of a Profession*. Cornell University Press, Ithaca, NY, pp. 17–33
- Stage S, Vincenti V B 1997 *Rethinking Home Economics: Women and the History of a Profession*. Cornell University Press, Ithaca, NY
- Teachman J D, Polonko K A, Scanzoni J 1999 Demography and families. In: Sussman M B, Steinmetz S K, Peterson G W (eds.) *Handbook of Marriage and the Family*, 2nd edn. Plenum Press, New York, pp. 39–75

- The conceptual framework for the 21st century 1994 *Journal of Home Economics* 86(4): 38
- Touliatos J, Compton N H 1988 *Research Methods in Human Ecology/Home Economics*. Iowa State University Press, Ames, IA
- Vincenti V B 1982 Toward a clearer professional identity. *Journal of Home Economics* 74(3): 20–25
- Vincenti V B 1997 Chronology of events and movements which have defined and shaped home economics. In: Stage S, Vincenti V B (eds.) *Rethinking Home Economics: Women and the History of a Profession*. Cornell University Press, Ithaca, NY, pp. 321–30

S. Y. Nickols

Family and Gender

The relationship between gender and the family has been a contested issue in the social sciences. Much of this controversy arises because even to speak about gender is, deliberately or inadvertently, to raise the issue of equality between men and women. Moreover, the issue of equality between the genders, in turn, is intimately linked to the paradoxical treatment of the family in classical liberal political thought (Eisenstein 1981).

1. The Classical Liberal Legacy—Patriarchal Antipatriarchalism

Classical liberal theory is one of the constitutive elements of modern society. Liberal political ideas emerged against the background of the collapse of the feudal social order and the emergence of a society based on market relations. Liberalism simultaneously promoted the ideals of freedom and equality, and strict separation between a public and a private realm. However, classical liberalism's assertion that principles of equality and freedom should be limited to the public realm has led many later thinkers to see the transformation of power in private family households as the last remaining objective of an incomplete revolution.

The most celebrated exponent of classical liberalism, John Locke, developed his doctrine in opposition to the doctrine of patriarchalism—a Latin term meaning the rule of the father. The patriarchalist Sir Robert Filmer had justified the absolute rule of the monarch by analogy to the 'divine and natural' authority of fathers over their households. In Filmer's social world, no distinction was drawn between the public and the private capacity of the king. King James I, for example, was expected to finance his own government from the revenue generated by his own estates, crown lands, and customs.

Locke opposed Filmer precisely because Filmer conflated the political and the familial. In Locke's view each institution had a separate purpose. Family, as an institution dealing with the rearing of dependent children, was organized around the inequality of parent and child. In a market society, politics, by contrast, was concerned with the coordination of independent individuals; that is to say, it was organized around the principle of equality. Thus, Locke rejects patriarchal doctrine, acclaiming the principle of public equality between reasoning adults, while at the same time endorsing patriarchy as the appropriate logic of the private realm of the family.

Locke and his contemporaries excluded women from the public realm because they assumed that rationality was an exclusively male characteristic. Locke recognized women's role in procreation and the preparation of children for political life but assigned these activities to the private sphere. Furthermore, Locke believed that a 'wife's subordination to her husband had a "Foundation in Nature" and that the husband's will must prevail in the household as he is naturally "abler and the stronger"' (Pateman 1989, p. 121).

Classical liberalism's legacy, among other things, is a continuing tension between the antipatriarchal demand for equality associated with rule by consent in public life and the creation of a separate private sphere still based on patriarchal prerogatives. This tension has given rise to a demand for equality between the genders, even in the private sphere of the family. Indeed, the family has often been seen as a key institution and, in some cases, the very locus of patriarchal rule (Hartmann 1981). However, the demand for gender equality has different variants depending on different analyses of the significance of the family for the subordination of women.

2. Liberal Feminism—Escaping the Doll's House

Liberal feminists accept broadly the classical liberal idea that citizenship is based on participation in public institutions, especially the (labor) market. Furthermore, liberal feminists also to a great extent accept that what happens at home is a private matter to be negotiated between husband and wife. Consequently, theorists of this persuasion argue that the key to women's subordination is the restriction of women to the private sphere.

Liberal feminism urges women to be like Nora, the heroine of Henrik Ibsen's play *The Doll's House*, and throw off domesticity and participate fully in the labor market, competing with men for the economic and social rewards available through paid work. It assumes that the traditional domestic division of labor, where the man is the provider and the woman the homemaker, follows from the fact that the homemaker is more dependent economically on the provider. It is the

dependence of the housewife on her husband, and not her gender, that is responsible for the sexual division of labor. The implication of this doctrine is that once women achieve economic independence, gender equality in the home will follow automatically.

From the perspective of liberal feminism, women's legal rights to property, including property in their own person, is the first step in the emancipation of women. Liberal feminism broadly accepts the proposition that contemporary advanced industrial societies are meritocratic, and that women as a group are not innately less talented or less diligent than men but are denied opportunities simply because of their sex. It shares with neoclassical economics the belief that discrimination distorts the market mechanism and produces undesirable outcomes, while disagreeing with the neoclassical economists' presumption that discrimination is an infrequent occurrence. Liberal feminism views wage discrimination and sex segregation of occupations in the labor market as a mechanism designed to exclude women from the labor market. Liberal feminist doctrine lies behind much of the legislation on antidiscrimination and equal opportunity. Discrimination against women, claims liberal feminism, often arises from stereotyped expectations.

These gender stereotypes and prejudices are often learned in the family. Like structural functionalist theory in sociology, liberal feminism sees the family as an institution that specializes in socialization. Children learn gender roles in the family, using role models available to them that are 'reinforced' by the wider society. Characteristically, liberal feminism implies that the inequitable influence of gender roles can be reversed by simply giving girls better role models, establishing equal opportunity in education, and encouraging high achievement aspirations. The mother's role in the traditional domestic division of labor models poor achievement aspiration to daughters but it is not itself a source of constraint. The fundamental gender neutrality of the domestic division of labor, which liberal feminism presumes, can be justified on two distinct bases, either on the claim that the current organization of the family is an economically rational specialization of roles, or on the basis of the exchange theory of power.

Households, according to economist Gary Becker (1981), allocate their resources of time and money rationally between the spheres of market work, non-market work, and leisure. Individuals specialize in those tasks for which they have a 'comparative advantage'; a decision based on relative productivities. Becker claims that small biological differences, such as lactation, give women an advantage in child rearing. Men become more productive at market work because gender specialization leads them to invest in their capacity to earn (human capital), through education and on-the-job experience. This leads to a situation where men's wages are generally higher than women's,

and any extra allocation of men's time to housework would oblige them to forgo a greater amount of income. By contrast, the 'shadow price' or opportunity cost of women's time spent in unpaid work at home is lower, because the cost of income forgone is smaller. This somewhat circular reasoning leads to the claim that men have a 'comparative advantage' in paid work, while women have a 'comparative advantage' in unpaid work. Because each member specializes in activities where their advantage (productivity) is greatest, and they pool and exchange the rewards of these separate activities, the household is said to benefit from 'gains of trade.' Assuming an equitable distribution of these gains, the household that specializes is collectively better off than the one that does not. The asymmetry in the allocation of domestic work in the family is seen as a freely chosen and rational response to the biological differences plus inequalities in the labor market. The contribution of discrimination to gender inequality in the labor market is de-emphasized, since neoclassical economic theory implies that discrimination should erode in competitive markets. The allocation of domestic tasks will become symmetrical if the barriers to women's economic emergence are removed, and their labor market situation becomes similar to that of men.

Sociological exchange theory does not require the assumption that everyone in the household strives to increase the household's joint welfare. On the contrary, it assumes that whoever brings the most valuable resources to a marriage can expect to be able to bargain for some kind of compensation in the form of release from menial activities. In the traditional family, the wife provides domestic labor in return for economic support. The fact that housework and child care are done by women follows, according to exchange theory, from her inability to contribute socially recognized resources, such as money, to the relationship. In other words, responsibility for housework and child care falls to her because she contributes significantly less money rather than because she is a woman. Improve the labor market position of women, argue liberal feminists, and the inequities in the domestic division of labor will disappear.

If liberal feminists allow for any 'stickiness' in patterns of family organization this is attributed to a process of lagged communication and the learning of new sex role expectations. Women must acquire the techniques of assertiveness that allow them to negotiate their relationships with their partners. Men must learn to accept new domestic responsibilities and become 'new fathers.'

3. Other Forms of Feminism

Other forms of feminism view the family as a key site in the determination of gender equality. Nonliberal

feminists have drawn attention to how the privatization of family life contributes to the subordination of women in two key respects. First, women's unpaid work around the home is private, unseen, and devalued. Moreover, the pattern of the distribution of resources is hidden from view. Second, by definition, privacy screens family relations from view, allowing domination, including physical forms of coercion, to displace rule by consent. In addition, modern family relations have been seen as the crucible in which gender identity is forged.

Since the 1940s it has been popular to view the family as a 'haven in a heartless world'—a secluded refuge from the ruthlessly impersonal world of market relations. However, according to its critics, this liberal view of the family as refuge relies on two propositions: the separation of home and work, and the idea that only market work is work.

Nonliberal feminists challenged the presumption of economics (before the development of Becker's new home economics)—that the family household is purely a site of consumption—pointing to the valuable stream of goods and services produced in the home. In place of the separation of home and work, feminists substituted the distinction between paid and unpaid work. In the traditional division of labor, the male provider returns from market work ('a hard day at the office') to seek comfort and rest at home. For his homemaker spouse, however, the home is the site of her work.

Just how nonmarket work (performed in private homes) became virtually invisible, even to social scientists, would make a fascinating study. It is possible, however, to make explicit some of the social conditions that make this illusion possible. In pre-modern times, households were largely self-sustaining (consuming mostly goods produced at home) and most work was nonmarket work. However, as soon as market exchange becomes the dominant principal of economic organization, most things start to be bought with money. A labor market develops where people buy and sell the capacity to work for a stipulated amount of money (a wage). Money becomes the universal measure of value. Since mothers and housewives receive no pay for their work, it is seen as valueless, as not actually work at all. By the middle of the twentieth century the idea of housework as an instrumental labor process had disappeared so thoroughly that most celebrated social theorists of the time thought that wives and mothers specialized in 'expressive' behavior (Parsons and Bales 1955).

In contrast to the gender-neutral theories of the allocation of unpaid work favored by liberal feminists, other varieties of feminism place great emphasis on the fact that unpaid work is allocated on the basis of gender, so that housework and child care continue to be 'women's work.' Women, not men, interrupt their careers to raise a family. Indeed, men with families are more devoted to careers than single men. Studies of the time devoted to unpaid work show that women

still do an overwhelming proportion of laundry, cleaning, cooking, and the physical care of children. The same studies show that men continued to specialize in mowing the lawn and polishing the car, and their chief contact with children is in the context of play.

Even when both partners are in full-time employment the sex segregation in unpaid work tasks does not diminish. The idea that gender relations within the family make it difficult to reassign the sex-stereotyped domestic responsibilities lies behind the anxiety about women performing a 'second shift' (Hochschild and Machung 1989).

Although responsibility for key domestic tasks is allocated to women, all members of the family enjoy the benefits of domestic work—live in clean dwellings, eat home-cooked meals, wear laundered clothes, receive emotional nurturance, etc. Although society as a whole depends on the reproduction of acculturated individuals, the costs of the activity of mothering, such as loss of lifetime earnings, reduced autonomy, and greater exposure to poverty, predominantly are borne privately by mothers.

The recognition of these facts about the distribution of domestic labor brings under suspicion the key presumption of the 'rational' specialization thesis, namely joint family welfare. These suspicions are reinforced by research into the relationship between women's share of family income and consumption of market goods. If families work to maximize their common (unitary) utility and pooled resources, then their consumption will be determined by total family income alone, and not by who earns or controls income. However, there is mounting evidence that, independent of the amount of total family income, increasing women's share of family income changes family expenditure patterns. The increasing women's share of family income results in greater expenditure on goods and services preferred by women (women's and children's clothing, restaurant expenditures) (Lundberg et al. 1997). Taken together, this evidence suggests that neither the sexual division of labor nor the distribution of market and nonmarket goods and services is consensual.

This raises the prospect that the family as an institution is constituted by systematically unequal exchange between the genders. Indeed, significant numbers of feminists believe that men as a social group actively or passively resist the reallocation of responsibility for domestic tasks because, as beneficiaries of this unequal exchange, it is not in their interests to change (Delphy and Leonard 1992).

Recognition of resistance leads feminists to explore the basis of capacity to resist, or to say the same thing in different words, to seek out the sources of male power. While liberal feminists tend to see men's dominance of the public sphere as the key to their power, radical feminists have often sought the basis of male power in family relations.

A simple, and long neglected, form of male power is physical coercion. The rise of second-wave feminism toward the end of the 1960s was accompanied by a growing recognition of the shocking extent of domestic violence. The significance of this kind of coercion is probably greater than the recorded incidence of violent acts, because in most social systems it is the threat of violence (terror) rather than direct acts of physical coercion that intimidates people into acquiescence. Susan Brownmiller used a similar logic when she claimed that rape is 'a conscious process of intimidation by which all men keep all women in a state of fear' (1975, p. 15). Emphasizing the crime of rape focused attention on the fact, a legacy of classical liberal thinking, that in most jurisdictions rape within marriage was not a crime in the middle of the twentieth century.

Domestic violence, and rape within marriage in particular, is so unexpected (even shocking) because they contradict the notion that family relations are relations between consenting equals. Much of the debate around sexual harassment and date rape also centers on the issue of consent. In classical liberal terms, it has become natural for us to allow the standards of liberty and equality developed as principles applying only to the public sphere to spill over to the private sphere. Many writers argue that the contemporary family is based on the progressive democratization of intimate relations.

Most social scientists believe that few systems of power can rely on coercion alone, and that the routine exercise of power must be seen as somehow legitimate. Sociologists believe that power is most successful when it becomes domination, that is, when it operates with the apparent consent of the dominated. Among the ideas that can be used to legitimate domestic inequality, a few have received the most consistent theoretical attention. The first is to present it as consensual and beneficial. Opponents of the 'rational' specialization thesis argue that this doctrine serves as an apologia for patriarchalism at home. The second focus has been the devaluation of women's work. Devaluation generally consists of diminishing the perceived importance of something, failing to recognize its true worth. The extreme case of devaluation is complete invisibility—which was a notable feature of the treatment of household productive activities as 'expressive' rather than 'instrumental.' The third idea that has received attention is the association of domestic work with femininity and its obverse, the association of masculinity with market work. Gender identity, many claim, is a major support for patriarchalism in the family and a significant obstacle to the achievement of gender equity.

The idea that the activities typically associated with women have been devalued has led to both attempts to revalue women's work, for example, through comparable worth campaigns, and attempts to reconsider the relationship between equality and difference. At

one level this has involved positive emphasis on 'women's values' (sensitivity, caring, and nurture) in opposition to the male values of competition, self-centeredness, aggression, and violence.

The celebration of difference has often been accompanied by an intellectual exploration of the socially constructed nature of gender identity which has taken priority over an analysis of family relations. Theories about language, and more broadly discourse, have been used to argue that signs used to represent things are arbitrarily self-referential. Self-referential because the meaning or one word is defined in terms of its relationship to other words—for example, 'light' is the opposite of 'dark,' while 'many' is defined against the background of the concept of 'one.'

It follows from this approach that the most powerful thing a social analyst can do is to deconstruct the discursive practices through which the social world is portrayed. The political objects of these more recent postmodern varieties of feminism have been representations of women and men, their bodies and relations in the various media of cultural transmission. Feminist theory has moved away from a preoccupation with things to emphasis on the salience of words and other cultural symbols.

Of course, the method of deconstruction can and has also been applied to the texts of feminist theory itself. This has made it problematic to assume that a political movement, with common objectives, can be rooted in a single category called 'women.' There has been an emerging sensitivity to the idea of difference among women, especially between 'women of color' and 'white middle-class feminists.' Hooks (1984) has argued that, for black American women, the family has been a source of solidarity and resistance to white racism and therefore less oppressive to black women than white American women. This emphasis on difference has been combined with a hostility derived from postmodernism to grand narratives or 'totalizing' conceptions. Generally, this has resulted in a political agenda that celebrates multiple identities, plastic sexuality, and diversity of household types ranging beyond the currently known groupings of family household.

From this perspective, sexual identity is an arbitrary cultural or social product. Gender identity, however, is seen as among the elements that contribute, in the deepest sense, to the conception of what one is. In 1978 Nancy Chodorow drew on a variant of Freudian psychodynamic theory (object relations theory) to explain gender differentiation and the different personality traits of adult men and women. Chodorow argued that because women are currently responsible for the primary care of children, this situation creates deeply internalized, unconscious desires in sons and daughters that reproduce male dominance and female mothering. When women 'mother,' all children have an initial attachment to and identification with a female. Because boys have a carer of a different sex,

achieving gender identity encourages emotional separation from the mother. Individuating from a female carer encourages boys to define their masculine gender identity that emphasizes difference from and rejection of femaleness, identification with emotionally absent (non-nurturing) males. For girls, raised by a same-sex carer, gender identity and individuation are not as mutually reinforcing. Thus, males grow up with an internalized need to emphasize separation and with misogynistic tendencies to define themselves in opposition to whatever is female; females grow up with a desire for emotional connection, including mothering. Chodorow draws the heartening conclusion that if men were to 'mother,' and therefore become more like women, this would change the masculine psyche. Chodorow's proposal to create greater gender equity ultimately turns on reforming the nurturing relations in the family.

Interest in the formation of gender identity, especially masculinity, has increased since Chodorow wrote *The Reproduction of Mothering*, using a different framework for understanding the process which places more emphasis on the role of culture and less on family relations. The most prominent alternative approach is social constructionism. In social constructionist variants, 'gender is not fixed in advance of social interaction, but is constructed in interaction' (Connell 1995, p. 35). It encourages study of the construction of masculinity in everyday life, acknowledges the existence of significantly different masculinities, and emphasizes the contradictory and dynamic nature of gender. This framework presupposes that Western culture typically valorizes the masculine over the feminine. The sexual division of labor at home has been viewed as a scaffold for the construction of gender identities.

4. Conclusion

The relationship between families and gender continues to be the subject of dispute in the social sciences. The 'linguistic turn' in the study of gender relations has directed attention to the most general cultural determinants of the gender identity of family members. The economic significance of nonmarket work has directed fresh attention to the process of home production. The classical liberal values of freedom and equality have proven themselves to be irrepressible, and although the split between the public sphere and the private sphere has endured, it has proven difficult to limit the demands for equality purely to the public sphere.

See also: Dissolution of Family in Western Nations; Cultural Concerns; Divorce and Children's Social Development; Divorce and Gender; Divorce, Sociology of; Families as Educational Settings; Family and Kinship, History of; Family and Psychiatric

Illness/Family Therapy; Family, Anthropology of; Family as Institution; Family Bargaining; Family Law; Family Processes; Family Systems and the Preferred Sex of Children; Family Theory: Feminist-Economist Critique; Family Theory: Role of Changing Values; Family Therapy, Clinical Psychology of; Fatherhood; Feminist Legal Theory; Feminist Theory: Psychoanalytic; Gender and Feminist Studies in Anthropology; Gender and Reproductive Health; Gender and the Law; Gender Ideology: Cross-cultural Aspects; Gender-related Development; Gender Role Stress and Health; Lone Mothers in Affluent Nations; Lone Mothers in Nations of the South; Love and Intimacy, Psychology of; Marital Interaction: Effects on Child Development; Marriage and the Dual-career Family; Cultural Concerns; Masculinities and Femininities; Men's Health; Motherhood: Economic Aspects; Motherhood: Social and Cultural Aspects; Parenthood and Adult Psychological Developments; Parenting: Attitudes and Beliefs; Parenting in Ethnic Minority Families: United States; Partnership Formation and Dissolution in Western Societies; Repartnering and Stepchildren; Rural Industrialization in Developing Nations and Gender

Bibliography

- Becker G S 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA
- Brownmiller S 1975 *Against Our Will: Men, Women and Rape*. Penguin, Harmondsworth, UK
- Chodorow N 1978 *Reproduction of Mothering: Psychoanalysis and the Sociology of Gender*. University of California Press, Berkeley, CA
- Connell R W 1995 *Masculinities*. Allen & Unwin, Sydney
- Delphy C, Leonard D 1992 *Familiar Exploitation: A New Analysis of Marriage in Contemporary Western Societies*. Polity Press, Cambridge, UK
- Eisenstein Z R 1981 *The Radical Future of Liberal Feminism*. Longman, New York
- Hartmann H I 1981 The family as the locus of gender, class and political struggle: The example of housework. *Signs* 6(3): 366-94
- Hochschild A, Machung A 1989 *The Second Shift: Working Parents and the Revolution at Home*. Viking, New York
- Hooks B 1984 *Feminist Theory: From Margin to Center*. South End Press, Boston
- Lundberg S J, Pollak R A, Wales T J 1997 Do husbands and wives pool their resources? Evidence from the United Kingdom Child Benefit. *Journal of Human Resources* 32(3): 463-80
- Parsons T, Bales R F 1955 *Family, Socialization and Interaction Process*. Free Press, Glencoe, IL
- Pateman C 1989 *The Disorder of Women: Democracy, Feminism and Political Theory*. Polity Press, Cambridge, UK

M. Bittman

Family and Kinship, History of

This article discusses the emergence and development of the historical study of the family and interprets them in the larger context of social history. It addresses the following issues: the emergence of the field of family history; household structure; developmental perspectives; kinship; the family's interaction with the process of industrialization; family strategies; and future directions in the field of family history.

1. *Emergence of the Field*

Recent historical research on the family has revised some widely held myths about family life in the past, as well as generalizations about the impact of the grand processes of social change on the family and society. Family history has complex roots in both the historical demography of the early 1960s and the 'new social history' of the same period. Particularly in the United States, it has shared with the latter a commitment to reconstructing the family patterns of ordinary people and viewing them as actors in the process of change. Contemporary historians of the family have sought to reintroduce human experience into historical research and to emphasize the complexity of historical change (see Hareven 1971, 2000a, Tilly 1987). Hence, research in the history of the family explores previously neglected dimensions of human experience such as growing up, courting, getting married, bearing and rearing children, living in families, becoming old and dying. The formidable goal has been to understand the family in various contexts of change and at different points in historical time. In short, the historical study of the family represents an effort to understand the interrelationship between 'individual time,' 'family time,' and 'historical time' (Elder 1978).

Before systematic historical research on the family began, various social science disciplines had generated their own myths and grand theories about continuities and changes in family behavior in the past. Sociologists in particular argued that, in preindustrial societies in the west, the dominant household form had contained an extended family, often involving three co-resident generations, and that the 'modern' family—characterized by a nuclear household structure, family limitation, the spacing of children, and population mobility—was brought about by industrialization. Associated with these generalizations was also the popular myth that industrialization broke up traditional family patterns and community life. By providing a perspective on change over time, as well as on family behavior within specific social and cultural contexts in discrete time periods, historical research on the family has led to the rejection of these assumptions and to the resulting questioning of the role of industrialization as a major watershed in this field for

American and European history (Wrigley 1972, Goode 1963).

Over the three-and-a-half decades of its existence, family history has moved from a limited view of the family as a static unit at one point in time to an examination of the family as a process over the life course of its members; from a study of discrete domestic structures to the investigation of the nuclear family's relations with the wider kinship group; and from a study of the family as a separate domestic unit to an examination of its interaction with the worlds of religion, work, education, correctional and welfare institutions, and with processes such as migration, industrialization, and urbanization (Vinovskis 1977, Hareven 1982).

As research in this field developed, new findings and approaches have led to the revision and augmentation of the pioneers' approaches. Efforts to explore decision-making processes within the family have led to an investigation of strategies and choices that individuals and family groups make. The life-course approach has added an important developmental dimension by focusing on age and cohort comparisons in ways that link individual and family development to historical events. Research also expanded chronologically to cover antiquity, and geographically to encompass southern Europe, eastern Europe, the Mediterranean, Japan, China, and Latin America, in addition to Western Europe, North America, and Japan. The cumulative impact of studies in the history of the family has been to revise simplistic views of both social change and family behavior (Reher 1997, Andorka 1994, Plakans and Wetherell 1999, Todorova 1997, Halpern et al. 1996, Hayami and Uchida 1972, Kuroso 1996, Lee and Campbell 1997, Lee and Feng 1999, Borges 1992, Kuznesof 1997).

The emergence of the history of the family as a special area of inquiry received its major impetus from the publication of Philippe Aries's *Centuries of Childhood* (1960 in French and 1962 in an English translation). Aries argued that childhood as we know it originated only in the early modern period and that its discovery was closely linked to the emergence of the 'modern' or conjugal family, in which parents' private relationships with their children were more important 'than the honor of a line, the integrity of an inheritance, or the age and permanence of a name' (Aries 1962, p. 393). By linking the 'discovery' of childhood to transformations in family and social structures as well as economic and demographic changes, Aries inspired a whole generation of scholars.

Along another line of research, historical demographers in France in the early 1960s provided family historians with a powerful 'new weapon' for the analysis of demographic processes. Louis Henry and Pierre Goubert had developed a family reconstitution technique in the 1950s. Using genealogies, and marriage, baptismal, and death records from parish registers, this method enabled demographers to recon-

struct aggregate patterns of fertility, nuptiality, and mortality for vast numbers of people and, in some instances, over several generations (Henry 1953).

From the 1960s on, historical demography and family history in France developed into two parallel but interrelated streams. One continued to concentrate on demographic analysis, along the lines of Henry and Goubert; the other, influenced by Aries, anthropology, and the French social history tradition, integrated demographic analyses with patterns of family and sexuality, linking community and social and cultural variables with *mentalité*, as exemplified in the work of Emmanuel LeRoy Ladurie (1978), André Burguière (1987), Jean-Louis Flandrin (1976), and Martine Segalen (1991), among others. Family reconstitution subsequently became a powerful tool in the hands of the Cambridge Group for the History of Population and Social Structure.

Established in 1964, the Cambridge Group adapted the family reconstitution method to English parish registers, while also pursuing analysis of seventeenth-century nominal household registers. Wrigley found in the decline in fertility in the seventeenth century and in its recovery in the eighteenth century, evidence that rural births and marriages responded to changing economic conditions. He concluded that the demographic transition involved a change 'from a system of control through social institution and custom to one in which the private choice of individual couples played a major part in governing the fertility rate' (Laslett 1977, Wrigley and Schofield 1981).

Such demographic analyses for France and England, and subsequently Sweden, Austria, Germany, and The Netherlands, revealed that, in the pre-industrial period, people married later than had been generally assumed, couples practiced some form of family limitation and child spacing as early as the seventeenth century, households were predominantly nuclear rather than extended, and preindustrial populations experienced extensive geographic mobility (Laslett 1965, Goubert 1977). Late marriage served as a method of family limitation. It was also closely related to the expectation that a newly wed couple would establish a separate household and be self-supporting, while continuing to contribute to their families of origin. Linking late age at marriage to the nuclearity of the household, John Hajnal (1965) developed his model of the 'West European Marriage Pattern,' which served as the basic model for the analysis of West European families until it came under criticism. Hajnal claimed that marriage patterns in historic Europe could be differentiated with a line drawn approximately from St. Petersburg on the Baltic Sea to the Adriatic Sea. West of this line marriages tended to be relatively late (late 20s) for both sexes, and substantial portions of the population did not marry at all. East of this line, marriage for both sexes was relatively early (late teens) and marriage was nearly universal. He identified the 'western' pattern as

the 'European' pattern and suggested that the 'eastern' pattern was shared by many developing countries (in the 1960s).

2. Household Structure

Using nominal census records, Laslett found evidence of continuity in nuclear household structure in England, at least since the sixteenth century. A landmark volume published by Laslett and Wall (1972), containing articles on household structure for various countries, including Japan, concluded that pre-industrial households in Western Europe and North America affirmed these findings and dispelled the previously held assumption that industrialization brought about a nuclear family form. Although Laslett's early work emphasized the persistence of a nuclear 'family,' the major unit discussed was the household, not the family. A household is not identical with a family, since the domestic group may include nonrelatives as well. Nor is a 'family' restricted to the household, since extended family ties transcended the household unit (Goody 1972, Hareven 1982).

Subsequent studies showed that co-residence with extended kin tended to increase, not decrease, after the 'industrial revolution,' because of the need of newly arrived migrants to industrial cities to share housing space (Laslett and Wall 1972, Hayami and Uchida 1972, Anderson 1971). David Herlihy (1987) and Richard Smith (1979), respectively, traced a nuclear household structure in medieval Tuscany back to the twelfth century, as well as in England; and, in the medieval village of Montailou, LeRoy Ladurie (1978) found a variety of forms of family structures in the *domus* (household), ranging from ordinary nuclear families to various complex configurations. Household structure varied over the life course of its members, in relation to the economy and migration.

Such early discoveries also led to the formation of new generations and stereotypes: by combining the findings about the nuclearity of the household with the Hajnal thesis on the Western marriage pattern, Laslett developed a typology of the 'western family,' which he characterized as having a simple nuclear household; a relatively late child-bearing age; a relatively narrow age gap between husband and wife, with a relatively high proportion of wives older than their husbands; and the presence in the household of 'life cycle' servants unrelated to the family with whom they were residing. By contrast, he characterized the South European and Mediterranean family type as having complex household structures and early marriage, in which nuclear households are formed by the breakup of extended ones, rather than through marriage (Laslett 1977). These typologies, along with the Hajnal thesis, later came under criticism. David Kertzer refuted the thesis that links early female marriage to joint household organization as a characteristic of the

'South European' family pattern. He showed variability in household structures in Italy, depending on types of landholding and sharecropping. Similarly, Haim Gerber, on the basis of his analysis of urban Turkish households in the seventeenth century, questioned stereotypes of the traditional 'Middle Eastern family' (Kertzer and Hogan 1989, Gerber 1989).

In New England, when interpreted in the context of economic and social institutions (such as landholding, inheritance, religion, community structure, and religion), demographic patterns served as the backbone of complex family and community analyses. In his analysis of household and family structure in colonial Andover, Massachusetts, Philip Greven revealed the power relations between fathers and sons and the limits of the sons' autonomy based on inheritance. He found that households were primarily restricted to members of the nuclear family and nonrelatives, but a 'complex web' of family connections reinforced by proximity of residence, often on the same land, permeated the community (Greven 1970). Demos, in his study of colonial Plymouth, Massachusetts, used family reconstitution, the analysis of nominal census records and a great variety of other sources to reconstruct family and community patterns (1970). He also found a predominance of nuclear households, but these differed considerably from contemporary ones. Families were larger then and included nonrelatives, such as servants, apprentices, and boarders and lodgers in the households. Age configurations were also different; in some families, the oldest child in the household would be an adult while the youngest was still at the breast.

A major critique of early studies of household structure was their reliance on 'snapshots' in the household schedules listed in the census at one point in time. On the basis of his reconstruction of peasant households in the villages of Heidenreichstein in northern Austria in 1763, Berkner (1972) provided evidence that household structure changed several times over the course of the family's life, as it moved through its cycle—from a stem family to a nuclear one, and later to a stem family again, in relation to inheritance. Households were not static. At the same time, Demos' reconstruction of the stages of the life cycle in Plymouth demonstrated dynamic movements of individuals within and across families and households, in relation to the religious and cultural prescriptions of Puritan society (Demos 1970).

The presence of 'life cycle' servants—young men and women in their teens who lived and served in other people's households in the transitional period between leaving home and marrying—which Laslett found in English villages, as well as the presence of other unrelated individuals in the household, proves the remarkable flexibility of households in the past. Nuclear households included nonrelatives such as servants, apprentices, boarders and lodgers. Similarly, in nineteenth-century American urban households,

Modell and Hareven (1973) found a high proportion of boarders and lodgers, who provided important substitutes for children who had left home. Households expanded and contracted over the life course and in accordance with the family's needs.

3. *Developmental Perspectives: The Life Course*

The discovery of boarders and lodgers in nineteenth-century households and of the overall fluidity of household structure led historians of the family to search for a framework that would allow them to capture the movement of individuals through various household forms over their lives and the changes in the composition of the family and the household under various historical conditions. The family cycle model, which was used originally to identify changes in family structure, became unsatisfactory, because the stages of the family cycle were derived from contemporary American middle-class families, and were *a priori*. As Glen Elder explained, the family cycle identifies stages of parenthood rather than the more dynamic aspects of individual transitions into and out of various family roles (Elder 1978).

Dissatisfaction with the family cycle model led historians and sociologists to develop the life-course approach, which has introduced a dynamic dimension into the historical study of the family, and has moved analysis from a simplistic examination of stages of the family cycle to individuals' and families' timing of life transitions in relation to historical time (Hareven 1978, Elder 1978). According to the life-course paradigm, the pace and definition of the timing of life transitions, such as starting work, leaving home, and getting married, are determined by their social and cultural context. On the familial level, timing involves the synchronization of individual life transitions with collective family ones (Elder 1978). For example, using both demographic evidence and prescriptive literature, Modell, Funstenberg, and Herschberg found a major historical change in American society over the past century, from the 'erratic' timing of life transitions governed by collective family needs to more individualized timing in conformity with age norms (Modell et al. 1976).

Life-course research also illuminates the links between behavior and perception. While the actual timing of life transitions can be reconstructed from demographic records, its meaning to the individual and family members undergoing these transitions hinges on the examination of qualitative, subjective sources (Neugarten and Hagestad 1976, Bourdelais 1993, Hareven and Masaoka 1988, Hareven 2000b). A life-course perspective has also made an important contribution to the study of kinship by directing attention to the changing configurations of kin with whom individuals 'travel' together over their lives. Such configurations are formed and re-formed; they

change in their composition, in their relationship to the individual and the nuclear family, and in relation to each other over the life course.

4. Kinship

During the second decade of the development in the field of family history, kinship ties outside the household received increasing attention. Rather than being isolated, nuclear households were embedded in extended kinship ties outside their confines. Members of the nuclear family were engaged in various forms of mutual assistance, collaboration, and rituals with extended kin. Even though aging parents did not generally reside in the same household with their married adult children, they lived in the vicinity, often on the same land (Greven 1970, Berkner 1972, Smith 1979, Sabean 1998).

For nineteenth- and twentieth-century urban populations, Anderson (1971) and Hareven (1982), respectively, have documented the central role of family members and more distant kin in organizing migration from rural areas to industrial cities, in facilitating settlement in urban communities, and in helping migrants adapt to new working and living conditions. In her study of late-nineteenth- and twentieth-century textile workers in Manchester, New Hampshire, Hareven found that kinship networks in communities of origin were reinforced by the back-and-forth migration of individual members, and the transfer of resources. Following 'chain migration' routes, villagers who went to work in industrial communities spearheaded migration for other relatives by locating housing and jobs. Those who remained in the communities of origin often took care of aging parents and other relatives who stayed behind (Anderson 1971, Hareven 1982). Hareven found that despite the scale of the modern factory system, workers' kinship networks cushioned the adaptation of immigrant workers to new industrial working conditions. Kin acted as brokers between the workers and their industrial employers. Rather than disrupting kinship ties, migration often strengthened them and led to the development of new functions for kin in response to changing economic and employment conditions in response to the requirements of the industrial system (Hareven 1982).

5. The Family's Interaction with the Process of Industrialization

An examination of the family's interaction with the process of social and economic change enables us to understand more precisely what occurred internally within the family and how such changes were accomplished on a societal level. It provides new insights into the process of industrialization and urbanization,

into how labor markets functioned, and how industrial work processes and labor relations were organized. Historical research on the family has led to the rejection of simplistic models of social change, especially of modernization theory. How the family both initiates change and adapts to it and how it translates the impact of larger structural changes to its own sphere are crucial questions for the understanding of the process of social change.

The important principle underlying these questions is a view of the family as an active agent. The family planned, initiated, or resisted change; it did not just respond passively. Historical research over the past two decades, highlighting the family's role as an active agent, has led to revising questions such as 'What was the impact of industrialization on the family?' to 'What was the family's impact on the industrial system?' (Hareven 1982, c.f. Goode 1963). The question has been further cast into a more complex form: Under what circumstances was the family more able to control its destiny and to affect the larger social processes, and under what circumstances did the family succumb to declining markets, changing modes of production, business cycles, and other external forces?

The family had a central role in charting the transitions to various modes of industrial production by following diverse routes: from a rural economy to a proto-industrial or industrial one, or from a rural economy directly to an industrial one. In all these cases, while following its own priorities, when possible, the family facilitated the advance of industrialization by releasing the labor force needed for the newly developing factories and by organizing migration to industrial centers (Smelser 1959, Tilly and Scott 1978, Quataert 1985). The family's interaction with the industrial system has received attention from studies of family production systems in artisanal and proto-industrial settings and the study of family and work in the context of large industrial enterprises.

Proto-industrialization was a household-based form of production in the countryside and in some urban areas in England, France, Belgium, Switzerland, and Austria, from the seventeenth to the nineteenth centuries that preceded the 'industrial revolution.' It was characterized by the production of goods in local cottages for a capitalist employer who sold the products in external markets. Rudolph Braun (who first identified these patterns in Switzerland) and other historians developed a model of demographic behavior they claimed was typical of 'proto-industrial' families. This consisted of earlier age at marriage, higher fertility, and delays in children's departure from the home, all of which were conducive to maximizing a family labor force. Marriage formed the key to the spread of the cottage industry, because it led to the formation of a new nuclear family unit, which was also a 'unit of labor' (Braun 1974, Mendels 1972, Medick 1976).

In the 1980s and 1990s, the demographic model of proto-industrialization has come under criticism. On the one hand, scholars have questioned whether demographic behavior associated with proto-industrialization had been sufficiently widespread to warrant such generalizations. They found that demographic behavior in response to new economic conditions 'differed according to the environment (i.e., land ownership, social structure, and inheritance) and the exact nature of the economic change which took place' (Gutmann and Leboutte 1984, p. 589). Regional variation, interaction with the local agricultural structure, as well as the type of cottage industry (whether weaving or spinning), made a difference (Mitterauer 1992).

Like research on the family in proto-industrial regimes, the study of family in the factory system has provided considerable evidence documenting the family's role as an active agent in its interaction with the process of industrialization. Such research effectively undermined the theory of 'social breakdown' that had haunted sociology and social history. According to this notion, migration from rural to urban centers had led to the uprooting of people from their traditional kinship networks. The pressures of industrial work and urban life caused a disintegration of the family unit, and adaptation to industrial life stripped migrants of their traditional culture (Parsons 1955, Wirth 1938).

Historians and sociologists alike have challenged the theory of social breakdown, by providing contrary evidence. Neil Smelser, in his pioneer study of the family in the early stages of industrialization in Britain, documented the recruitment of entire family groups as work units in the early textile factories. Fathers contracted for their children, collected their wages, and were responsible for their performance in the factory. Entire families depended on the factory as their employer; the factories, in turn, depended on the recruitment of family groups to maintain a continuous labor supply (Smelser 1959). Anderson documented a continuity of recruitment in family units and in the vital role of kin among workers in the textile industry in Lancashire, England, at least through the middle of the nineteenth century. The practice of family members working together in the factory was carried over into the United States, as the factory system developed there, and was still present among immigrant workers in the twentieth century. The very success of the early industrial system depended on the recruitment of labor along kinship lines (Anderson 1971, Hareven 1982). As mentioned above, Hareven's study of family and work in the Amoskeag Mills in Manchester, New Hampshire, has documented the active role of the family and the kinship group in relation to the industrial corporation. It showed that the family type most 'fit' to interact with the factory system was not an 'isolated' nuclear family, but rather one embedded in extended kinship ties (Hareven 1982).

The functions that kin fulfilled in their interaction with the factory system were not merely archaic carryovers from rural society but a selective use of pre-migration patterns of kin assistance patterns in response to the needs dictated by industrial conditions. The encounter of immigrant workers with the modern factory system led neither to the abandonment of pre-migration family traditions, nor to a rigid adherence to them. Rather, the workers adapted their customs and social organization to the new conditions they confronted. In doing so, they addressed the factory system on its own terms. The family selected those aspects of its pre-migration culture that were most useful in coping with industrial conditions and adapted them to new needs. These patterns of selectivity called into question the linear view of social change advanced by modernization theory. Family behavior did not modernize at the same pace as workers' conduct in the factory. The family was both a custodian of tradition and an agent of change (Hareven 1982).

6. *Family Strategies*

To better understand the family's role in these processes of social change, it is necessary to examine how families charted their strategies in relation to external opportunities and constraints. Family strategies involved both the decisions individuals or families made and the actual timing of such decisions in relation to opportunities and constraints. Strategies included, at times, calculated trade-offs in order to find employment, achieve solvency, buy a house, facilitate children's education or their occupational advancement, control or facilitate a child's marriage, save for the future and provide for times of illness, old age, and death. Strategies were part of a larger life plan. As people encountered new circumstances, they modified and reshaped their plans and strategies in the context of their own culture and traditions (Hareven 1990).

As research on family strategies developed further, the concept also underwent important revisions. Under the influence of feminist historians, the question was raised: 'Whose strategies are family strategies?' This question is still part of an important research agenda about the process of decision making within the family, particularly concerning women's participation. Recent research suggests, though, that there may have been a considerable variation along gender-lines and among family members in the forging of strategies.

7. *Future Directions*

The multidimensional research effort in family history has produced impressive results and identified new

lines of investigation. Research still needs to pursue established topics that have not received sufficient attention, including the family's relationship to social space and a more systematic study of the family's relationship to religion, the state, and the legal system.

There is a need for more systematic studies of the 'edges' of family life, such as solitary individuals, orphans and foundlings, and the process of family breakdown through divorce or death and of family reconstitution through remarriage. A second direction is the forging of more systematic linkages between demographic behavior and household structure with internal family dynamics; a closer integration of the study of the household with nonresident kin; a more careful linkage of kinship and household patterns with various processes such as work and migration; and the relationship between demographic and family patterns with cultural dimensions.

In addition to this specific agenda, the larger questions of historical changes in the family over time still need to be explored systematically in several directions: First, if industrialization did not cause the emergence of 'modern' family behavior in Western Europe and the United States, what did cause it? One plausible explanation that has been advanced is the commercial revolution of the sixteenth and seventeenth centuries. This does not explain, however, the existence of nuclear household patterns in the Middle Ages. While Stone (1977), like Aries, dated the emergence of the 'modern family' in the late seventeenth and early eighteenth centuries, Shorter placed it in the late eighteenth, and early nineteenth centuries. Carl Degler (1980) dated the emergence of the American modern family in the late eighteenth or early nineteenth centuries as well. Following in Aries's footsteps, Stone, Shorter (1976) and Degler have focused on the rise of 'affective individualism' as the major criterion of 'modern' family life. They generally agreed that the 'modern family' is privatized, nuclear, domestic, and based on emotional bonding between husband and wife, and between parents and children. The correlates of its emergence were the weakening influence of extended kin, friends, and neighbors on family ties and an isolation of the family from interaction with the community. There is some disagreement and, at times, lack of clarity about which class initiated these changes. Aries, Stone, and, more implicitly, Degler viewed the bourgeoisie and the gentry as the vanguard, while Shorter assigned a crucial role to peasants and workers.

Related to the emergence of the family as a private, intimate entity is the widely accepted interpretation that this transformation was based on the family's surrender of earlier-held functions of production, welfare, education, and social control to other institutions and its withdrawal into domestic privacy has become one of the standard clichés of family history and sociology. We need to understand the process by which these functions were transferred to other insti-

tutions. A systematic exploration of the family's relationship to public agencies and institutions—specifically institutions of education, welfare, and social control, after the transfer of its functions—will help historians escape the trap of viewing the family as an isolated domestic entity.

Historical studies in Europe in the 1990s have emphasized an integrated view of the family both as a private entity and an object of the state. *The History of Private Life*, under the general editorship of Philippe Aries and Georges Duby, which focuses primarily on French society, deserves particular attention. The fourth volume, *From the Fires of Revolution to the Great War*, edited by Michelle Perrot, presents the experiential history of the family in relation to public life. The authors view the family's private life and sexuality as inseparable from the state (Perrot 1990).

As further research continues to emerge, it becomes necessary to develop a comprehensive model of change in family behavior that does justice to its complexity. The most important dimension still absent from studies of long-term changes in the family involves more systematic distinctions by social classes, rather than generalizing for the entire society on the basis of the middle class. We need a more detailed understanding of the historical process by which other classes adopted middle-class family behavior, if indeed that was the case, and of what class differences have survived such a process. How did patterns of family behavior that first emerged in the middle class transfer to other classes in society and by what processes? Equally important, at the end of that process, had rural and working-class families really adopted middle-class family forms, or did differences persist underneath apparent similarities?

The question about when and how change takes place needs to be raised again. Various strands of change in family history still need to be traced individually to the point of their culmination as visible transformations. The process is rendered more difficult by the fact that not all the strands undergo change at the same pace. Since the family is not a monolith, different members within it may initiate change or accept change at different points. For example, women were agents of change in certain areas of family life, such as in the practice of family limitation, while men were innovators in others. Even children were innovators in certain areas. They may have brought literacy into the family, new perceptions of behavior they learned in school or new work habits and new technologies. Thus, when we examine the family's role in the process of social change, we need to differentiate among the respective roles of members within the family, in interacting with these processes. This is an important task for future decades.

See also: Divorce and Gender; Family as Institution; Gender History; Industrialization, Typologies and History of; Kinship in Anthropology; Labor History;

Labor, Division of; Life Course in History; Sociology, Epistemology of; Work: Anthropological Aspects; Work, History of

Bibliography

- Anderson M S 1971 *Family Structure in Nineteenth-century Lancashire*. Cambridge University Press, Cambridge, UK
- Andorka R 1994 The historical demography of a proper Hungarian village: At any in the eighteenth and nineteenth centuries. *Journal of Family History* 19: 311–31
- Aries E 1962 *Centuries of Childhood: A Social History of Family Life*. Baldick R (trans.). Knopf, New York
- Berkner L K 1972 The stem family and the developmental cycle of the peasant household: An eighteenth century Austrian example. *American Historical Review* 77: 398–418
- Borges D E 1992 *The Family in Bahia, Brazil, 1870–1945*. Stanford University Press, Stanford, CA
- Bourdelaïs P 1993 *La Nouvelle Age de la vieillesse: Histoire du vieillissement de la Population*. O. Jacob, Paris
- Braun R 1974 Early industrialization and demographic change in the Canton of Zurich. In: Tilly C (ed.) *Historical Studies of Changing Fertility: Quantitative Studies in History*. Princeton University Press, Princeton, IN
- Burguière A 1987 The formation of the couple. *Journal of Family History* 12: 39–56
- Degler C N 1980 *At Odds: Women and the Family in America from the Revolution to the Present*. Oxford University Press, New York
- Demos J 1970 *A Little Commonwealth: Family Life in Plymouth Colony*. Oxford University Press, New York
- Elder G H Jr 1978 Family history and the life course. In: Hareven T K (ed.) *Transitions: The Family and the Life Course in Historical Perspective*. Academic Press, New York
- Flandrin J-L 1976 *Families in Former Times: Kinship, Household, and Sexuality*. Cambridge University Press, Cambridge, UK
- Gerber H 1989 Anthropology and family history: The Ottoman and Turkish families. *Journal of Family History* 14: 409–21
- Goode W 1963 *World Revolution and Family Patterns*. Free Press, New York
- Goody J 1972 Evolution of the family. In: Laslett P, Wall R (eds.) *Household and Family in Past Time*. Cambridge University Press, Cambridge, UK
- Goubert P 1977 Family and province: A contribution to the knowledge of family structures in early modern France. *Journal of Family History* 2: 223–36
- Greven P 1970 *Four Generations: Population, Land, and Family in Colonial Andover, Massachusetts*. Cornell University Press, Ithaca, NY
- Gutmann M, Leboutte R 1984 Rethinking protoindustrialization and family. *Journal of Interdisciplinary History* 14: 587–607
- Hajnal J 1965 European marriage patterns in perspective. In: Glass D V, Eversley D E C (eds.) *Population in History*. E Arnold, London
- Halpern J M, Kaser K, Wagner R A 1996 Patriarchy in the Balkans: Temporal and cross-cultural approaches. *History of the Family: An International Quarterly* 1: 425–43
- Hareven T K 1971 The history of the family as an interdisciplinary field. *Journal of Interdisciplinary History* 2: 273–99
- Hareven T K (ed.) 1978 *Transitions: The Family and the Life Course in Historical Perspective*. Academic Press, New York
- Hareven T K 1982 *Family Time and Industrial Time: The Relationship Between the Family and Work in a New England Industrial Community*. Cambridge University Press, New York
- Hareven T K 1990 A complex relationship: Family strategies and the process of economic and social change. In: Friedland R, Robertson A F (eds.) *Beyond the Marketplace: Rethinking Economy and Society*. Aldine de Gruyter, New York
- Hareven T K 2000a *Families, History, and Social Change: Life-Course & Cross-Cultural Perspectives*. Westview, Press, Boulder, CO
- Hareven T K 2000b Formal timing and subjective timing: Historical changes and the cultural construction of the life course. In: Bideau A, Bourdelaïs P, Brunet G (eds.) *Structures Par Âge et Âges de la Vie: Deux Visions du Monde?* Cahier de Annales de Démographie Historique, Paris
- Hareven T K, Masaoka K 1988 Turning points and transitions: Perceptions of the life course. *Journal of Family History* 13: 271–89
- Hayami A, Uchida N 1972 Size of household in a Japanese country throughout the Tokugawa era. In: Laslett P, Wall R (eds.) *Household and Family in Past Time*. Cambridge University Press, Cambridge, UK
- Henry L 1953 Une richesse démographique en fiche: Les registres paroissiaux. *Population* 8: 281–90
- Herlihy D 1987 The family and religious ideologies in medieval Europe. *Journal of Family History* 12: 3–18
- Kertzer D I, Hogan D 1989 *Family, Political Economy, and Demographic Change: The Transformation of Life in Casalecchi, Italy, 1961–1921*. Life Course Studies. University of Wisconsin Press, Madison, WI
- Kurosui S 1996 Leaving home in a stem family system: Departures of heir and non-heirs in pre-industrial Japan. *History of the Family* 1: 329–52
- Kuznesof E A 1997 Who were the families of 'natural' children in nineteenth-century Rio de Janeiro, Brazil? A comparison of baptismal and census records. *The History of the Family: An International Quarterly* 2: 171–82
- Laslett P 1965 *The World We Have Lost*. Methuen, London
- Laslett P (ed.) 1977 *Family Life and Illicit Love in Earlier Generations: Essays in Historical Sociology*. Cambridge University Press, Cambridge, UK
- Laslett P, Wall R (eds.) 1972 *Household and Family in Past Time*. Cambridge University Press, Cambridge, UK
- Lee J, Campbell C 1997 *Fate and Fortune in Rural China: Social Organization and Population Behavior in Lioning 1774–1873*. Cambridge University Press, Cambridge, UK
- Lee J, Feng W 1999 *One Quarter of Humanity: Malthusian Mythology and Chinese Realities, 1700–2000*. Harvard University Press, Cambridge, MA
- LeRoy Ladurie E 1978 *Montaillou: The Promised Land of Error*. Bray B (trans.). Braziller, New York
- Medick H 1976 The proto-industrial family economy. *Social History* 3: 291–315
- Mendels F 1972 Proto-industrialization: The first phase of the industrialization process. *Journal of Economic History* 32: 241–61
- Mitterauer M 1992 Peasant and non-peasant family forms in relation to the physical environment and the local economy. *Journal of Family History* 17: 139–59
- Modell J, Furstenberg F, Hershberg T 1976 Social change and transitions to adulthood in historical perspective. *Journal of Family History* 1(1): 7–32
- Modell J, Hareven T K 1973 Urbanization and the malleable household: Boarding and lodging in American families. *Journal of Marriage and the Family* 35: 467–79

- Neugarten B, Hagestadt G 1976 Age and the life course. In: Binstock R, Shanas E (eds.) *Handbook of Aging and the Social Sciences*. Van Nostrand Reinhold, New York
- Parsons T 1955 The American family: Its relation to personality and to the social structure. In: Parsons R, Bales R (eds.) *Family, Socialization, and Interaction Process*. Free Press, Glencoe, IL
- Perrot M (ed.) 1990 *A History of Private Lives*. Vol. 4, *From the Fires of Revolution to the Great War*. Goldhammer A (trans.). Beklnap Press of Harvard University Press, Cambridge, London
- Plakans A, Wetherell C 1999 Fertility and cultural diversity: An exploratory essay concerning the borderlands problem in the nineteenth century Baltic area. *Continuity and Change* 1: 33–56
- Quataert J 1985 Combining agrarian and industrial livelihood: Rural households in the Saxon Oberlausitz in the nineteenth century. *Journal of Family History* 10: 145–62
- Reher DS 1997 *Perspectives on the Family in Spain, Past and Present*. Oxford University Press, Oxford, UK
- Rogers J 1993 Nordic family history: Themes and issues, old and new. *Journal of Family History* 18: 291–314
- Sabean DW 1998 *Kinship in Neckarhausen, 1700–1870*. Cambridge University Press, Cambridge, UK
- Segalen M 1991 *Fifteen Generations of Bretons: Kinship and Society in Lower Brittany, 1720–1980*. Cambridge University Press, New York
- Shorter E 1976 *The Making of the Modern Family*. Basic Books, New York
- Smelser N 1959 *Social Change in the Industrial Revolution: An Application of Theory to the British Cotton Industry*. University of Chicago Press, Chicago
- Smith R 1979 Kin and neighbors in a thirteenth-century Suffolk community. *Journal of Family History* 4: 219–56
- Stone L 1977 *The Family, Sex, and Marriage in England, 500–1800*. Harper and Row, New York
- Tilly C 1987 Family history, social history, and social change. *Journal of Family History* 12: 319–30
- Tilly L, Scott J 1978 *Women, Work, and Family*. Holt, Rinehart and Winston, New York
- Todorova M 1997 *Imagining the Balkans*. Oxford University Press, New York
- Vinovskis M 1977 From household size to the life course: Some observations on recent trends in family history. *American Behavioral Scientist* 21: 263–87
- Wirth L 1938 Urbanism as a way of life. *American Journal of Sociology* 44: 1–24
- Wrigley E 1972 The process of modernization and the industrial revolution in England. *Journal of Interdisciplinary History* 3: 225–60
- Wrigley E, Schofield RS 1981 *The Population History of England, 1541–1871*. Harvard University Press, Cambridge, MA

T. K. Hareven

Family and Psychiatric Illness/Family Therapy

The term family therapy refers to the involvement of family members in the treatment of the patient. One participating member, as in the case of a spouse or

partner, is sufficient to justify the use of the term. Family members included in the therapy may share the same household with the patient or live apart, and may be seen by therapists with or without the patient. Family therapy was introduced as a means of altering family relationships that were believed to contribute to the development of the patient's illness. Recently, interventions have been mounted to help relatives cope better with long lasting illnesses without any implication that they have played a part in their causation.

1. Theories of Family Causation of Schizophrenia

Schizophrenia is one of the most severe mental illnesses and although relatively uncommon, affecting 1 percent of the population, can cause lifetime disability. Its causes are still unknown, although there is strong evidence for a contribution from inheritance (see *Mental Illness, Genetics of*). This evidence was scanty in the 1940s when psychoanalysts began to identify features of parental behavior that they linked to the development of schizophrenia. One of the earliest papers was by Hajdu-Gaines (1940), based on his psychoanalysis of four women with schizophrenia. He characterized their mothers as cold, rigorous, and sadistic and their fathers as soft, indifferent, and passive. Fromm-Reichman (1948), when discussing the problems patients with schizophrenia pose in psychoanalysis, coined the term 'schizophrenogenic mother,' which became common currency among psychiatric professionals. Coming from a similar background, Lidz et al. (1957) studied 14 families with a member suffering from schizophrenia and concluded that the parents' marriages were either unbalanced in terms of power ('marital skew') or exhibited an emotional distance between the partners ('marital schism'). Bateson et al. (1956) developed a theory that the irrational thinking of people with schizophrenia was induced by exposure to contradictory messages from their parents. They worked on the theory and came up with the term 'the double bind' that was an early formulation of catch 22. The double bind was defined carefully as a contradiction within the verbal content of a message, or between the words used and the feelings expressed. In addition, the recipient of the message is not allowed to comment on the contradiction (the doubling of the bind). Like Fromm-Reichman's term, double bind rapidly infiltrated the psychiatric vocabulary, and indeed went beyond into usage by the lay public. The basis of its popular appeal was in fact its greatest weakness as an explanation for the origin of schizophrenia, namely the fact that it was recognized immediately as occurring in everyday interchanges.

All the foregoing theorists were American, but Britain produced its own charismatic figure, R. D. Laing, who achieved popular fame during his lifetime. His formulation of schizophrenia was that it was a

completely understandable response to the family's efforts to drive one of its members mad (Laing and Esterson 1964). He believed the family selected one of its number as a scapegoat to bear the brunt of the family's emotional tangles. His ideas also spread far beyond his colleagues to be taken up by young people in particular, who found the role of being their parents' victim immensely appealing. A film based on his theory, called *Family Life*, was a popular success.

All four theories described above are flawed for the same reason: they were developed from clinical experience with highly selected groups of patients, and were not subjected to scientific tests of their validity by their originators. This would have entailed developing reliable measures of the phenomenon described, e.g., cold, rejecting mothers; double bind messages; and comparing their occurrence in representative samples of families of people with schizophrenia and in control families. One pair of theorists did just this: Lyman Wynne and Margaret Singer. Like Bateson's group, they focused on distorted communication in the family, only they developed an extensive schema for rating varieties of what they termed 'communication deviance.' Using this method to rate transcripts of subjects' speech they were able to distinguish parents of people with schizophrenia from parents of people with other psychiatric disorders and from normal subjects with uncanny accuracy (Singer and Wynne 1966). The results were so striking that they demanded replication and Hirsch and Leff (1971) attempted to do this using a British sample. They learned to make ratings of communication deviance reliably and assessed transcripts of parents of patients with schizophrenia and of parents of patients with neuroses. The former scored significantly higher than the latter but there was substantial overlap in the scores of the two groups. Thus, the parents of people with schizophrenia do seem to have an increased tendency to communicate in an unclear fashion, but this finding by no means proves that this lack of clarity induces schizophrenia in their offspring. An equally plausible explanation is that deviant communication in the parents is a manifestation of the genes they share with the ill family member. As a consequence of shared inheritance, 10 percent of parents of people with schizophrenia develop the illness themselves.

The theories presented address different parental qualities: their emotional responsiveness, the balance of power in their relationship, their ability to communicate. Each group of theorists has tended to formulate schizophrenia in terms of the particular abnormality they have focused on in the parents, rather than starting with a generally agreed concept of the nature of the illness. For example, Bateson's group conceptualized schizophrenia as a disturbance of logical thinking, while Laing saw it as an attempt to escape from the role of victim. In fact schizophrenia produces pervasive disturbances in the entire mental life of the sufferer: thinking, feeling, perception, and

communication may all be affected. None of the theories accounts for all the varied manifestations of this puzzling condition.

2. Family Therapy

2.1 The Introduction of Family Therapy

Family therapy developed initially in relation to schizophrenia, fuelled by the widely held view among psychiatric professionals that parents induced the illness in their offspring. Experienced psychoanalysts brought their expertise to bear on families caring for a person with schizophrenia. This in itself was an unconventional use of psychoanalysis, which was traditionally centered on the relationship between the analyst and a single client. Although scientific evaluations of this adaptation of psychoanalysis are almost nonexistent, the pioneers realized after some years that they were making little impact on the illness. Rubinstein (1974), one of the most respected therapists in the field, wrote that 'One of our earliest preconceptions, which proved to be unwarranted with the passage of time and our increase in experience, was that schizophrenic behavior could easily be modified if we treated the family conjointly, and if we helped the members sort out some of their distorted interpersonal dynamics.' The disillusionment of therapists with this work can be documented by calculating the decline over time in the number of papers on schizophrenia in one of the premier journals dedicated to family therapy. In the first volume of *Family Process*, published in 1962, exactly half the papers were concerned with schizophrenia. By 1975, the proportion of papers dealing with this condition had dropped to 10 percent, and by 1990 to 6 percent.

While psychoanalytic therapists turned their attention to families with other types of problems, a few charismatic individuals continued the work with schizophrenia, developing their own approaches and techniques, which often departed radically from the tenets of psychoanalysis. Therapists who are convinced of the value of their work rarely submit their efforts to scientific evaluation and no rigorous studies have been conducted of these individualistic therapists.

2.2 Shift from Schizophrenia to Other Diagnostic Groups

The focus of family therapy shifted to disorders of children, such as bedwetting, tantrums, and school refusal. These disturbances proved to be much more responsive to a family approach than schizophrenia, partly because there is not likely to be a biological basis to these behavioral problems, and partly because

children are much more under the influence of their parents than young adults. This point is underscored by the results of family therapy for eating disorders. This has beneficial results when the patient is a child or adolescent, but is less effective for adult patients, who do better with individual therapy (Russell et al. 1987).

2.3 The Influence of Systems Theory

While psychodynamic approaches to family work continue to have their adherents, they have been eclipsed by the rise of systemic family therapy. This was based on systems theory as enunciated by Bertalanffy (1968). The family is conceptualized as a system that attempts to maintain its integrity by responding to change in one part of the system with compensatory changes in other parts. Homeostasis, then, is the overriding concern of the family system (Jackson 1957). When a child develops a disturbance such as bedwetting, this is seen as compensating for a disturbance between other members of the family, say a conflictual relationship between the parents. When the parents bring the child for treatment, the child's symptoms are taken to signify dysfunction in the family system as a whole. For this reason the individual presenting for treatment is often termed the 'designated patient,' with the implication that this person has been selected by the family (unconsciously) as a messenger to the health professional that the whole family is in need of help. There is an echo here of Laing's view of the person with schizophrenia as the scapegoat for the rest of the family.

2.4 Family Therapy Approaches

Structural family therapy was developed by Minuchin and co-workers and entailed a basic assumption about the way in which a normative family should function (Minuchin 1974). The therapist takes an active role in trying to mold the family as closely as possible to this ideal image. As it involved a view of the father as the head of the household, it was criticized as gender biased and ethnocentric. In response to this, structural therapists now request families to state how they would like to function, and then help them to achieve the necessary changes, thus incorporating cultural sensitivity.

Strategic therapy, by contrast with systemic therapy, takes the presenting symptom at face value and involves a narrow focus and a time-limited intervention. An important innovation of strategic therapists is 'reframing,' in which the problem as presented by the family is reformulated, often changing its negative connotations into positive ones. This can free the family from established patterns of behavior and enable them to respond to the 'problematic' member in a different way.

Systemic family therapy received a boost from a group working in Italy who became known as the Milan School. The key therapists were Selvini Palazzoli, Boscolo, and Cecchin. They introduced innovative techniques such as paradoxical injunctions, and focused on patients with eating disorders and with schizophrenia (Selvini Palazzoli et al. 1978). They developed an interviewing technique of circular questioning, which has become recognized by family therapists as a useful tool. An example would be the therapist asking one member to describe the relationship between two other family members. This will produce responses from all concerned and will prompt the therapist to formulate the next question. This process makes evident to the family the circularity of interaction, as hypothesized by systems theory. The fame of the Milan therapists was such that clients came to them from all over Italy. This made follow-up very difficult and no convincing outcome research has been conducted on their patients.

2.5 Research on Family Therapy

In fact systemic formulations about the family's contribution to mental illness are difficult, though not impossible, to test scientifically. A number of instruments have been devised to measure the ways in which family members function when together, but these are not entirely satisfactory in grappling with the problem, since when several people are interacting they produce a vast amount of data. As an example, Mishler and Waxler (1967) recorded interactions between two parents and a single offspring. The recording, transcribing, coding, and scoring of the interviews required 129 hours of staff time for every hour of interview. For this reason greater success has been achieved by focusing on specific aspects of family interactions. The introduction of videotaping of family sessions, allowing replay, pausing, and slow motion, facilitated the study of interactions in great detail. One of the most productive lines of research has developed from the use of a measure of relatives' emotional response to a sick family member, known by the term Expressed Emotion (EE).

3. Relatives Expressed Emotion

This assessment instrument was introduced to investigate one of the consequences of deinstitutionalization (e.g., see *Deinstitutionalization, Psychiatry of*). In the UK it was found that patients discharged from psychiatric hospitals to live with close relatives were more likely to be readmitted than those who went to live with more distant kin or with landlords. George Brown, a sociologist, had a hunch that it was to do with the emotional relationship between patients and their close relatives, and with Michael Rutter, a child

psychiatrist, developed a technique to measure these. It is based on a semistructured interview with the relative enquiring about the patient's behavior and symptoms in the previous three months. The interview is audiotaped and ratings are made later from the tape. These depend partly on what the relative says and partly on the way their voice is used to convey emotion. The key ratings are of critical comments, reflecting how angry the relative is with the patient, hostility, representing rejection of the patient, overinvolvement, comprising overprotection and overidentification, and warmth.

3.1 Relatives EE and Schizophrenia

The initial studies were of patients with schizophrenia, since this was the most common diagnosis in patients discharged from long-term hospital care (Leff and Vaughn 1985). It was found that relatives' negative emotions—critical comments, hostility, and over involvement—were each associated with a greater likelihood of the patients to relapse. Consequently, they were combined into an index of Expressed Emotion (EE) and relatives were classified as high or low EE. In most parental households both parents were congruent in being either high or low EE. When they differed in their ratings, the highest rating was taken as characterizing the household. By contrast with the ratings of negative emotion, warmth was associated with a better outcome for the patients. Thus, it appeared that the family members could influence the course of schizophrenia either in a positive or a negative way.

The early studies of EE aroused a great deal of interest since they demonstrated that family members' emotional relationship with the patient could be measured reliably, and furthermore that these relationships appeared to determine the outcome of the patient's illness. It should be emphasized that, unlike the theories of causality reviewed above, it was not claimed that high levels of EE antedated the appearance of schizophrenia or played any part in its etiology. If there was any causal influence, it was on the subsequent course of the illness once it was established. Many groups have replicated the link between EE and the outcome of schizophrenia, with the result that it is now one of the best evidenced associations in psychiatry (Butzlaff and Hooley 1998). These replications have been carried out in a number of languages and cultures, including India, China, and Japan, showing that the association is independent of variations in family structure and function.

3.2 Family Interventions for Schizophrenia

The research on relatives' emotional responses and schizophrenia led to a revival of interest in working with the families of such patients. However the

approach was quite different from that of the earlier generation of therapists. Instead of viewing the family as pathogenic and in need of treatment, the new attitude was one of respect for the carers and a genuine desire to work with them as allies in the struggle against the illness. This in itself was revolutionary since for decades family members had been subject to vilification and rejection by mental health professionals. Instead of the patients being scapegoated by their families, as Laing had claimed, it was the relatives who were treated as scapegoats by the professionals. Relatives were denied information about the patient's illness on the grounds of confidentiality, which unnecessarily hindered their attempts to understand what was happening to the patient. Partly as a response to this treatment, relatives formed national organizations, such as the National Schizophrenia Fellowship in the UK and the National Alliance for the Mentally Ill in the US. This development raised the profile of relatives, gave them access to political power, and increased their confidence in dealing with professionals (e.g., see *Mental Illness: Family and Patient Organizations*).

Controlled trials of family interventions for schizophrenia began to appear in print in 1978 and over the next two decades more than a dozen such studies were published (Lam 1991). They all employed the same design, namely a comparison of family intervention plus antipsychotic medication against medication alone. This design reflects the important assumption that antipsychotic medication is an essential basis for the treatment and maintenance of a person with schizophrenia, but that working with the family to improve the management of the condition confers an added benefit. Indeed, this series of studies provided strong evidence for that proposition. Patients' relapse rate over one year was reduced from an average of 50 percent with medication alone to 20 percent or less by the addition of family interventions.

Despite a multiplicity of names, the interventions share much in common. They all start with an educational program which conveys to the family information about the causes of schizophrenia, the symptoms, the likely outcome, and the treatment and management. This is followed by instruction in problem solving, improving communication, reducing the intensity of negative emotions, utilizing the family's natural support networks, and developing realistic expectations for the patient. The general aim of the approach is to build on the family's strengths and to facilitate their mutual co-operation in coping with the illness.

3.3 Testing Systems Theory

Some programs were delivered by sessions in the family's own home, a radical departure from traditional family therapy, in which treatment often took place in a special clinic room equipped with a one-way

mirror to allow the whole team to monitor the interactions between therapist and family. Home visits are an effective way of engaging families, as well as enabling the therapists to observe families in their natural surroundings. These advantages have to be weighed against the extra staff time taken up by travel. Other programs set up groups for relatives that were run by professionals on the same lines as the family sessions. Voluntary organizations for families often establish self-help groups, but when they invite professionals along it is to provide information and not to institute therapy. Indeed they reject the term 'family therapy' as implying that relatives are pathological and in need of treatment themselves.

One American therapist, Bill McFarlane, based his program on multifamily groups that included patients as well as relatives. He has produced some evidence that these are more effective than groups comprising only relatives (McFarlane et al. 1993). Relatives groups provide the opportunity to test one of the basic tenets of systems theory, namely that altering one part of the system can effect change in the entire system. In the course of evaluating family interventions for schizophrenia, Leff et al. (1989) compared the effectiveness of family sessions with a relatives group from which patients were excluded. They found that relatives who attended the group showed a significant reduction in EE and this was accompanied by a lowered relapse rate for the patients. Hence, altering one component of the system, the relatives, led to the desired change in another component, the patient.

3.4 Relatives EE and Other Illnesses

Since the original research on schizophrenia, the measurement of EE has been applied to a variety of other conditions, both psychiatric and physical. In the realm of psychiatry, high levels of EE have been found in relatives of patients with manic-depressive psychosis, depressive neurosis, dementia, learning difficulties, alcoholism, eating disorders, and post-traumatic stress disorder, and in families in which children have suffered physical or sexual abuse. High EE has been found to be linked with outcome in all these conditions, with the exception of dementia and learning difficulties. In the domain of physical illness, high EE attitudes have been recorded in relatives of patients with epilepsy, heart disease, diabetes, inflammatory bowel disease, and Parkinson's disease. From this it appears that high EE attitudes can develop in response to any long-standing or recurrent illness, be it psychiatric or physical. However, an association with outcome is much stronger for psychiatric than for physical disorders. Regarding the latter, only for diabetes has a link been demonstrated and the evidence is equivocal.

It seems from this series of studies that the emotional attitudes of family members have a much greater

impact on people suffering from disorders with an emotional basis than on people with physical illnesses where there is a demonstrable pathology. The exception might seem to be schizophrenia, in which brain abnormalities have been demonstrated (e.g., see *MRI (Magnetic Resonance Imaging) in Psychiatry*). However, a high proportion of patients with schizophrenia experience depressive symptoms and they are responsive to the same adverse events in their life as are people with a depressive illness. Hence, the underlying brain disorder does not negate their sensitivity to the social environment.

4. Couple Therapy for Depression

Marital therapy, as it used to be called before marriage went out of fashion, has a long history, and developed to help couples resolve their relationship difficulties or to part amicably. In the UK, a voluntary organization, Relate, trains counselors in the skills needed for couple therapy. However, there is not usually an identifiable psychiatric illness in either member of the partnership. In recent years, a small number of studies have tested the value of couple therapy for people with a depressive illness. In most of these, couple therapy has been compared with cognitive therapy, a treatment of known effectiveness for depression that focuses on the individual patient. In general, couple therapy has been found to be as effective as cognitive therapy for improving depression, but has stronger effects on the couple's relationship.

A recent study tested couple therapy against the best possible regime of antidepressant medication for depressed patients who were in a stable relationship. Both treatments were continued for one year and then stopped, after which the subjects were followed up for a further year. This design was chosen because most treatments in use improve depressive symptoms: the challenge is to reduce the liability for the symptoms to recur. The couple therapy was given by two systemic therapists who focused on the relationship between the partners as being the means by which the depression was maintained. The patients who received couple therapy showed a greater improvement in depression than those assigned to medication, and this advantage for couple therapy persisted over the second year after treatment had been stopped (Leff et al. 2000). Hence, changing the relationship between patient and partner was more effective both for treatment and prevention than altering the biochemistry in the patient's brain.

5. The Family as a Therapeutic Resource

Since family therapy was first introduced, there has been a major shift in professional attitudes from holding relatives responsible for the patient's illness to

viewing them as allies in a struggle to help the patient overcome symptoms and disabilities. Alongside this sea change, western families have been altering in their composition, structure, and function, requiring a greater flexibility in professional inputs. Several schools of family therapy have developed, each with their own distinctive concepts of the family, and have introduced novel techniques. Not all of these have proved their worth, but the emerging evidence of family influences on the course of a wide range of psychiatric disorders is a spur to test existing methods of working with families and to develop innovative ones.

See also: Cognitive and Interpersonal Therapy: Psychiatric Aspects; Family Therapy, Clinical Psychology of; Fatherhood; Mental Illness: Family and Patient Organizations; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychotherapy: Ethical Issues

- to clinical practice: Dissemination of New York State's family psychoeducation project. *Hospital and Community Psychiatry* **44**: 265–70
- Minuchin S 1974 *Families and Family Therapy*. Tavistock, London
- Mishler E G, Waxler N E 1967 Family interaction patterns and schizophrenia: A multi-level analysis. In: Romano J *The Origins of Schizophrenia*. Excerpta Medica International Congress Series No. 151, Excerpta Medica, Amsterdam
- Rubinstein D 1974 Techniques in family psychotherapy of schizophrenia. In: Cancro R, Fox N, Shapiro L E (eds.) *Strategic Intervention in Schizophrenia*. Behavioral Publications, New York
- Russell G F, Szmukler G I, Dare C 1987 An evaluation of family therapy in anorexia nervosa and bulimia nervosa. *Archives of General Psychiatry* **44**: 1047–56
- Selvini Palazzoli M, Boscolo L, Cecchin G, Prata G 1978 *Paradox and Counterparadox*. Jason Aronson, New York
- Singer M T, Wynne L C 1966 Communication styles in parents of normals, neurotics and schizophrenics. *Psychiatry Research Reports* **20**: 25–38

J. Leff

Bibliography

- Bateson G, Jackson D D, Haley J, Weakland J H 1956 Toward a theory of schizophrenia. *Behavioral Science* **1**: 251–64
- von Bertalanffy L 1968 *General Systems Theory*. Braziller, New York
- Butzlaff R L, Hooley J M 1998 Expressed emotion and psychiatric relapse: A meta-analysis. *Archives of General Psychiatry* **55**: 547–52
- Fromm-Reichmann F 1948 Notes on the development of treatment of schizophrenics by psychoanalytic psychotherapy. *Psychiatry* **11**: 73–8
- Hajdu-Gaines I 1940 Contributions to the etiology of schizophrenia. *Psychoanalysis Review* **27**: 421–38
- Hirsch S R, Leff J P 1971 Parental abnormalities of verbal communication in the transmission of schizophrenia. *Psychological Medicine* **1**: 118–27
- Jackson D 1957 The question of family homeostasis. *Psychiatric Quarterly* **31**: 79–90
- Laing R D, Esterson D 1964 *Sanity, Madness and the Family*. Tavistock, London
- Lam D H 1991 Psycho-social family intervention in schizophrenia: A review of empirical studies. *Psychological Medicine* **21**: 423–41
- Leff J, Berkowitz R, Shavit N, Strachan A, Glass I, Vaughn C 1989 A trial of family therapy vs. a relatives group for schizophrenia. *British Journal of Psychiatry* **154**: 58–66
- Leff J, Vaughn C 1985 *Expressed Emotion in Families: Its Significance for Mental Illness*. Guilford, New York
- Leff J, Vearnals S, Brewin C R, Wolff G, Alexander B, Asen E, Dayson D, Jones E, Chisholm D, Everitt B 2000 The London intervention trial: An RCT of antidepressants versus couple therapy in the treatment and maintenance of depressed people with a partner: Clinical outcome and costs. *British Journal of Psychiatry* **177**: 95–100
- Lidz T, Cornelison A R, Fleck S, Terry D 1957 The intrafamilial environment of the schizophrenic patient II. Marital schism and marital skew. *American Journal of Psychiatry* **114**: 241–8
- McFarlane W R, Dunne E, Lukens E, Newmark M, McLaughlin-Toran J, Deakins S, Horen B 1993 From research

Family and Schooling

There is an obvious difficulty in considering relationships of the family to other institutions at a time when it is changing rapidly and radically in structure, functions, and internal dynamics. While some investigations have addressed aspects of those changes, much research on family–school relationships does not take account of them. Perhaps this research is more properly regarded as providing information on homes, which can encompass a variety of family types, rather than on families.

Up to relatively recently, the family took major responsibility for the care, education, and training of children. In it, children acquired skills to deal with day-to-day tasks, including those of work. During the Industrial Revolution, as work moved into factories and became more specialized and skilled, and knowledge accumulated, homes could no longer provide children with the required knowledge and skills, and the need for an institution dedicated to formal learning arose. Increases in migration, industrialization, and urbanization served to underline the need, and so nations came to accept education and training in school as their responsibilities.

Children still spend most of their waking time outside school, however, and it would be surprising if the continuous influence of the family, both before children enter school and when they are attending school, did not play a major role in their learning. Although one would expect families and schools to play complementary and mutually reinforcing roles in children's education, they may not always do so

because of differences in their goals, structure, demands, activities, values, and the competencies they foster. When these differences are large, children may experience a serious dysfunction in socialization when they go to school, which will affect their ability to adapt and to learn. To deal with this problem, concerted efforts have been made in several countries in recent years to increase parents' involvement in the formal education of their children.

This article provides a summary of the findings of research that indicate that children's home circumstances are related to school learning. Proposed explanations of the findings are presented. Finally, programs to increase parental involvement in the education of children living in families and communities characterized by high rates of educational failure are considered.

1. The Findings of Research on Family Characteristics and Children's School Learning

Research on the family factors that influence children's development, much of it in the context of policies of providing equality of educational opportunity, has been carried out since 1900 in several disciplines: education, psychology, sociology, economics, and behavior genetics. There has been little integration of the work across disciplines.

Findings of the empirical studies can be summarized in eight general conclusions (Kellaghan et al. 1993, White 1982): (a) level of socioeconomic status (or social class) is positively but not very strongly related to children's performance on measures of scholastic ability and achievement. Children from high socioeconomic status homes tend to perform better than children from low socioeconomic status homes; (b) differences in cognitive skills between children from varying socioeconomic and ethnic backgrounds are in evidence before children go to school; (c) the gap between children from different backgrounds widens as children progress through school; (d) the length of time a student spends in the education system is a function of the socioeconomic status of the student's family. This is so even for students of the same level of general scholastic ability; (e) in countries in which secondary schools are differentiated, major selection in terms of socioeconomic background occurs at the point of transfer from primary to secondary school. Students from high socioeconomic backgrounds tend to transfer to an academic type of secondary school which will lead to third-level education, while children from low socioeconomic backgrounds tend to go to technical, vocational, or 'short-cycle' schools; (f) variables which describe the ethos and teaching style of families predict students' academic achievement better than do measures of the socioeconomic status of families. Thus, what parents do is more important

than their status; (g) children's learning benefits from parental involvement. Such involvement is more likely to occur if parents construe their role as involving activities related to children's education, and if they believe that involvement will have a positive influence on their children and will be welcomed by schools (Epstein 1991, Hoover-Dempsey and Sandler 1995); (h) many studies reflect a low negative correlation between family size and educational abilities, achievements, and attainment. Increased spacing between children tends to reduce the decrement associated with increase in family size, while the effects of size are more marked in homes in which the father has a low occupational level.

Methodologies have been employed which reject the 'objective' quantitative approach of empirical studies, adopting the view that since acts imply purposeful constructions by actors, they can only be understood from the actor's point of view. In this context, ethnographic and case study procedures have been used to explore individuals' 'interpretations' of home-school relationships (see Marjoribanks 1991). The approach is supported by the findings of psychological research that perceptions of family supports are better predictors of personal adjustment than are actual supportive behaviors provided by network members (Pierce et al. 1996).

Research on the effects of variation in family structure and function on children's school learning does not provide clear-cut findings, which is not surprising, given the complexity and variety of the conditions being investigated. It is, for example, difficult to design studies that will isolate the effects of family circumstances (e.g., divorce, maternal employment, blended families, lone parenting) from the effects of many conditions frequently associated with these circumstances. One consistent finding, however, is that lone parenting is often associated with school problems for children when the parent (usually a mother) is young, financially poor, and poorly educated (Kellaghan et al. 1993).

Changes in family structure, which are associated with a decrease in the monitoring and supporting of children's behavior by parents and other adults, and in which the cultivation of one's own well-being replaces interest in others, are perceived by some commentators to have a negative impact on children's educational development (see Coleman 1987). The changes are represented as reflecting an erosion of 'social capital' (see Sect. 2).

2. Explaining the Family's Influence on Children's Development

Using a concept analogous to financial capital, the term 'cultural capital' has been used to conceptualize some of the influences of families, and the communities

in which they are embedded, on children's academic development. It exists in three forms: in cognitive and noncognitive competencies derived from past experiences (particularly familial ones) which actively organize future experience; in an 'objectified' state in cultural goods (pictures, books, instruments); and in an institutionalized state (educational qualifications). Language forms an important part of cultural capital since, in addition to being a means of communication, it provides, together with a richer or poorer vocabulary, a system of categories, which enables one to decipher and manipulate complex logical and aesthetic structures (Bourdieu 1986).

Separate strands of research in education (Kellaghan et al. 1993) and in psychology (Moos and Moos 1994) have identified a number of behaviors and conditions in the family that contribute to cultural capital: (a) modeling (in use of complex language; in planning and organization to ensure that time and space are well structured and used; intellectual-cultural orientation in activities; moral-religious emphasis); (b) providing motivation and reinforcement (encouraging and rewarding school-related activities and independence in decision-making); (c) holding high academic aspirations and expectations; (d) providing direct instruction (guiding and supporting academic work; helping with homework); and (e) ensuring that the activities engaged in are developmentally appropriate.

A number of points have been made about cultural capital: (a) its value will vary with the 'markets' in which it can be used advantageously; (b) capital required for success in school is defined by the dominant social groups in society; (c) ability or talent is the product of an investment of time and cultural capital; and (d) the notion of cultural capital helps explain the unequal scholastic achievement of children originating from different social classes. If you bring the right kind of capital to school, you will do well. If you do not, you are likely to find yourself at an academic, social, and ideological disadvantage in most schools (Bourdieu 1986).

In separate, but mutually supporting research in psychology and sociology, interpersonal relationships have been identified as another area in which the family's influence on children's development can be discerned. In psychology, the focus is on cognitive aspects of support (that incorporate expectations about how others will respond when assistance is needed) and on actual supportive behavior (Pierce et al. 1996). Sociologists, on the other hand, speak of 'social capital,' which is perceived to inhere in the structure of relations between and among individuals and organizations. Three forms have been identified: shared obligations and expectations; useful information that is passed between members of a social network; and a value or norm consensus between adults, coupled with a willingness to monitor behavior and enforce norms that control and shape children's

activities. These conditions are seen as contributing to the development in children of basic attachments, attitudes, willingness to make certain efforts, and concepts of self, all of which work to children's advantage when they go to school. A variety of individuals and organizations can help children with these developmental tasks: siblings, grandparents and other members of the extended family, neighbors, churches, and various social and cultural groups (Coleman 1987, Garbarino et al. 1997).

These explanations of the factors in social and familial contexts that affect school learning do not help us understand how it is that children from the same family can differ considerably in their success at school. The answer would seem to lie in the fact that, although certain aspects of children's genetic endowment and environment (inside and outside the home) are shared by children in a family, other aspects are unique to each child. It may be that non-shared influences (e.g., the specific treatment of a child by parents and siblings) contribute more to variation in children's characteristics than family variables that are common to all children (e.g., socioeconomic status, general parenting style) (Rowe 1994). The complexity of the interaction between genetic and environmental factors becomes obvious if it is the case that compared to their less 'able' counterparts, genetically more 'able' parents create more stimulating environments, and genetically more 'able' children take a more active role in shaping their environment.

3. Home-School Partnerships

Recent efforts to improve home-school relationships may be attributed to concern about achievement levels in schools, sometimes arising out of the results of comparative international studies of achievement; the limited success, despite considerable investment, of efforts, such as 'Headstart' in the USA and similar programs elsewhere, to raise significantly the educational achievements of children living in areas described as economically and socially disadvantaged; and realization of the contribution that factors outside the school in homes and communities, particularly in the early years, make to the development of competence in children, a realization that was inspired by evidence relating to the importance of early childhood experiences, and of the effects of home background on achievement.

In the USA, there is a long tradition of involving parents and communities in children's formal education in, for example, parent-teacher associations. In Europe, the situation is more complex because of differences between countries in legal arrangements and ideology. However, in many countries, recent efforts, in some cases supported by legislation, go well beyond traditional parent-teacher associations in the role they propose for parents, involving them in

curriculum decision-making, and in teaching and learning activities in the classroom (e.g., as classroom assistants) or in the home (e.g., in developing children's skills in problem solving and language, and in supervision of homework). Many of the efforts are directed towards families in 'disadvantaged' areas, where children face economic, educational, and social problems, and are considered to be at risk of educational failure, school dropout, and intergenerational poverty. As is the case with family-support initiatives in general, programs are designed to improve the developmental capacity and educational options available to families. They are based on a variety of models and approaches, have few theoretical frameworks, and vary in the extent to which empirical evidence is available on their effectiveness.

Most programs can be described as instrumental-technical, in that they are designed to provide parents with a range of skills that are considered relevant to their children's education. Alternative approaches are based on an interpretative paradigm, which, recognizing the uniqueness of parents' experience, attempt to facilitate them in interpreting and understanding that experience. Yet other approaches go a step further, again helping parents interpret their situation, but to do so critically in the context of presumed determining social structures and ideologies (e.g., Morgaine 1992).

One of the mechanisms used in the USA (Becker et al. 1997) and in the UK (Bastiani and Wyse 1999) to involve parents requires them to sign home-school contracts. Such agreements describe, among other things, the reciprocal responsibilities of schools, parents, and students in relation to student learning and school discipline.

A number of problems associated with efforts to promote partnerships have been identified: (a) despite rhetoric regarding choice, partnership, and parents' role as 'primary educators,' many efforts mirror the traditional home-school power relationships that they were designed to alter; (b) in many cases, schools limit the involvement of parents, denying them any real input into the construction of 'shared understandings'; (c) many approaches are based on an implicit view of parents as being inadequate in parenting skills, or as being 'hard to reach'; (d) many parents are untouched by the partnership, and remain distanced and alienated; (e) when involvement is based on contracts which require agreement before a student will be admitted to a school, only families with a certain level of commitment may enrol their children; (f) the extent of parental influence depends on parents' social, cultural, and material resources; (g) finally, strategies generally fail to reach fathers (Vincent and Tomlinson 1997).

While evaluations of individual initiatives to improve family involvement have reported a variety of effects on parents and children, it is difficult to generalize about the effectiveness of home-school

partnerships because of variation in the goals of programs, in the situations in which programs are delivered, in the delivery of services, and in the recipients of services.

4. Conclusion

Research has gone a considerable way in identifying family characteristics that are associated with children's successful school learning. We may expect future research to be more interdisciplinary, to focus more on individuals' perceptions of family-school relations, to attend to the role of individuals' unique nonshared attributes and experiences, and to support the development of parent involvement programs that are based on the concept of equality of the partners.

See also: Divorce and Children's Social Development; Families as Educational Settings; Family as Institution; Learning and Instruction: Social-cognitive Perspectives; Nontraditional Families and Child Development; Parenting: Attitudes and Beliefs; Parenting in Ethnic Minority Families: United States; Parents and Teachers as Partners in Education

Bibliography

- Bastiani J, Wyse B 1999 *Introducing Your Home-School Agreement*. Royal Society for the Encouragement of Arts, Manufacture and Commerce, London
- Becker H J, Nakagawa K, Corwin R G 1997 Parent involvement contracts in California's charter schools: Strategy for educational improvement or method of exclusion? *Teachers College Record* **98**: 511–36
- Bourdieu P 1986 The forms of capital. In: Richardson J G (ed.) *Handbook of Theory and Research for the Sociology of Education*. Greenwood Press, Westport, CT, pp. 241–58
- Coleman J S 1987 Families and Schools. *Educational Researcher* **16**(6): 32–8
- Epstein J L 1991 Effects on student achievement of teachers' practices of parent involvement. In: Silvern S B (ed.) *Advances in Reading/Language Research. Literacy through Family, Community, and School Interaction*. JAI Press, Greenwich, CT, Vol. 5, pp. 261–76
- Garbarino J, Kostelny K, Barry F 1997 Value transmission in an ecological context: The high-risk neighborhood. In: Grusec J E, Kuczynski L (eds.) *Parenting and Children's Internalization of Values. A Handbook of Contemporary Theory*. Wiley, New York, pp. 307–32
- Hoover-Dempsey K V, Sandler H M 1995 Parental involvement in children's education: Why does it make a difference? *Teachers College Record* **97**: 310–31
- Kellaghan T, Sloane K, Alvarez B, Bloom B S 1993 *The Home Environment and School Learning. Promoting Parental Involvement in the Education of Children*, 1st edn. Jossey-Bass, San Francisco
- Marjoribanks K 1991 Families, schools, and students' educational outcomes. In: Fraser B J, Walberg H J (eds.) *Educational Environments: Evaluation, Antecedents and Consequences*, 1st edn. Pergamon, Oxford, UK, pp. 75–91

- Moos R H, Moos B S 1994 *Family Environment Scale Manual*. Consulting Psychologists Press, Palo Alto, CA
- Morgaine C A 1992 Alternative paradigms for helping families change themselves. *Family Relations* 41(1): 12–17
- Pierce G R, Sarason B R, Sarason I G 1996 *Handbook of Social Support and the Family*. Plenum Press, New York
- Rowe D C 1994 *The Limits of Family Influence. Genes, Experience and Behavior*. Guilford Press, New York
- Vincent C, Tomlinson S 1997 Home-school relationships: 'The swarming of disciplinary mechanisms?'. *British Educational Research Journal* 23: 361–77
- White K R 1982 The relation between socioeconomic status and academic achievement. *Psychological Bulletin* 91: 461–81

T. Kellaghan

Family, Anthropology of

At intervals during the last 50 years, anthropologists have pointed out the deficiencies of the concept of the family as an analytical tool (e.g., Fortes 1958, Yanagisako 1979, La Fontaine 1985). One of the more recent, Robertson, starts with an account of the importance of the idea of the family in Western culture; later, he draws attention to social scientists' 'fixation on the family' (1991, p. 154). These two phenomena are undoubtedly connected. An account of anthropological understandings of the family must distinguish between its significance in Western society and its use in the comparative analysis of social life. Comparing the two may then expose the fundamental ethnocentricity of the concept and its weakness in cross-cultural analysis.

1. The Family in Western Society

The significance of the family in Western societies is obvious in daily life. Children in primary school are asked to describe their families as a writing exercise and politicians vie with one another to be 'the party of the family.' Christianity, the majority religion of Western society, celebrates the Holy Family and its priests proclaim the duty of Christians to emulate it in their domestic life. The terms used by members of a family for one another are not applied to any other relatives and the relationships they designate are seen as unique. The family encloses the private life of individuals and, as the building block of society, is given equal weight with the public institutions of the state, since the division into public and private is one of the fundamental social distinctions in Western culture. Yet this important institution is also the subject of controversy. There are arguments over the past and future of the family, over whether it is universal or merely an adaptation to the modern world, over the possibility of its decline and the effect on it of political and economic measures. The family is not taken for granted.

Despite its importance, the word 'family' has no clearly defined meaning. In twentieth century England, for example, in ordinary speech it may evoke any one of a cluster of meanings, depending on context (Harris 1983, p. 31). As well as parents and children, 'family' can mean: a network of kin ('all the family'); a succession of related people ('the family has lived here for 400 years'); children ('they have not brought the family with them'); birth status ('he comes from a good family'); household ('children leave the family as they grow up'); relatives generally ('she has no family in London'); etc. It is difficult to claim that any of these meanings is the 'true' meaning and others merely metaphorical usages, although some people might claim that the unit of parents and children is the primary sense of the term. However, there is evidence to show that this narrower meaning is relatively recent in English history. Davidoff and Hall (1987) trace its importance to the rise of the industrially based middle-class in the nineteenth century.

Three elements are *interrelated* in complex ways in the various uses of the word family: marriage, relatedness, or kinship and domestic or household organization. Each usage of the term carries different combinations of these elements. The 'elementary' or 'nuclear' family, as social scientists label parents and children, combines all three and it is this form that carries such symbolic weight in twentieth-century Western culture. At some point in the history of anthropology each of these three features has been used separately, and effectively, in cross-cultural comparison (see *Marriage; Kinship in Anthropology; Household in Anthropology*). The three principles are quite distinct in the abstract but separating them in analysis is not easy. However, it is the confusion between, for example, kinship and residence or marriage and parentage that causes problems in using 'the family' as a technical term.

In the West marriage is the foundation of 'the family.' As a legally established relationship the marital tie may be contrasted with the 'biological' links between parents and children or between siblings, which are referred to as blood ties. In Western societies marriage entails co-residence, although this is not a universal consequence of marriage (e.g., Fortes 1949). Where the norm is for a newly-married couple to set up their own household as in much of North America and Western Europe, this qualifies the newly-married couple to be referred to as a 'family' in anticipation of the children that are expected to follow. The death of a spouse or a divorce are individual breaks in the expected pattern but remarriage 'reconstitutes' the family once more.

The centrality of relatedness to the notion of family can be seen in the fact that not all residential units qualify to be called families. Individuals living on their own, boarding schools, prisons, and colleges are never referred to as families, nor are groups of unrelated people sharing a flat. In the case of the last three, this

may indicate that children are essential to a family, but children's residential homes are not considered families either. It is the nature of the ties among them that qualifies people who live together to be called a family.

The relationships internal to a family are thought of as emotional bonds, permanent ties that reflect ideals of sharing and altruism by contrast with the short-lived rational contracts of the market or the self-serving nature of political alliances. The authority of parents over children and the role of the father as 'head of the family' are believed to derive from the reproductive functions of the family and to reflect the 'natural' differences of generation and gender within it. The assumption that living together is necessary for the family to fulfill its reproductive function underlies the interchangeability of 'family' and 'household.'

However, Western cultures also make implicit distinctions between families and households. Where children are expected to leave the parental household at adulthood, this may, like the death or divorce of parents, be said to 'break up' the original family. Nevertheless, married adults continue to be related to their parents and siblings by genealogical ties that may be referred to collectively as belonging to a 'family,' although one which is not a household. Husband and wife belong, as offspring, to distinct 'families' that the marriage can be seen as linking. Individuals are thus members of at least two 'families': one in which they are children and (usually) siblings and one in which they are parents and spouses. Sociologists have labeled these the family of origin and the family of procreation, which emphasizes the different methods by which the focal individuals are recruited, birth and marriage; the labels incorporate the folk view. From a more analytical perspective the network of genealogical ties linking individuals can be seen as cross-cut by living arrangements that group them into different households.

Other usages of the term 'family' used in a variety of contexts show different combinations of kinship and residence. The term extended family may be used to refer to all individuals linked genealogically but not living together, giving this use of the term 'family' a purely kinship connotation, unlike its use to mean close kin who do live together. Alternatively, a large number of related households in one community may be referred to as an extended family, making their common locality a significant factor. Finally, 'extended family' may mean large households containing three or more generations, which can still be found in some rural communities in Europe. All these uses of the term are also current in academic works (e.g., Janssens 1993, 1fn).

The range of meanings that attach to the concept of the family make clear its great importance in Western culture. However, its complexity also raises a series of problems when it is used to analyze other societies. Simplifying the concept by identifying one or other of

its component elements as fundamental has not solved the problems in cross-cultural analysis but has preserved the idea that 'the family' is a useful concept.

2. The Family in Anthropological Analysis

The issues with which the anthropology of the family have been associated reflect Western notions. At the outset the focus was on marriage. In the nineteenth century, the scholars who developed anthropology regarded monogamous lifelong marriage as characteristic of 'civilization,' the culmination of human evolution. The Western family was a late development not to be found among peoples who represented earlier stages of the evolutionary process. These latter were characterized by plural marriage and a failure to understand the nature and significance of biological links. Nineteenth century evolutionism has long been recognized as lacking a sound empirical basis for its conjectures, but the idea that the present form of 'the family' is a product of social changes over time was not entirely abandoned. It survived in modified form in the view that the present form of 'the family' is a product of modernization, of capitalism, or of industrial society and in the myth that Europe's peasant forebears all lived in extended families.

By the middle of the twentieth century, some anthropologists argued instead that the family could be said to be universal. The basis for this contrary view was the assumption that 'the family,' so far from being associated with marriage in a particular cultural form, was based on a universal, natural function: reproduction. It was this biological 'fact' that formed the basis of the family, a view widespread in the societies from which anthropologists came. 'The family,' a couple with their children, might be incorporated in various different domestic arrangements, but was itself a biological 'given.' Whether any individual society recognized this or not depended on their culture, a view not unlike that of the evolutionists. Malinowski, an early proponent of this functional view, nevertheless cautioned that '...the essential features of the individual family...depend upon the general structure of a given society...' (1913, p. 6). More boldly, Murdock (1949) proclaimed the nuclear family to be universal, a co-residential unit with 'sexual, economic, reproductive, and educational' functions. His conclusions depended on the application of the label 'family' to a wide variety of social units that were not strictly comparable. The procedure showed all too clearly the looseness of the term.

Anthropological arguments over the incest taboo paralleled those over whether the family was a biological or social entity (La Fontaine 1990, pp. 20-43). The view that the incest taboo reflected a natural avoidance of sexual relations between related people assumed that incest referred to sexual relations within the 'nuclear family'; these relationships were such that

sexual attraction naturally did not arise. Other anthropologists pointed to documented cases of sexual prohibitions of an identical nature involving relationships outside the 'nuclear family,' as was also the case in Western society. Incest taboos, considered as total systems of prohibitions, varied widely cross-culturally, and could not therefore be considered biologically determined or 'natural' (Goody 1956). The later development of sociobiology resulted in claims that the taboo could also be observed to exist among the higher primates and was therefore genetic, rather than a feature of human family relations. This and other forms of biological determinism were vigorously contested (Schneider 1965, Sahlin 1976); in particular it was pointed out that paternity cannot be identified among higher primates and among human beings it is always socially determined. Few social or cultural anthropologists now accept that the incest taboo is universal, but the ideas that among close kin it is only rarely broken and that incest is 'unnatural' remains. At the end of the twentieth century, evidence that breaches of the incest taboo were infrequent but not rare in Western societies became public knowledge there. This offered proof that the 'taboo' was neither genetic nor unbroken. The reality of incest was not easily accepted, whether by the public at large or by social scientists, showing that the assumptions underlying it and also entailed in the concept of the family were deeply rooted in Western culture.

Knowledge of the variation in incest prohibitions came from the study of kinship systems that expanded rapidly in the middle decades of the twentieth century. The development had a further impact on views of the family. A solution to the theoretical problems raised by the data on domestic organization was to distinguish between 'the family' as a reproductive unit and the household or domestic group. One of the earliest to do this was Fortes (1958) who used the term domestic group, referring to the 'family' as the reproductive nucleus within it. His distinction emphasized a number of contrasts between the concepts: 'the family' consists of parents and children, whereas the domestic group might include more than two generations; the family is recruited through marriage and kinship ties, whereas many different kinds of tie may draw individuals into living together. The family is concerned with social and physical reproduction, the domestic group 'is essentially a householding and housekeeping unit' (Fortes 1958, p. 8). According to Fortes, the changes in the size and structure of domestic groups that can be observed in a particular society do not reflect breaches of any 'rule of residence' as had previously been thought, but derive partly from the effect on household constitution of the political and economic structure but also from the developmental cycle of growth and maturation in the reproductive nucleus. These processes and the cultural reactions that they entail are the driving force behind household changes.

Fortes' distinction also drew attention to production as a household activity. Property rights and labor needs were shown to influence the form of the domestic group as much as the developmental cycle of the reproductive core. Households were shaped by the demands for different kinds of labor associated with their subsistence activities and the association of related people within different households concerned economic viability as well as kinship obligations. Goody went further and showed that interpersonal relations in the 'reproductive core' were also affected by these factors (Goody 1956). Marxist anthropology subsequently attempted to construct different modes of domestic production and reveal the power structures entailed in them, but these largely ignored reproductive processes.

Distinguishing 'family' from 'household' avoided rather than solved the problems inherent in the concept of the family (La Fontaine 1985). The notion of the family as a reproductive unit conflates child-bearing and child-rearing with co-residence so that in effect the family is actually defined by residence, a return to the Western folk concept (Smith 1973). It is not surprising that some anthropologists have subsequently reverted to using family and household indistinguishably. The distinction also maintained the segregation of domestic kinship from the network of kin as a whole, which was particularly misleading where spouses were chosen from among them.

The French anthropologist Lévi-Strauss' revolutionary approach to kinship disposed of some of these questions. He argued that exchange of goods and of women, not biological reproduction, was the basis of social life. Hence, marriage rules are integral to the wider social structure, not merely the basis of the family or a means of recruitment to households. Systems where marriage with relatives was the ideal showed how the patterns of distribution of women in marriage constituted social structures. His 'atom of kinship' (1958 (1963)) contained the same elements as 'the nuclear family,' but rearranged so as to emphasize the relation between a man and his wife's brother created by marriage. Moreover its elements, pairs of opposites, reflected the oppositions in systems of thought, the deep structure of culture.

Under the rather more materialist influence of feminism, anthropologists in the 1970s showed how the 'natural' processes of childbirth seemed to be very widely used to symbolize women and legitimize their subordination to men. Within the domestic sphere, also associated with women, the division of labor was seen as reflecting the natural capabilities of men and women; the products of their labor were shared with their children. However, the distribution of goods and the allocation of labor within the household actually reflected its internal power structure rather than a pooling of family resources or a 'natural' division of family labor. The realities of power were presented as unquestionable by being associated with the natures of

men and women. Concepts of nature were thus shown to be more than mental constructs; they were also intimately bound up with the social distribution of power and authority, which they helped to reproduce. The 'family' was revealed as a similar social representation legitimizing household organization (Harris 1981).

Both of these approaches stimulated research into the meanings of folk concepts and into local systems of knowledge (Barnes 1973, Harris 1981, Carsten 1997). The social constructions of the person in folk theories made it clear that whatever the common physical nature of human beings, it is perceived through these culturally variable understandings. The image of shared blood or common physical substance are only two among many ways of representing relatedness and the concept of the family is a folk concept among others.

While Levi-Strauss' structuralist approach undoubtedly revealed the significance of cultural understandings of the world, he was not interested in distinguishing between the meanings drawn from the analysts' own Western culture and those of the societies they were studying. Recognition of this ethnocentricity led to a wholesale rejection of such oppositions as domestic and political, natural, and social that underpinned the notion of the family and kinship (Schneider 1984, Carsten 1997, p. 20). Others accepted that certain distinctions might be culturally specific, but argued that the topic of relatedness could be approached through the folk concepts of the society studied (Carsten 1995, p. 225). This left 'the family' as a Western folk concept, limited in its usefulness to an understanding of the West.

A final strand in late twentieth century anthropological thinking on 'the family' has been contributed by research in Europe and on issues of public concern (La Fontaine 1990, Simpson 1998). While there is a general acceptance that the concept of the family is analytically unhelpful, the term continues to be used. Thus, Segalen and Zonabend, having noted that under the influence of anthropology the 'overfluid' concept of the family has been replaced by analytical concepts such as household and kinship (1987, p. 112), then use 'family' unreflectively and do not discuss why or in what sense. Rather than refer to kinship networks established after divorce, Simpson retains the term family, though qualifying it as 'the unclear family.' He argues for the study of divorce, not as a social problem or a symptom of general breakdown, but as the reorganization of interpersonal relationships in new forms of kindred. His rejection of 'the family' as a theoretical model is explicit, but his continued use of the term implies, as does that of Segalen and Zonabend, an attachment to certain connotations of it that have yet to be explored. The concept of the family may have been discarded from the repertoire of many anthropologists but it is too soon to regard its relegation to the scrap-heap as final.

See also: Dissolution of Family in Western Nations; Cultural Concerns; Divorce and Children's Social Development; Divorce and Gender; Divorce, Sociology of; Family and Gender; Family and Kinship, History of; Family and Psychiatric Illness/Family Therapy; Family as Institution; Family Law; Family Processes; Family Systems in Europe; Family Theory; Feminist-Economist Critique; Family Therapy, Clinical Psychology of; Fatherhood; Female Genital Mutilation; Feminist Legal Theory; Gender and Feminist Studies in Anthropology; Gender and Reproductive Health; Gender Role Stress and Health; Household in Anthropology; Household Production; Kinship in Anthropology; Kinship Terminology; Lone Mothers in Affluent Nations; Love and Intimacy, Psychology of; Male Dominance; Marital Interaction: Effects on Child Development; Marriage; Marriage and the Dual-career Family: Cultural Concerns; Matrifocality; Men's Health; Nation-states, Nationalism, and Gender; Parenthood and Adult Psychological Developments; Parenting: Attitudes and Beliefs; Parenting in Ethnic Minority Families: United States; Partnership Formation and Dissolution in Western Societies; Repartnering and Stepchildren

Bibliography

- Barnes J 1973 Genitor: genitrix: nature: culture. In: Goody J (ed.) *The Character of Kinship*. Cambridge University Press, Cambridge, UK
- Carsten J 1995 The substance of kinship and the heat of the hearth: feeding, personhood, and relatedness among Malays in Pulau Langkawi. *American Ethnologist* 22: 223-41
- Carsten J 1997 *The Heat of the Hearth*. Clarendon Press, Oxford, UK
- Davidoff L, Hall C 1987 *Family Fortunes*. University of Chicago Press, Chicago
- Fortes M 1949 Time and social structure: an Ashanti case-study. In: Fortes M (ed.) *Essays Presented to A. R. Radcliffe-Brown*. Clarendon Press, Oxford, UK
- Fortes M 1958 Introduction. In: Goody J (ed.) *The Developmental Cycle in Domestic Groups*. Cambridge University Press, Cambridge, UK
- Goody J 1956 A comparative approach to incest and adultery. *British Journal of Sociology* 7: 286-305
- Goody J 1983 *The Development of the Family and Marriage in Europe*. Cambridge University Press, Cambridge, UK
- Harris O 1981 Households as natural units. In: Young K, Wolkowitz C, McCullagh R (eds.) *Of Marriage and the Market*. CSE Books, London
- Harris C 1983 *The Family in Industrial Society*. George Allen and Unwin, London
- Janssens A 1993 *Family and Social Change*. Cambridge University Press, Cambridge, UK
- La Fontaine J S 1985 Anthropological perspectives on the family and social change. *The Quarterly Journal of Social Affairs* 1(1): 29-56
- La Fontaine J S 1990 *Child Sexual Abuse*. Polity Press, Cambridge, Oxford, UK
- Levi-Strauss C 1958 (1963) *Structural Anthropology*. Basic Books, New York

- Malinowski B 1913 *The Family Among the Australian Aborigines*. University of London Press, London
- Murdock G P 1949 *Social Structure*. Macmillan, New York
- Robertson A 1991 *Beyond the Family*. Polity Press, Cambridge, UK
- Sahlins M 1976 *The Use and Abuse of Biology*. University of Michigan Press, Ann Arbor, MI
- Scheider M A 1984 *A Critique of the Study of Kinship*. University of Michigan Press, Ann Arbor, MI
- Schneider D M 1965 Kinship and biology. In: Coale A J, Levy M J, Fallers L A, Scheider M A, Tomkins S S (eds.) *Aspects of the Analysis of Family Structure*. Princeton University Press, Princeton, NJ
- Segalen M, Zonabend F 1987 Social anthropology and the ethnology of France: The field of kinship and the family. In: Jackson A (ed.) *Anthropology at Home* [trans. Benthall J]. ASA Monographs 25. Tavistock, London
- Simpson B 1998 *Changing Families*. Berg, Oxford, UK
- Smith R T 1973 The matrifocal family. In: Goody J (ed.) *The Character of Kinship*. Cambridge University Press, Cambridge, UK
- Yanagisako S J 1979 Family and household: The analysis of domestic groups. *Annual Review of Anthropology* 8: 161–205

J. S. La Fontaine

Family as Institution

The family is one of the foundational social institutions in all societies, although the definition of ‘the family’ varies from place to place and from time to time. Thornton and Fricke (1989, p. 130) suggest an inclusive definition that provides a good starting point: ‘family [is] a social network, not necessarily localized, that is based on culturally recognized biological and marital relationships.’ In most times and places, families were responsible for production, distribution, and consumption of commodities, for reproduction and socialization of the next generation, for co-residence and transmission of property. And families generally still are. This article examines the structure of the contemporary family and changes in that structure. It describes alternative family forms and it looks at the processes producing families and dissolving them.

1. Family Mode of Social Organization

Under the family mode of social organization, kin groups pooled resources, including their labor, specialized in particular tasks, coordinated their activities, and connected to the larger community as a unit. This family mode of social organization is often associated with agricultural production, but it appears in a wide range of economic environments (Thornton and Fricke 1989).

The family mode of social organization has altered, however, with other, far-reaching social changes,

including the rise of the market economy, vast increases in productivity with concomitant increases in real income (Fogel 2000), urbanization, changes in ideology toward greater individualization (Lesthaeghe 1983), and changes in the structure of education. All of these changes have shifted decision making and social control away from the family, toward the individual or toward other social institutions. As families have less control over the time and resources of children, they are less able to influence marriage choices—whether, when, and whom to wed. As more people support themselves through wage-based employment rather than through work on a family farm or small business, families have less stake in the property and family connections that a potential marriage partner brings and young adults acquire more autonomy in marriage choices (Caldwell et al. 1983). Urbanization and electronic communication have made the family a less important source of companionship and entertainment now than when most people lived on farms or in villages (Burch and Matthews 1987).

In developed industrial societies such as the USA, the family retains responsibility for reproduction, socialization, co-residence, and transmission of property across generations. It is the main unit of consumption and often also produces considerable amounts of goods and services. Families provide care and support for both the young and the old. Although older adults receive financial transfers and access to medical care from the government in many societies, family members still provide the vast majority of their help and support (Logan and Spitze 1996) and children are almost entirely dependent on their families for financial, emotional, and instrumental support.

2. Structure of the Family

In the USA and many industrialized societies, the structure of the family looks quite different than it did in the mid-twentieth century. In fact, fewer people live in families as traditionally defined and more live in nonfamily households. The rise in nonfamily living can be traced to earlier nestleaving by young adults (Goldscheider and Goldscheider 1993), to delayed marriage and to nonmarriage, to continued high rates of marital disruption with lower rates of remarriage (Cherlin 1992), and to increases in independent living at older ages (Michael et al. 1980). In 1998, 15 percent of all people lived in nonfamily households, 10 percent alone (US Bureau of the Census 1999, Table 16), compared to 6 percent in nonfamily households in 1950 (US Bureau of the Census 1955).

3. Marriage

In the USA, men and women are delaying marriage into their mid-to-late twenties, often entering a cohabitation first. Divorce rates are high and stable, but rates of remarriage have fallen, so that a larger

proportion of adults are unmarried at the beginning of the twenty-first century than in the past. In 1970, unmarried people made up 28 percent of the adult population. In 1996, 40 percent of all adults were unmarried. Seventy-one percent of women born in the early 1950s had married by age 25, compared to 54 percent of those born in the late 1960s (Raley 2000). In fact, the shift away from marriage has been so dramatic for blacks that now a majority of black men and women are not married, compared to about a third of white men and women (Waite 1995).

Similar changes in marriage patterns have taken place in most European countries; recent cohorts are marrying at older ages and over a wider range of ages than in the past. In addition, European countries differ substantially in marriage ages. The Nordic countries of Sweden, Denmark, and Iceland show the highest average ages at marriage for women (around age 29) and the Eastern European countries of Bulgaria, the Czech Republic, Hungary, and Poland the lowest (around age 22). Since societies with relatively high age at marriage also tend to be those in which many people never marry, this diversity suggests that marriage is a more salient component of family in some European countries than others (Kiernan 2000).

Countries in Europe also show a great deal of variation in the proportion of women in marital unions. Marriage is most common in Greece and Portugal, where over 60 percent of women aged 25 to 29 are married, and least common in the Nordic countries, Italy, and Spain where a third or less are.

Age at marriage also seems to be rising in many countries in Africa, including the northern African countries of Egypt, Sudan, Morocco, and Tunisia, and in Kenya, Mauritania, Nigeria, Senegal, Togo, Uganda, and Zimbabwe. Countries in northern Africa also show a rise in age at first birth that is not apparent in the rest of Africa, suggesting an increase in nonmarital childbearing in those countries in which age at marriage is rising (van de Walle 1993). The proportion of women who are currently married is 55 percent in Morocco, 61 percent in Kenya, 58 percent in Haiti (both formal and informal unions), 68 percent in Indonesia, and 63 percent in Bangladesh (Macro International 2000).

4. Cohabitation

Declines in marriage are closely linked to increases in cohabitation, although it is difficult to untangle the nature of the association. In the USA cohabitation has become an increasingly common step in the courtship process; only seven percent of the women born in the late 1940s cohabited before age 25 compared to 55 percent among those born in the late 1960s (Raley 2000). Most couples begin their intimate life together by cohabiting rather than by marrying, so that the form of the union has changed more than its existence. But even when we consider both marriage and cohab-

itation, young adults are less likely to have formed a union now than in the past. Among young women born in the early 1950s about a quarter had not formed a union by age 25, compared to a third of those born in the late 1960s (Raley 2000).

Although a number of European countries have experienced similar increases in cohabitation, some have experienced much more and some much less. Cohabitation is strikingly common in the Nordic countries of Denmark, Sweden, and Finland, and France also shows fairly high levels, with about 30 percent of the women aged 25–29 in cohabiting unions. A group of countries that includes The Netherlands, Belgium, UK, West and East Germany, and Austria shows moderate levels of cohabitation—from 8 to 16 percent of women aged 25–29 are in this type of union. And in the Southern European countries and Ireland cohabitation is rare with less than three percent cohabiting among women aged 25–29 (Kiernan 2000).

In many European countries, women typically are in either cohabitational or marital unions by their mid-to-late twenties. However, over 60 percent of Italian women and 50 percent of Spanish are single, neither cohabiting nor married at these ages, compared to around one in three Portuguese and Greek women. In the Nordic countries and France, about a third of women aged 25–29 are cohabiting, a third are married, and a third are single. Marriage is much more common than cohabitation in all other European countries (Kiernan 2000).

5. Unmarried Childbearing

Changes in marriage have played a central role in increases in unmarried childbearing, which has reached historically unprecedented levels (Bachrach 1998); in 1996 32.4 percent of all births and 44 percent of all first births in the USA occurred to women who were not married (Ventura et al. 1999). But over a quarter of unmarried mothers are cohabiting with the child's father at the time of the birth so their children are living in 'intact,' if unmarried, families (Bumpass et al. 1995).

The proportion of births to unmarried women depends on the share of all women who are married, the fertility of unmarried women, and the fertility of married women. Marital fertility is relatively low in the USA and many European countries. Where nonmarital fertility is also low, as it is in Spain, Italy, and Japan, total fertility is substantially below replacement, but those low-fertility countries with fairly high levels of nonmarital childbearing, like the USA and France, tend to have higher total fertility (Rindfuss et al. 2000).

Unmarried childbearing varies substantially among racial and ethnic groups in the USA. The percentage of births to unmarried women is highest for black women (69 percent), and lowest for Chinese Americans (7 percent), with whites intermediate between

these two extremes at 26 percent. Rates of unmarried childbearing also vary a good deal within Hispanic origin groups, with rates for Puerto Rican women approaching rates for blacks (59 percent), whereas rates for Cuban-origin women (24 percent) approximate those of non-Hispanic whites (Ventura et al. 1999).

As we might expect from differences in the proportion of women who are single and in the proportion in cohabitational unions, unmarried childbearing also varies dramatically across countries in Europe. The Nordic countries, which have a high rate of cohabitation, also have quite high percentages of births to unmarried women; in Norway, for example, 46 percent of births at the end of the twentieth century occurred to unmarried women, most of whom were cohabiting. More than half of births in Sweden and almost half in Norway occur to unmarried women. At the other extreme, only 10 percent of births in Spain, 8 percent in Italy and 3 percent in Greece are to unmarried women. Countries such as Ireland, Belgium, Germany, Portugal, and The Netherlands fall in between (Rindfuss et al. 2000). The social implications of unmarried childbearing depend on both the extent to which these births take place within stable, socially recognized cohabiting couples, like in the Nordic countries, and the extent to which social welfare programs cushion the financial impact on the family of having a single parent. Where few supports for single parents exist, as in the USA, children in these families do less well than children raised in two-parent families, but even under these circumstances, most children raised by single parents are happy and successful as adults (Amato and Booth 1997).

6. Marital Disruption and Union Dissolution

A substantial proportion of all marriages end in divorce or separation due to marital discord. The divorce rate, which reflects the number of divorces in a year relative to the number of married people, rose continuously for more than a century in the USA and many similar industrialized countries, then leveled off at a fairly high level in about 1980 (Goldstein 1999). In the USA, the best estimates suggest that around half of all marriages will be disrupted (Cherlin 1992). The marriages most likely to end include those with no children, with children from a previous union or older children (Waite and Lillard 1991), marriages begun at a young age, and marriages between partners with relatively low levels of education (Martin and Bumpass 1989).

Although high divorce rates make marriages seem unstable, other types of unions are much more likely to dissolve. Cohabital unions show quite high chances of disruption, with a quarter ending in separation within three to four years compared to only five percent of marriages, according to one study (Wu and Balakrishnan 1995). Many cohabitations become

marriages, but these show lower stability than marriages not preceded by cohabitation (Lillard et al. 1995).

7. Alternative Family Structures

The married, two-parent family has been the most common family form in the USA and other industrialized countries for some centuries. But even at the height of the married couple family, many people lived in other types, most often due to the death of one member of the couple before all the children were grown (Watkins et al. 1987). When death ended many marriages relatively early in life, remarriage and stepfamilies were common, as were single-parent families caused by widowhood. The rise of cohabitation and nonmarital childbearing have meant that unmarried-couple families and never-married mother families are now common alternative family forms.

8. Sex

In spite of the sexual revolution, marriage circumscribes the vast majority of sexual relationships. Almost all married men and women are sexually active and almost all have only one sex partner—their spouse. Unmarried men and women have much lower levels of sexual activity than the married, and frequently have no sex partner at all. Cohabiting couples are at least as sexually active as married couples, but are much less likely to be sexually exclusive (Laumann et al. 1994). Thus, the married couple remains the locus of the vast majority of sexual activity.

9. Working Families

Perhaps as dramatic and far-reaching as the alternations in the structure of the family are the changes in the way its members use their time. In the early 1960s in the USA, among those in the prime working ages most married couples followed the male breadwinner/female homemaker model; 56 percent had only one earner. The dual-income family was uncommon—both spouses worked full-time in 21 percent of married couples. By 1997, only a quarter of married couples had one earner. In 44 percent of married couples both spouses worked full-time, and in another 24 percent one worked full-time and one part-time. The shift toward the dual-worker family was even more dramatic for couples with children (Waite and Nielsen 2001). Even by the beginning of the 1970s, most children living with a single parent did not have a parent at home full-time; now most children in married-couple families do not either.

Many public commentators and some scholars of the family argue that the family is in decline and under siege from legal, economic and social change (Popenoe 1993). As evidence, they point to low marriage rates,

high divorce rates, low marital and high nonmarital fertility. And this evidence is compelling, as far as it goes. But most adults are married (although it is a second marriage for many), most have children, most rate their marriage as very happy and place a high value on family life, and most have only one sex partner—their spouse. Most single parents raise happy and successful children. This evidence suggests that the family remains a key social institution, even in its altered state.

See also: Divorce and Gender; Divorce, Sociology of; Families as Educational Settings; Family and Gender; Family and Kinship, History of; Family Bargaining; Family Processes; Family Theory: Economics of Childbearing; Family Theory: Economics of Marriage and Divorce; Family Theory: Role of Changing Values; Gender-related Development; Kinship in Anthropology; Marriage; Marriage and the Dual-career Family; Cultural Concerns

Bibliography

- Amato P R, Booth A 1997 *A Generation at Risk: Growing Up in an Era of Family Upheaval*. Harvard University Press, Cambridge, MA
- Bachrach C 1998 The changing circumstances of marriage and fertility in the United States. In: Moffitt R A (ed.) *Welfare, the Family and Reproductive Behavior: Research Perspectives*. National Academy Press, Washington, DC, pp. 9–32
- Bumpass L L, Raley R K, Sweet J A 1995 The changing character of stepfamilies: Implications of cohabitation and nonmarital childbearing. *Demography* **32**: 425–36
- Burch T K, Matthews B J 1987 Household formation in developed societies. *Population and Development Review* **13**(3): 495–511
- Caldwell J C, Reddy P H, Caldwell P 1983 The causes of marriage change in South Asia. *Population Studies* **37**: 343–61
- Cherlin A J 1992 *Marriage, Divorce, and Remarriage*. Harvard University Press, Cambridge, MA
- Fogel R W 2000 *The Fourth Great Awakening and the Future of Egalitarianism*. University of Chicago Press, Chicago
- Goldscheider F K, Goldscheider C 1993 *Leaving Home Before Marriage: Ethnicity, Familism, and Generational Relationships*. University of Wisconsin Press, Madison, WI
- Goldstein J R 1999 The leveling of divorce in the United States. *Demography* **36**: 409–14
- Kiernan K 2000 European perspectives on union formation. In: Waite L, Bachrach C, Hindin M, Thomson E, Thornton A (eds.) *Ties that Bind: Perspectives on Marriage and Cohabitation*. Aldine de Gruyter, New York, pp. 40–58
- Laumann E O, Gagnon J H, Michael R T, Michaels S 1994 *The Social Organization of Sexuality*. University of Chicago Press, Chicago
- Lesthaeghe R 1983 A century of demographic and cultural change in Western Europe: An exploration of underlying dimensions. *Population and Development Review* **9**(3): 411–35
- Lillard L A, Brien M J, Waite L J 1995 Pre-marital cohabitation and subsequent marital dissolution: Is it self-selection? *Demography* **32**: 437–58
- Logan J R, Spitze G D 1996 *Family Ties: Enduring Relations between Parents and Their Grown Children*. Temple University Press, Philadelphia, PA

- Macro International 2000
www.macoint.com/dhs/pubs/fs/nav.cfm?fn=bd93e.cfm
- Martin T C, Bumpass L L 1989 Recent trends in marital disruption. *Demography* **32**: 509–20
- Michael R T, Fuchs V R, Scott S R 1980 Changes in the propensity to live alone: 1960–1976. *Demography* **20**: 433–47
- Popenoe D 1993 American family decline, 1960–1990: A review and reappraisal. *Journal of Marriage and the Family* **55**: 527–55
- Rindfuss R R, Benjamin K, Morgan S P 2000 The changing institutional context of low fertility. Working paper, Carolina Population Center, University of North Carolina, Chapel Hill
- Raley R K 2000 Recent trends in marriage and cohabitation. In: Waite L, Bachrach C, Hindin M, Thomson E, Thornton A (eds.) *Ties that Bind: Perspectives on Marriage and Cohabitation*. Aldine de Gruyter, New York, pp. 19–39
- Thornton A, Fricke T E 1989 Social change and the family: Comparative perspectives from the West, China and South Asia. In: Mayone Stycos J (ed.) *Demography as an Interdisciplinary*. Transaction, New Brunswick, NJ, pp. 128–61
- US Bureau of the Census 1955 Household and family characteristics: March 1954. *Current Population Report*. US Government Printing Office, Washington, DC
- US Bureau of the Census 1999 Household and family characteristics: March 1998 (Update). *Current Population Report*. US Government Printing Office, Washington, DC
- van de Walle E 1993 Recent trends in marriage ages. In: Foote K A, Hill K H, Martin L G (eds.) *Demography Change in Sub-Saharan Africa*. National Academy Press, Washington, DC, pp. 117–52
- Ventura S J, Martin J A, Curtin S C, Mathews T J 1998 Report of final natality statistics, 1996. *Monthly Vital Statistics Report*, Vol. 46, No. 11, Suppl. National Center for Health Statistics, Hyattsville, MD
- Ventura S J, Martin J A, Curtin S C, Mathews T J 1999 Births: Final data for 1997. *National Vital Statistics Report*, Vol. 47, No. 18. National Center for Health Statistics, Hyattsville, MD
- Waite L J 1995 Does marriage matter? *Demography* **32**: 483–508
- Waite L J, Lillard L A 1991 Children and marital disruption. *American Journal of Sociology* **96**: 930–53
- Waite L J, Nielsen M 2001 The rise of the dual-worker family: 1963–1997. In: Hertz R, Marshall N (eds.) *Women and Work in the Twentieth Century*. University of California Press, Berkeley, CA
- Watkins S C, Menken J A, Bongaarts J 1987 Demographic foundations of family change. *American Sociological Review* **52**: 346–58
- Wu Z, Balakrishnan T R 1995 Dissolution of premarital cohabitation in Canada. *Demography* **32**(4): 521–32

L. J. Waite

Family Bargaining

1. Introduction

The concept of bargaining refers to a particular approach to decision-making in situations of conflict. Such conflicts could be resolved in other ways: dictatorially, for instance, through the use of force or authority, or democratically through majority vote. Bargaining, on the other hand, refers to the attempts

by the different parties to a dispute to threaten, cajole, or persuade others to settle it in their favor. Bargaining takes different forms because it occurs in different contexts. This article reviews different social science approaches to the analysis of bargaining in the specific context of family-based households and the insights they provide into its distinctive features. Two broad approaches will be distinguished: the neoclassical and the 'institutional.'

2. *Neoclassical Economics and Household Decision-making*

Neoclassical economists tend to explain human behavior through abstract mathematical models and econometric approaches to data analysis. They therefore place considerable value on 'parsimony,' or the ability to explain complex phenomena through a very limited set of variables. This is evident in early models of household decision-making which assumed it to be the responsibility of the household head, acting as 'benevolent dictator': 'benevolent' because he was assumed to take account of the welfare of all members in his decisions and a 'dictator' because he was able to impose his decisions on them (such models implicitly assumed a male household head, an assumption which is often but not invariably appropriate). However, while this dual assumption considerably simplified attempts to model household decision-making, it ruled out the possibility that power and conflict might have any role to play in the process (Becker 1981)

Two bodies of empirical evidence challenged this representation of household decision-making. The first showed that access to resources by individual members influenced the allocation of intrahousehold resources, suggesting that household decision-making processes were susceptible to the economic leverage exercised by different members (Thomas 1993, Hoddinott and Haddad 1995). The second documented phenomena such as domestic violence as well as systematic inequalities in household welfare along age and gender lines (Momsen and Townsend 1987), further undermining the assumption of conflict-free, welfare-maximizing, decision-making. Bargaining models of household decision-making emerged in response to such evidence. Their starting premise was that individuals cooperated to form households because the resulting benefits exceeded those obtainable from living alone. Where members had conflicting views about certain decisions, they might seek to bargain over the final outcomes. However, not all members had equal bargaining power and bargaining models largely were concerned with establishing the determinants of differentials in bargaining power within the household. These were seen to depend on the kind of 'threats' that different members could make in order to influence the decision-making process. The ultimate threat available was the threat of

'exit,' which in the context of the household, implied withdrawal from its membership.

However, the credibility of such a threat depended on the kind of resources that members were able to command as individuals rather than as household members. Initially, the 'threat' position was equated with access to individual economic resources, for instance, labor market earnings. It was expanded subsequently to include aspects of the social context which were likely to impinge on the threat positions of individuals as social groups. For female members, for instance, individual threat positions might be influenced by sex ratios in the relevant marriage market, laws concerning alimony/child support settlements, women's ability to return to their natal homes after marital breakdown, and the cultural acceptability of outside work.

Quantitative tests of bargaining models supported the assumption that individual economic characteristics, such as wages and incomes, as well as broader structural factors, such as legal provision relating to divorce settlements, did influence decision-making outcomes within the household (Hoddinott and Adam 1998, McElroy 1990, Schultz 1990, Jones 1986). They confirmed the feminist argument that, by and large, male household members exercised greater bargaining power than females, a reflection of their higher earning capacity as well as gender asymmetries in the legal structure. They also confirmed anthropological findings that men used their stronger bargaining position to claim a disproportionate share of household resources for their own consumption as well as more leisure (Dwyer and Bruce 1988). They pointed to the importance of women's access to paid employment for helping to shift the balance of decision-making power within the household. However, the insights provided by economic models have been limited by their focus on the economic determinants of bargaining power and their failure to consider its social dimensions. Yet it could be argued that it is its social dimensions which most help to distinguish the process of bargaining within the family from bargaining in any other arena of social life.

3. *Institutional Approaches and Family Bargaining*

Institutional approaches to household bargaining refer to attempts by nonneoclassical economists, sociologists, and anthropologists to account for the specificities of the household as an institutional form. These see households as an institutional response to the human need for long-term stable environments in which to bear and bring up children, to care, and be cared for, through sickness, disability, and old age and to plan for the future in a world characterized by uncertainty. What gives the household its institutional advantage in achieving these goals is the close inter-

twining of emotions and interests that characterize its relationships. Despite the different forms households take in different parts of the world, its relationships are generally familial. Its members are related by blood or marriage, they have lived together over considerable periods of time, they care for each other, and in any case, trust each other more than they trust strangers.

However, cooperative behavior within the household is not left entirely to the spontaneous impulse of members. It is also underwritten by a series of 'implicit contracts' that spell out the claims and obligations of different members to each other. Given the importance of the activities carried out within the family to the reproduction of social life in the wider society, these contracts typically embody social norms and beliefs about the meaning of the family in different cultural contexts. And given the importance of gender relationships in the key activities of the family, contracts between women and men within the family typically embody social norms and beliefs about gender differences in the wider society.

As long as household members have more to gain from their membership of the household than from going it alone, they have an incentive to abide by the terms of intrahousehold contracts, including its rules about decision-making. However, when these gains are less certain, how household members respond, whether or how they seek to achieve more favorable outcomes, will depend on what they are stand to gain from such action compared with what they stand to lose. Institutional approaches to cooperation and conflict within the household differ from neoclassical approaches in two important ways. First of all, they explain the rationale for cooperation within the family in affective and contractual terms rather than purely instrumentalist ones. Household members do not make decisions and seek to bargain as unrelated individuals, differentiated only by their resource endowments, but as members of families, with ascribed roles and responsibilities. Familial connections also mean that what individual members value about their membership of particular households, and hence what they stand to lose, cannot be captured by a materialist calculus alone. Nelson (1996), for instance, suggests that along with concerns with material well-being, household members will also seek to balance their need for 'affiliation,' the need to love and belong, with the value they attach to 'agency,' the capacity to define and act in their own interests. These different dimensions of well-being often involve trade-offs so that an individual may gain greater 'agency' by withdrawing from the household but have to sacrifice affectivity or material well-being.

At the same time, there may be situations where members may feel strongly enough about a decision to seek to influence its outcome. Because such action can, in certain cases, lead to open conflict, the likelihood of it happening will be determined partly by the extent to which the member, or members, in question feel able

to jeopardize their membership of the household, in other words, on the strength of their fall-back positions. The second important difference between institutional and neoclassical approaches to bargaining relates to how they conceptualize differences in fall-back positions. While the primary focus of neoclassical approaches has been on differences in individual economic resources, institutional analysis emphasizes the much wider range of constraints and possibilities which reflect the household as a specific institutional form and create social differentials among individual members, or categories of members. These points have been articulated most clearly in the literature dealing with the gender dynamics of household decision-making and will be elaborated by drawing from this literature.

4. Gender and the Structure of Constraint

Neoclassical approaches to household economics tend to treat differences in resource endowments as the product of individual efforts or as 'given' by genetic features like the sex of individuals. Institutional approaches, by contrast, suggest that these differences are influenced significantly by social rules and norms which prescribe the distribution of resources, the division of labor and responsibility for decision-making within the household. While it is largely men who are favored by these rules and norms, gender inequalities vary considerably across societies, leading to wide variations in whether, and how, women have influence in household decision-making. For instance, most inheritance laws and customs favor male members of family but vary considerably in how they treat women. In some societies, women can own land and property in their own right. In others, they have rights of access or use, but only as members of particular households. In yet others, they are effectively property-less. Similarly, social norms tend to assign primary responsibility for domestic work to women in most societies, but vary considerably in whether they permit women to engage in economic activities as well. In societies that practice female seclusion, for instance, women are proscribed culturally from undertaking public forms of employment. They are consequently likely to be far more dependent on male members of the household, and their capacity to influence household decision-making processes correspondingly curtailed, compared with societies where women have a socially-sanctioned role in production.

Finally, there are also cross-cultural variations in the access that women enjoy to extrahousehold social networks which will also influence the extent to which they are dependent on their immediate household members for provision and security. For example, in the Indian context, the ability of women to retain contacts with their natal families after marriage has been found to be an important factor in determining

the degree of agency they are able to exercise within marriage.

These variations in the contractual relations of the family create very different sets of gender interests in different social contexts and very different possibilities for women, as the subordinate category, to pursue their own interests. In contexts where women have recognized access to resources and economic opportunities, as in parts of sub-Saharan Africa, the relationship between spouses is often characterized by an overt bargaining element. Attempts to infringe on women's economic autonomy can result in open conflict, often ending with women leaving their husbands to set up their own households. By contrast, in many parts of South Asia, where women are denied access to critical productive resource or the ability to engage in economic activities, there is less scope for overt bargaining with more powerful male members. Conflicts are likely to be suppressed and dissent expressed covertly through what have been described as the 'weapons of the weak': deception, manipulation, and subversion.

5. *Gender and the Structure Of Preferences*

Neoclassical approaches to household decision-making do not pay a great deal of attention on what motivates people in their behavior. It is assumed that each member will have their individual preferences, depending on what gives them 'utility' or satisfaction. Because these are subjectively determined, they will be idiosyncratic to individuals and distributed randomly across the membership. However, evidence that women tend to behave in systematically different ways from men within the household, that they are more likely to put the collective interests before their own and to allocate resources under their jurisdiction to collective, rather than individual forms of consumption, has been given different explanations.

One set of explanations focus on differences in interests and fallback positions. Thus, although women often appear to give value to the collective well-being over their own as a result of prevailing ideologies of maternal altruism, Whitehead's (1981) work in Ghana leads her to suggest that such altruism often contained a dimension of self-interest. Women's fortunes are bound up with the fortunes of the household collectivity to a larger extent than those of men and their longer-term self-interests better served through altruistic forms of behavior which helped to preserve household solidarity. According to Agarwal (1997), evidence from South Asia that women tend to subordinate their own well-being to the well-being of others reflected their awareness of the weakness of their fall-back positions and a resulting caution in acting overtly in their own self-interest. An alternative set of explanations suggest that these differences in behavior reflect actual differences in the values and

priorities, differences which in turn reflect social definitions of gender identity. Social norms and beliefs do not simply determine gender differences in resource endowments; they also justify and legitimate these differentials by ascribing very different aptitudes and dispositions to men and women, differences which are internalized by them as aspects of their gender identities.

For instance, studies in the US and Europe suggest that the processes of socialization through which men and women acquire their gender identities tend to stress agency and 'separative' notions of selfhood for men, and affiliation and 'connected' notions of selfhood for women (Chodorow 1978). In the South Asian context, it has been argued that gender identities are linked to differentiated notions of individual self-worth (Sen 1990). Women are seen to accept the lesser status accorded to them by a society and to define their own sense of well-being in terms of the well-being of other family members.

If men and women within the family have different values and preferences, bargaining can take the form of trade-offs between different dimensions of individual well-being noted earlier. Thus, a wife may agree to love, honor, and obey her husband in exchange for his promise to love, honor, and protect her: both value affiliation but the wife trades some amount of agency in return for material security. Alternatively, a husband may accept greater agency for his wife (and perhaps less for himself) by agreeing to her paid employment but only on condition that she does not neglect the housework and hence his material comforts. However, these negotiations and trade-offs will not be conducted on equal terms because the gender differentials in values and priorities do not have symmetrical implications for the bargaining process. If women do indeed attach greater value to affiliation or less value to their own self-worth, it is likely to work to men's advantage in situations of conflict. First of all, women will be less willing to disrupt household solidarity by open expression of dissent. Second, even if they do decide to engage in explicit bargaining, they will find it harder to hold out for their own self-interest or to threaten to use the exit option, if things do not go their own way.

6. *Conclusion*

Institutional approaches to family bargaining differ from neoclassical bargaining approaches because they focus on households as sets of social relationships, rather than a collection of individuals, on the structural determinants of bargaining power rather than the individual, and on the role of norms and values in shaping preferences rather than taking preferences as given. They consequently offer important insights into the specificities of households as social institutions and into what makes bargaining within the family different from bargaining in other arenas of life. Because

families are constructed around relations of unequal, but intimate, interdependence, power within the family is rarely exercised in a naked form. Violence, and the threat of violence, exist, but more often, men are able to resort to the authority vested in them in their capacity as household heads or as primary breadwinners to protect their own interests. The association of authority with male roles within the family explains why decision-making roles are not reversed automatically in households with unemployed husbands and employed wives or where women earn more than their husbands.

Equally, however, despite their stronger fall-back position, men within the family do not always get their own way. Bargaining processes within the family encompass forms of 'threats' as well as 'persuasions' that cannot be reduced to the economics of fall-back positions. They reflect the fact that the family is an intensely personal arena of life and that family relationships are suffused with feelings and emotions to a much greater extent than any other institution in society. Because families are based on ideologies of affectivity and shared interests, bargaining within the family rarely takes the form of open self-interest. Instead, it is conducted as negotiations over meanings and interpretations of, for instance, what constitutes the collective interests of family members, what their proper roles and responsibilities should be, and the basis of claims on collective household resources.

This means that, despite their subordinate position, women have a valuable resource on which they can draw in devising their bargaining strategies: their intimate knowledge of other family members, of their feelings and emotions, as well as their values and priorities. This knowledge can be translated into various forms of 'threats,' including persistent complaining, ridicule, withdrawal into silence, and withholding sex from husbands. Alternatively, they may employ more persuasive tactics. These may be discursive in form, for instance, appeals to shared interests in children's welfare as the justification for pursuing some preferred course of action. Or they may be more practical, involving strategies of 'wielding and yielding,' trading concessions in some areas in some areas to win concessions in others.

However, the fact that, despite the weakness of their fall-back positions, women are able to engage in a variety of bargaining maneuvers to achieve their priorities, should not be taken to negate the importance of improvements in their fall-back positions. Such improvements can bring about qualitative transformation in the asymmetries of the bargaining process. They can, for instance, lead to more open forms of bargaining between equals in place of the covert tactics that make up the 'weapons of the weak.' They can also open up for negotiation and contestation of aspects of household inequality that were previously considered nonnegotiable, for instance, in the domestic division of labor.

Research is needed to establish which are the most critical opportunities and constraints which make up women's fall-back position in different societies, how changes in these opportunities and constraints come about, and what role can be played by policy and public action to facilitate the right kind of changes. Women's paid activities has been found to be one of the most consistent predictors of their capacity to exercise agency within the household, both in advanced industrialized countries as well as in poorer ones. However, their ability to take advantage of employment opportunities is itself a matter for bargaining and negotiation, particularly in societies which practice female seclusion. The fact that women have, nevertheless, been able to do so suggests that other factors also came into play in negotiating these opportunities. Economic resources are important, but they may require changes of the kind brought about by enabling policies and social movements before their transformatory potential can be realized.

In addition, research is needed on the pathways by which changes in the wider structures of constraints and opportunities transform the unequal interdependencies of the family. It is possible that they will vary considerably according to local beliefs and values. In cultures that are strongly individualist in their outlook, women may use improvements in their fall-back position to seek greater independence, perhaps exercising their exit options. On the other hand, in cultures that emphasize collective values, they may prefer to bargain for greater equality in decision-making, the democratization of household structures. Attempts to evaluate the transformatory potential of different kinds of change in women's fall-back position will 'misread' the nature of the process if they are premised on an inappropriate set of values.

See also: Conflict and Conflict Resolution, Social Psychology of; Conflict/Consensus; Domestic Life and Material Culture; Economic Development and Women; Exchange in Anthropology; Exchange: Social; Family and Gender; Family and Kinship, History of; Family Theory: Feminist-Economist Critique; Feminist Economics; Gender, Economics of; Group Decision Making, Cognitive Psychology of; Household in Anthropology; Household Production; Infancy and Childhood: Emotional Development; Kinship in Anthropology; Land Rights and Gender; Love and Intimacy, Psychology of; Marriage and the Dual-career Family; Cultural Concerns; Rural Industrialization in Developing Nations and Gender; Temperament and Human Development

Bibliography

- Agarwal B 1997 'Bargaining' and gender relations: Within and beyond the household. *Feminist Economics* 3(1): 1-51
Becker G 1981 *A Treatise on the Family*. Harvard University Press, MA

- Chodorow N J 1978 The reproduction of mothering: Psychoanalysis and the sociology of gender. University of California Press, Berkeley, CA
- Dwyer D, Bruce J 1988 *A Home Divided: Women and Income in the Third World*. Stanford University Press, Stanford, CA
- Hoddinote J, Adam C 1998 Testing Nash bargaining household models with time series data: Divorce law reform and female suicide in Canada. *Food Consumption and Nutrition Division, Discussion Paper 52*. International Food Policy Research Institute, Washington, DC
- Jones C 1983 The mobilisation of women's labor for cash crop production: A game theoretic approach. *American Journal of Agricultural Economics* 65(5): 1049–54
- Kabeer N 2000 *The Power to Choose: Bangladeshi Women and Labor Market Decisions in London and Dhaka*. Verso, London and New York
- Kandiyoti D 1988 Bargaining with patriarchy. *Gender and Society* 2(3): 274–90
- McElroy M B 1990 The empirical content of Nash-bargained household behavior. *Journal of Human Resources* 25(4): 559–83
- Momsen J H, Townsend J 1987 *Geography of Gender in the Third World State*. University of New York Press, Albany, NY
- Nelson J A 1996 *Feminism, Objectivity and Economics*. Routledge, London
- Schultz T P 1990 Testing the neoclassical model of family labor supply and fertility. *Journal of Human Resources* 25(4): 599–634
- Sen A K 1990 Gender and cooperative conflicts. In: Tinker I (ed.) *Persistent Inequalities*. Oxford University Press, Oxford, UK
- Thomas D 1993 The distribution of income and expenditure within the household. *Annales D'Economie et de Statistique* 29: 109–35
- Whitehead A 1981 I'm hungry, Mum: The politics of domestic budgeting. In: Young K, Wolkowitz C, McCullagh R (eds.) *Of Marriage and Market: Women's Subordination in International Perspective*. CSE Books, London

N. Kabeer

Family Business

Family business is an interdisciplinary field involving all professionals who advise, study, and work in family firms addressing their structural and transitional problems. Family firms are companies of various sizes controlled by one or more owners tied by family relationship or solid alliances (see Corbetta 1995, Aronoff et al. 1996, Neubauer and Lank 1998). Even if this control-based definition is widely accepted, some authors propose others based on the owners' managerial involvement. Other authors do not consider either firms controlled only by the founder or those controlled by unrelated partners as family businesses. Another group tends to identify family businesses with small and medium-size enterprises where the overlapping of family and firm is particularly high. Whichever the accepted definition may be, it includes

various types of family firms that can be classified using different dimensions. Family business specialists' fields of origin are equally varied and include such disciplines as business, law, psychology, sociology, and history.

A study carried out on the articles published in *Family Business Review* from 1988 to 1997, shows that over 70 percent of the academic articles were written by authors in the field of business management (Dyer and Sánchez 1998). In this discipline, family business literature mainly overlaps with studies on entrepreneurship and management issues in small and medium enterprises (SMEs) (see *Entrepreneurship*). However, it distinguishes itself from such studies on at least three levels. First of all, family business literature investigates family firms' evolution from succession to the founding owner up to following stages in which the firm may have become very large and even public but it may still be controlled by one or few families tied by family relationship or solid alliances (Gersick et al. 1997). Second, family business management literature deals with the entrepreneur's characteristics and entrepreneurship training and developing processes in a peculiar way.

On one hand, these topics are always related to owning families' characteristics and dynamics, and on the other hand much more attention is devoted to the issue of how to pass on entrepreneurial capabilities rather than on how to start a new venture. Finally, family business literature deals with topics that are not investigated by the streams mentioned previously, such as managing the ownership group, governing the family firms and managing the family–firm relationship.

Family businesses are dynamic systems that include two subsets, the family and the firm, which are characterized by their own evolutionary patterns and are linked to each other by many dynamic relations. Hence a scholar may be considered a scholar at family business when they deal with the whole system: a psychologist interested in parents–children relationships addresses family business issues only when they consider how these relationships' dynamics impact the way the business operates. A business management scholar interested in strategy deals with family business only when investigating the relationships between the owning families' structure and evolution and business continuity.

1. Origins and Development of Family Business

Even if some family business articles appeared in the 1950s, 1960s, and 1970s, only at the beginning of the 1980s was family business born as a field acknowledged by a number of North American scholars and practitioners. A fundamental contribution was made by a management researcher (Richard Beckhard, Management and Organization Behavior Professor at

MIT Sloan School of Management for more than 20 years), by a psychology researcher (Harry Levinson, professor at the Psychiatry Department at Harvard Medical School), and by a consultant (Léon Danco, CEO of the Center for Family Business in Cleveland). In the mid-1990s they have been recognized publicly as family business pioneers and interviewed in the 1994 *Family Business Review* issue (Vol. 2) dedicated to the field's history; in these interviews they all admit to how difficult it was to make the scientific community accept the existence of family business as a special discipline.

Until then business and family (and related studies) had been dealt with as two separate, autonomous fields: it was difficult to convince scholars in either that it would be useful to get acquainted and cooperate. Moreover, the first studies in family business were case descriptions without any interpretative framework, which made it even more difficult to open a scientific debate. Finally, most family business professionals were practitioners at that time; as such, they were not well respected as researchers by the scientific community. Further, they were not very committed to publishing in order to share their knowledge and experiences with others.

Generally speaking, family firms traditionally have been considered—and not only by academics—as a residual of the past to overcome, beginning from the study by Berle and Means (1932) who considered separation between ownership and control as a fundamental condition of development for modern economies. Not surprisingly, the media were used to dealing with family businesses only to highlight how disputes among family owners or nepotistic behaviors negatively impacted the business.

In 1984 something of decisive importance in the development of family business occurred: a group of North American scholars and practitioners, coordinated by Barbara Hollander, gave rise to the Family Firm Institute (FFI). By means of its various services, FFI has become a special forum for knowledge sharing and development. Particularly, annual conferences and the *Family Business Review*—founded by Ivan Lansberg in 1988—have been fundamental tools that helped develop theory and practice. In the leading article of the first issue, Lansberg himself provided significant data about how early the stage of the field's development was: according to a study conducted on the Social Science Citation Index from 1977 to 1987, only 53 articles had been published in the US on family business issues, as compared with the 813 on entrepreneurship.

In 1990, the Family Business Network (FBN) was founded in Europe. Despite its different membership policies (there are more family business members and less professionals), FBN fulfills a similar mission by spreading family business knowledge and fostering a scientific and cultural debate on this topic.

The increasing number of monographic issues on family business published in such reviews as *Organi-*

zational Dynamics (summer 1983), *Business History* (October 1993), and *Entrepreneurship Theory and Practice* (N. 1, 1994) shows that rising attention has been devoted to this new discipline by scholars in other fields as well.

In the US and in other parts of the world, some universities followed the increasing interest on the part of scholars and other stakeholders towards family business. Nowadays more than 100 universities and business schools in the world offer specific research and training programs on family business issues (Aronoff 1998). Most of these programs have been developed in family business centers, which—even if they are often based in management departments—usually get contributions from various disciplines.

Different factors have persuaded individuals and institutions to invest resources in exploring family business. Among these is the diffusion of data about the importance of family businesses in modern economies, evidence that family firms can grow big and cannot just be identified with SMEs, and consciousness of the difficulties for family firms to last beyond the first generation. Depending on the adopted definition, family firms represent from 30 percent to 70 percent of the GDP in the main industrialized countries, and above all, they seem to be the enterprises with the highest rate of growth in employment.

Looking into the future, family business seems to be in a delicate phase. In order to gain importance in the academic context this new discipline needs to be acknowledged from an institutional point of view as well; for this reason, a key challenge—at least in Europe—is to increase the number of universities and business schools that create chairs in family business. Probably this will happen mainly in management departments, as it occurred in the closely related field of entrepreneurship.

Concerning the practitioners' sphere, the family business field is growing very rapidly, with many advisors coming from a variety of professions and entering the field. The main problem in the near future is the difficulty in defining the minimum standards necessary to qualify people as family business professionals. The FFI is working on a research project that started with identification of four major content areas—Behavioral Science, Finance, Law, and Management Science—and is going on with the definition of fundamentals that competent family business advisors, consultants, and educators should know to work effectively in those subfields.

2. Research Areas and Achieved Results

Family businesses are a composite world including different archetypes. Many authors have proposed classification criteria. According to a widely accepted distinction there are family firms owned and managed either by a single owner or by two or more siblings or

by cousins' teams. A fourth archetype consists of those family firms where equity has been opened up to partners other than the founding family (Corbetta 1995). Usually, changing from one type to the others the company grows in size. But notable exceptions make it impossible to significantly correlate company size with ownership structure. The distinction mentioned here cannot be confused with the generation-based perspective, for example, first generation family enterprises can be controlled either by a single owner or by a sibling partnership or by a cousin consortium.

For every archetype identified, a variety of research topics can be defined, such as:

- (a) family structure and management;
- (b) ownership structure and management;
- (c) company governance;
- (d) company management;
- (e) relations among family, ownership, governance and management; and
- (f) transitions into another type of family business.

Other general topics of research concern the comparison among different types of family businesses, between family and nonfamily businesses (in terms of values, culture, strategies, and performances), among family enterprises operating in different countries, the relationship between enterprises and consultants, and the role of women.

In terms of necessary competencies, many family business scholars believe that most research topics need interdisciplinary team effort; in particular, researchers in management, psychology, and sociology should cooperate to capture the distinctive facets of family business.

2.1 Studies Before the 1980s

The first family business studies mostly concerned founder succession issues, that is either transition from the controlling owner archetype to another or recycling of the same archetype with different actors (Danco 1975). Such studies typically had the following characteristics:

- (a) they concerned small and medium sized firms with simple management structures;
- (b) they took into consideration only the simplest family situations—one founder, one heir—overlooking more complicated but also more widespread circumstances, for example, those of founding partners' teams or those where the transition occurs between multiple founders and multiple heirs;
- (c) they centered their attention on the two people involved in leadership transition—the founder and the heir—without considering other family and business members;
- (d) they almost exclusively dealt with the issue of passing on managerial roles overlooking ownership changes; and
- (e) they considered the family as an obstacle to the effective and efficient management of the firm and

suggested a clean distinction between family and enterprise as an indispensable condition for business continuity.

Even if many of these studies, which were clearly influenced by some streams of rationalism (Hollander and Elman 1988), have been revisited or even become outdated, they have actually achieved some important results. First of all, they have opened up a new, interesting stream by explicitly recognizing the differences—in terms of roles, goals and needs—and the mutual influence between the family and the business. Moreover, these studies suggested some interpretative frameworks still valid today such as: succession considered as a multiphase process and not an event; the difficulties of this transition coming from the variety of roles—parent, owner, and manager—simultaneously played by the founder; the importance of training for people; the need to consider the firm's continuity as a fundamental goal to be achieved to be able to consider generational transitions as successful ones.

Many of the first scholars were focused on small and medium sized enterprises, which made them also study managerialization processes in relation to company growth patterns.

2.2 The Subsequent Studies

In the following years the family business field was consolidated; its boundaries were clarified and new topics were identified. In particular, several contributions concerned the following:

- (a) the definition of family businesses and their classification, going beyond complete overlapping with small and medium sized enterprises;
- (b) the definition of the main dynamic relationships between family and business using interpretative frameworks based on different stages of the firms, people, and families' life cycles (Davis and Tagiuri 1982, Beckhard and Dyer 1983, Gersick et al. 1997);
- (c) succession processes, widening the range of considered situations and variables; and
- (d) new topics such as management of extended families, family business governance, and strategy. Concerning the first topic, numerous studies have been focused either on the role and way of implementing family protocols or on the way of functioning of specific bodies like the family council. With regard to the second topic, attention has been centered on the structure and the functioning of boards of directors (Ward 1991, Neubauer and Lank 1998). On the third topic, interesting contributions have been published on internationalization processes and strategic management in general (Ward 1988).

During this period, scholars have reached a greater consensus about the definition of family business and about the classification of different family firms' archetypes. The field has experienced significant development and has involved scholars from other

disciplines (e.g., finance and corporate governance), which has generated new research opportunities.

One very important achievement is that the family is no longer considered an obstacle to the firm's development. Longitudinal studies of family business cases and quantitative researches definitely have shown that the owning family can be a strong point for the firm as it provides such key factors as long-term orientation, education to entrepreneurship, values, commitment and personal sacrifice, reputation, speed in decision-making, and lower transaction costs (Gallo 1995). Social scientists' studies are considered an important basis to build up theories and models that can be helpful for family business success in the long run. Theories on family life cycles and adult development have been used to analyze family firms' stages of development, which are seen as the joint result of family, ownership, and business dynamics (Gersick et al. 1997). Family systems' theories are used to suggest how to better manage interpersonal relations, communications, and to handle conflicts through their predictable stages in the family business context (Barnes 1988, Kaye 1991). Interpretative models have recently been proposed considering the family just as important as the company as a determinant of family business success; so, both institutions are included in the model at the same level of detail, as purposive social systems transforming resources and constraints into outcomes by interpersonal and resource transactions (Scannel Trent 1999). The same studies have investigated family firms' diffusion by surveying households instead of companies.

Concerning succession, which is still the hottest topic in family business literature, research findings highlight the need to plan the transition: the importance of successors' personal development both inside and outside their own family firms; the critical role of the entrepreneur and of their ability to delegate; the importance of communication; and the useful contribution of nonfamily members either as outside directors or as tutors and mentors to the younger generation. Efforts have been made to broaden the focus of research: a variety of situations has been addressed in addition to the 'simple' one featuring one entrepreneur with only one heir; ownership transitions and their financial implications have emerged as relevant succession topics besides the classical one concerning company leadership. Efforts have also been made to offer deeper insights on the 'classical' issue of succession to the founder. Numerous studies have built upon the existing literature by exploring the complexities that characterize this type of succession as well and that had been overlooked so far. Relationship between founder and the whole organization, and how organizational culture can be created and passed on so as to enable company continuity have been investigated (Schein 1983). The broad range of stakeholders involved besides the founder and the successor involved has been analyzed, highlighting

how each of them may resist coping with the generational transition issue and by this means create obstacles to the whole process (Lansberg 1988). The issue of how entrepreneurial attitudes and capabilities can be learnt and taught has been further explored (Aronoff et al. 1996).

Finally, relevant insights have been provided on family business governance, particularly on the structure and the functioning of boards of directors and on their contribution to company development.

2.3 Future Developments

A recent article that appeared in the *Family Business Review* has identified some trends in family business studies (Aronoff 1998). Succession will remain one of the most interesting areas in the broader perspective of generational transition. A new research topic could concern possible roles for retiring entrepreneurs. Generational transitions will increasingly lead to both owners and family managers' teams; conditions under which such teams can effectively work will be one of the hottest topics in the nearest future. Companies' strategic context is significantly changing because of the impact of new technologies and globalization. Family businesses' capability to compete in such a context will make a relevant research topic as well. Another key issue will be the impact of increasing managerial and financial sophistication on both the culture and behaviors of family businesses. The other topics mentioned by Aronoff concern the roles of women, consultants, and educators.

Interesting developments may also come from the increasing focus on big family businesses that characterizes some studies in progress. Certainly this will lead not only to identifying and offering more sophisticated solutions to the already well-known problems of smaller family businesses, but also to investigating different ways to govern and manage large family fortunes in terms of both personal attitudes and necessary skills.

Studies on family businesses' developmental stages have not yet deeply explored transitions from one stage to another. Relevant insights on this topic could be provided by joint efforts of management and psychology researchers.

Finally, other topics are likely to undergo important developments. We can mention, for instance: comparisons among different archetypes of family firms, among family businesses from different countries, and between family and nonfamily enterprises in terms of resources, strategies, and performances; nonfamily managers' characteristics and roles; family businesses' selling processes; and the impact of some variables on family firms' continuity. Research efforts could also address family values and ownership structures.

Research topics still seem to be driven by both family enterprises and family business professionals, necessary because they are the main sponsors of

research projects. On the other hand, the strong link with family businesses is helpful in making the relationship between theory and practice mutually beneficial. Obviously scholars should avoid focusing on short-term needs and commit themselves to identifying long-term research goals and strategies.

3. Methodological Issues

Practitioners, as it was stated before, have played a predominant role in the formation of family business as a discipline. At least until the early 1980s, as a consequence, research methodologies were not emphasized. Even those scholars who have been in the field from the very beginning were much more concerned in defining boundaries and topics rather than in carrying out methodologically acceptable research projects. In the first issue of the *Family Business Review*, in 1988, only 14 percent of articles were research based. Notably, 64 percent of the articles published in the 10th issue of the same review reported qualitative or quantitative research findings (Dyer and Sánchez 1998).

In any case, family business research still has to cope with at least three specific problems. The first and most important one is that family business, as should be evident, requires that contributions be provided by professionals from different disciplines, particularly social sciences such as sociology, psychology, and business management. On one hand, it is not easy that a single person dominates both; at the same time, interdisciplinary collaboration is not well developed, probably because of the high costs of joint research due to the very different starting points in terms of knowledge, languages, and methodologies. On the other hand, since in many fields radical advances come from collaboration between authors in different disciplines and professions, such collaboration should be strongly encouraged.

The second problem concerns information availability. Entrepreneurs and their families are known to be very reserved and protective of their privacy. Moreover, most family businesses do not have to disclose much information about their structure or about their performances. Researchers are trying to overcome this problem by collaborating with institutions such as banks that can provide family business databases or by designing research projects based on primary sources and making use of interviews, questionnaires, and longitudinal case studies.

The third relevant problem regards the specific need of this field to encourage collaboration between scholars and practitioners. Increasing interest in family business among younger Ph.D.'s is definitely improving the quality of research. Some difficulties could arise as young scholars often focus their attention on topics that are not considered particularly relevant by family businesses, which may hinder communication between academics and other stake-

holders. In other fields such as entrepreneurship, such incommunicability has led to very disappointing results (Brockhaus 1994).

Finally, given the interdisciplinary nature of the family business field, relevant developments might result from the application of theoretical frameworks and analytical tools developed in other knowledge areas such as the system dynamics approach (which might be useful in explaining the dynamics of the family-business system) or the resource-based view (which might lead to a better understanding of the relationships between the business and the family's resources and the firm's performances).

See also: Business History; Entrepreneurship; Entrepreneurship, Psychology of; Ethnic Groups/Ethnicity: Historical Aspects; Family as Institution; Firm Behavior; Retail Trade

Bibliography

- Aronoff C E 1998 Megatrends in family business. *Family Business Review* 3: 181–5
- Aronoff C E, Astrachan J H, Ward J L 1996 *Family Business Sourcebook II*. Business Owners Resources, Marietta, GA
- Barnes L B 1988 Incongruent hierarchies: daughters and younger sons as company CEOs. *Family Business Review* 1: 9–21
- Beckhard R, Dyer Jr. W G 1983 Managing continuity in the family-owned business. *Organizational Dynamics* 12(1): 5–12
- Berle A, Means G C 1932 *The Modern Corporation and Private Property*. Macmillan, New York
- Brockhaus R H 1994 Entrepreneurship and family business research: Comparisons, critique and lessons. *Entrepreneurship Theory and Practice* 19(1): 25–38
- Corbetta G 1995 Patterns of development of family businesses in Italy. *Family Business Review* 4: 255–65
- Danco L A 1975 *Beyond Survival: A Business Owner's Guide for Success*. University Press, Cleveland, OH
- Davis J A, Tagiuri R 1989 The influence of life stage on father-son work relationships in family companies. *Family Business Review*, Spring
- Donnelley R G 1964 The family business. *Harvard Business Review* 42(3): 93–105
- Dyer Jr. W G, Sánchez M 1998 Current state of family business theory and practice as reflected in family business review 1988–1997. *Family Business Review* 4: 287–95
- Gallo M A 1995 *Empresa Familiar*. Textos y Casos, Editorial Barcelona
- Gersick K E, Davis J A, McCollom M, Lansberg I 1997 *Generation to Generation: Life Cycles of the Family Business*. Harvard Business School Press, Boston
- Hollander B S, Elman N S 1988 Family-owned businesses: an emerging field of inquiry. *Family Business Review* 11: 145–64
- Kaye K 1991 Penetrating the cycle of sustained conflict. *Family Business Review* 1: 21–44
- Lansberg I 1988 The succession conspiracy. *Family Business Review* 2: 119–43
- Martinelli A 1994 Entrepreneurship and management. In: Sunclser N J, Swedlord R (eds.) *The Handbook of Economic Sociology*. Princeton University Press, Princeton, NJ
- Neubauer F, Lank A G 1998 *The Family Business. Its Governance for Sustainability*. Routledge, New York

- Scannell Trent E (ed.) 1999 Family business from the household perspective. *Family Business Review Special Issue* 3: 197–281
- Schein E H 1983 The role of the founder in creating organizational culture. *Organizational Dynamics* 12(1): 13–28
- Ward J L 1988 The special role of strategic planning for family businesses. *Family Business Review* 2: 105–17
- Ward J L 1991 *Creating Effective Boards for Private Enterprises: Meeting the Challenges of Continuity and Competition*. Jossey-Bass, San Francisco

G. Corbetta

Family Health

The very diversified topic of family health is currently receiving increasing attention from different disciplines and different directions within psychology. This is especially evident in the areas of clinical and family psychology, gender-specific research, and health psychology (see *Health Psychology*). In this article, important theories and empirical results from these fields are summarized in a systematic overview which is divided into the following sections: (a) the bi-directional connections between family dimensions and the health of family members (e.g., the mother), (b) systemic concepts like ‘psychosomatic families’ or ‘family mental health,’ (c) family resources in terms of coping capabilities or ‘salutogenetic’ and regenerative variables of family life, and (d) differential aspects of family health, that is, differences between the developmental stages of the family life course, differences between family structures, and main results from gender-specific research.

1. Introduction

The article will first of all locate the topic of ‘family health’ within some main psychological traditions, before a systematization of the topic is proposed and selected empirical results are described briefly.

The field of family health can be seen at the crossroads between the disciplines of family psychology and health psychology. The connections between health and family factors are numerous, varied, and important. Nevertheless, the fields of ‘family health psychology’ and ‘family systems medicine’ are still in the early stages of development, although family psychology and health psychology are established disciplines (Akamatsu 1992).

Despite the fact that the family is the basic social context in which health behaviors are learned, clinical and health promotion theory and research traditionally have focused primarily on individual behavior. In contrast, family psychology has been studying inter-individual relationships or ‘systems’ from its beginn-

ings. Yet, within health psychology, family systems considerations are currently gaining recognition.

As health is a very important value for most individuals and families, it has always been of major interest within the discipline of family psychology, though not called ‘family health.’ Important contributions in research on family and health can be accredited to family stress research. In the beginning, this field of study focused on the familial crises brought about by major life events. At the same time, studies on the psychological processes of individuals were also ‘pathogenetically’ oriented. Antonovsky’s ‘salutogenetic’ approach offered an alternative to the pathogenetical perspective.

Antonovsky (1979) builds on earlier developments of Humanistic Psychology and the theory and practice of prevention, and he focuses on health-promoting factors and processes as well as on the characteristics of extremely healthy people or systems. In contrast to earlier studies on health and on the family, the current disciplines of family and health psychology are paying more and more attention to these salutogenetic views and models (see Olson and Stewart 1991).

The early notions of optimal family functioning were inferred from experiences with families or family members in treatment. As the practice of family therapy became institutionalized, integrative models of family health were developed increasingly. Generally speaking, theories and models of family psychology that can be regarded as unequivocally salutogenetic, either (a) conceptualize a health ideal of families, or (b) consider health processes within the context of individual or familiar resources that contribute to the state of health of the whole family or its individual members.

In the following sections, health is conceptualized as a state of total physical, psychological, and social well-being, and not only the absence of disease and disability (see Antonovsky 1979). As this definition contains a physical, a psychological, and a social component, it meets the requirements of a ‘biopsychosocial’ perspective in health psychology (see Schwarzer 1997).

Psychological definitions of family point out that a family is a special group of persons living in a close personal relationship, in the past, in the present, and in the future. It is an economic unit bound together by emotional ties (see Schneewind 1991).

In the following, the two topics of family and health are combined and a systematic overview of the field of ‘family health’ is offered.

2. Family Health: A Systematic Overview and Selected Empirical Results

The very diversified subject of family health was dealt with by totally different directions in theory and research, such as stress research or feminist research.

Table 1

Systematic overview of essential facets of family health in theory and research

1. *The family and the health of individual family members or dyads*
 - (a) Family stress and the health of individual family members
 - (b) The family as context for socialization
 - (c) Intergenerational theory of personal development within the family system: the family of origin in its relevance for the current health of its members
 - (d) Influences of a chronically or acutely sick individual upon other family members
2. *The family as resource and health-promoting institution*
 - (a) Family coping; family variables as 'stress buffers'
 - (b) The couple and family as direct resource and source of satisfaction
3. *Differential aspects of stress and health within the family*
 - (a) Gender-specific family psychological research regarding health and stress
 - (b) Different family structures, e.g., two-parent vs. divorce- and stepfamilies
 - (c) Health and stress in different phases of the family life course; intergenerational differences in health variables
4. *Systemic view of family health*
 - (a) Family health in the form of 'family competence' or 'family mental health'
 - (b) Family ideals
 - (c) Health-relevant characteristics of dyads within the family or of family structure
 - (d) Systemic interventions

Note: The numbering 1 to 4 is parallel to the order in the Sects. 2.1 to 2.4

Following the systematic overview in Table 1, some of these different aspects of family health will now be described briefly.

2.1 *The Family and the Health of Single Family Members*

Studying the links between family characteristics and the health of individuals, one has to distinguish whether (a) the effects of certain family characteristics or situations on system members are being considered, (b) the effects of certain individuals on the family, or whether (c) interaction processes and the resulting effects are being taken into account.

Effects of familial characteristics or situations are studied intensively by family stress research. Within this field, the family is viewed as a source of stress or a mediating factor. 'Stress' is considered either as 'life events,' family transitions, caregiving, or economic strains or stressors resulting from family interactions. Furthermore, extremely adverse familial influences such as violence and abuse are examined. All these forms of familial stress can affect psychological and physical health. Yet the extent of negative effects is highly dependent on coping processes and on individual and familial resources (see Sect. 2.2).

The study of familial influences on child development leads one into the very wide field of socialization within the family. Generally, the family is regarded as the main matrix of personality development, laying the foundations for mental health in childhood. From a clinical perspective, there are certain familial factors which put a child's devel-

opment at risk. Some of these risk factors are psychological disorders of a parent, parental delinquency, strong marital conflicts, and poor educational level of the parents. These variables have been found to correlate with emotional and behavioral disorders and academic failure during childhood. Moreover, childhood problems such as 'internalizing disorders' (e.g., depression) and 'externalizing disorders' (e.g., conduct problems) that stem in part from poor caregiving practices, are again strong predictors of mental disorders in adulthood.

What distinguishes effective from ineffective caregiving? Scientists have identified two fundamental dimensions of caregiving that are particularly important for children's adjustment. The first is, how much warmth, nurturance, and acceptance (vs. hostility and rejection) caregivers convey to children. The second is how much control, structure, and involvement (vs. permissiveness and detachment) caregivers display towards their children. Schneewind and Ruppert (1998), for example, demonstrate by longitudinal data that the family's atmosphere with the primary dimensions of positive emotionality, activation, and level of organization is a relevant predictor of family members' emotional, personality, and social development as well as the child's locus of control (see Schneewind 1995). Furthermore, studies have found that the family's emotional expressiveness also plays a role in the development of social competence in children. Intergenerational family systems theory of personal development within the family views well-being within the context of a multigenerational model. The quality and dynamics of these significant relationship patterns are conceptualized as having far-

reaching implications, not only for psychological functioning, but also for physical well-being, ability to cope with stress, and willingness to engage in positive health-related behavior. All of these assumptions are verified empirically. For practical reasons, nearly all of the studies rely on retrospective data, and most of them address only two generations. These studies, which consider the relationships between important psychological dimensions of the family of origin and the current health of family members, find, for example, moderate correlations between the degrees of autonomy and coherence, which were experienced by grown-up members of their family of origin, and their current psychological adaptation. Studies on the role of parental separation and divorce for child development have shown evidence of impairments of a child's social and personality development, mainly during the first year after the parent's separation. But the effects measured have been proven to be minimal, and the differences between children of intact families and divorced families almost vanish as time goes by.

The study of effects of individual characteristics on the family has a theoretical tradition in the field of family therapy, despite the transactional thinking of this systems approach. In earlier family therapy theories, a family in which an individual member became schizophrenic or developed a psychosomatic illness was known as a 'schizophrenic' or 'psychosomatic family.' Virginia Satir's experiential model of family therapy stresses the importance of the individual's self-esteem to the family's overall level of health. There are also empirical verifications of these interrelationships between self-esteem and perceived family functioning (e.g., Heaven et al. 1996).

This direction of effects from an individual's state of health on the couple or familial relationship has also been examined in cases of acute or chronic illness of a child, a parent, or a spouse. Using longitudinal data, Booth and Johnson (1994), for example, demonstrated that health impairments of a spouse diminish the quality of the couple's relationship quality, especially on the part of the sick person's partner. Nevertheless, the couple's or family's coping processes are very relevant for the system's adaptation to the situation.

2.2 The Family as a Resource and Health-promoting Institution

Although the family and its patterns of interaction can be an enormous stressor, scientists also know about the positive and regenerative aspects of the family. These 'family resources' are fundamental for the functioning of every society. They include a large variety of aspects, ranging from material resources, over social support, dimensions of recovery, and familial coping competencies, to 'daily uplifts' within the family's interaction.

An essential part of the health-promoting aspects of the family lies in the regenerative and protective effects of a good marriage or couple relationship (e.g., Gove et al. 1983). Even family status alone (single, married, separated, divorced, or widowed) shows significant connections with emotional well-being and life expectancy (e.g., see Cotten 1999, Gove et al. 1983). Despite a higher percentage of overweight persons among the married, living together with a spouse is on the whole health-promoting, especially for men. But these differences (between the groups and between the sexes) are diminishing currently. Furthermore, relationship quality is a moderating variable in the correlation between marital status and health, so that 'good,' satisfying marriages are more health-promoting (see Ross et al. 1990).

The correlation between socioeconomic status of families and mental and physical health appears consistently in the literature. Thus, the economic benefits within 'dual-income' families can be seen as a resource, as well as the emotional benefits which many women experience when they are working. These benefits could 'buffer' the stress of the 'double burden.'

Individual stress research has shown that the way people cope with stressors such as critical 'life events' is even more important than their occurrence (e.g., see Antonovsky 1979). Furthermore, there is evidence that also on the family level the quality of coping (e.g., with acute or chronic illness), is extremely relevant. Is the family able to 'buffer' the potentially negative effects of the stressor, or does the family use dysfunctional strategies which only serve to worsen the situation? Adequate coping is a resource which can lead to the establishment of other resources, which then can be useful in comparable or totally different situations in the future.

For couples or parents, the emotional support of a spouse is evidently very important (see *Social Support and Health*). Process analyses in the daily life of couples show that only in cases where a relatively high level of spouse support was evident does a decline in work-related stress occur during the evenings. Interestingly, allowing the partner to withdraw has proven to be one of the forms of spouse support (see Repetti 1989).

Furthermore, high quantity as well as quality of marital interactions has been connected with less risk-taking behaviors, such as less alcohol or drug use (see Wickrama et al. 1995). Similar findings regarding families demonstrate that the lifestyle of persons who live in families is more solid and healthier.

2.3 Differential Aspects of Family Health

Especially in the case of traditional role sharing, we expect gender differences in the experiences of family life and—in relation to these differences—in health-relevant variables.

Studies on the situation of the so-called 'dual income' families show that it is the women who experience a 'double burden' and difficulties in coordinating work and family life. In addition, for women the experienced inequality is more strongly connected with their couple satisfaction (see Kluwer et al. 1997). While women more often experience family stress in the form of their spouses' behaviors (e.g., not enough support), men more often feel stressed by their partners' criticism. Furthermore, as women tend to be supporting and do the larger portion of the household tasks, they have less time to 'unwind' after a stressful work day than their husbands (see Repetti 1989).

In summary, studies in Western nations have found that, although men show more risk-taking behavior compared to the more preventive behavior of women (Schwarzer 1997), women nevertheless have less favorable scores of physical and psychological health. This fact is to be seen in the context of the women's work and family roles, and the overload that is so often connected with these (e.g., see Barnett and Marshall 1991). This strain from work, parental, and couple domains has negative effects on women's well-being and health, and on the functioning of the family.

The lower health scores of older persons that were found by a large number of cross-sectional and longitudinal studies (e.g., see Schneewind and Ruppert 1998) can also be linked with family variables. Apart from biological factors, the stress of family transitions, such as disruptions of marital relationships through divorce or death, also bear aversive emotional and immunological consequences.

Another differential aspect of family health is the family structure or type. While it has often been thought that family structure (i.e., intact, divorced, stepfamily) was a key determinant of well-being, especially among children and adolescents, recent investigations have concluded that family process variables, such as conflict dynamics, cohesiveness, and support, are more important (e.g., Heaven et al. 1996; see above). Furthermore, cultural, social, and economic factors have widespread influences on the health of families.

Differential examinations of the different phases within the family life course show that the level of family stress is highest during the adolescent stage of family life, while family cohesiveness and satisfaction is lowest (see Aldous and Klein 1988).

2.4 Systemic View of Family Health

Scientists who are dealing with an overall, systemic concept of family health try to meet the demands of the systemic approach. They look at the topic in a more holistic manner and identify the structures and processes of the 'whole,' the family system which is embedded in other contexts (e.g., society). Conceptions that

consider the health of the whole family are, for instance, the psychological constructs of 'family competence' or 'family mental health' (e.g., see Pruchno et al. 1994), as well as all family ideals.

In terms of 'family health' measures, there are some self-report assessment instruments and additional measures which are designed to quantify the quality of familial functioning from the perspective of an external clinical observer. Questionnaires that consider an overall systemic construct, for example, 'family mental health,' usually found their operationalizations on sum scores of the individual family members as well as on discrepancy scores between them. Families with higher discrepancies in their perceptions of family reality normally have a lower score on individual well-being. Obviously, converging realities in the subjective views of the system members are functional, whereas larger divergences are a characteristic of clinical families.

Another concept which is of interest, not only from a systemic, family-centered perspective, but also from the perspective of health psychology, is the one of 'family health cognitions': but until now, the empirical basis regarding the impacts of family cognitions on the use of health services and on personal health actions is weak and needs further attention.

An example for a family ideal is the 'balanced' family type within the 'Circumplex model' (Olson 1989). These families have relatively high scores in cohesion (but without being 'enmeshed'), and relatively high scores in their capability to adapt to new situations (without being 'chaotic'). These characteristics are connected with better communication, more favorable coping, and higher family and personal satisfaction and health.

A further systemic perspective of family health concerns health-relevant characteristics of the system structure and the characteristics of dyads within the family system. Such a link between systemic-structural characteristics and family health is represented by the variables of generational boundaries and parental coalitions. Clear—yet not too rigid—boundaries between the parental subsystem on the one hand and the children's subsystem on the other have proven to be a characteristic of healthy family systems. Similarly, coalitions between the parents, the 'architects of the family,' are more functional than strong transgenerational coalitions (e.g., mother-daughter).

Furthermore, a systemic view of the whole family's health is practiced in all family interventions, that is, in family-oriented prevention or health-promotion programs and in family therapies. The insight that the health of the whole family should be considered when one member has a symptom was the birth of family therapy. Family health-promotion programs help families to sustain or enhance the social, emotional, and physical well-being of the family system or its members (see Schneewind 1991). These different kinds of family interventions have become well institutional-

ized. Nevertheless, researchers should pay more attention to evaluations of these different family interventions.

3. Conclusions

Family life experiences affect deeply the competence, resilience, and well-being of everyone. The family shapes the quality of our lives, but—to a certain extent—we also shape the quality and health of our families, especially when we become parents.

Research is paying closer attention to salutogenetic family-centered variables of coping and health. Empirical evidence supports this trend. Yet, in the field of research a great amount of work still lies ahead in order clearly to establish an overall systemic view of the family, whereas in the domain of family therapy this perspective has been the major characteristic from the very beginning.

As patterns of marriage and family life are currently changing in the Western world, scientists should be aware that changes in the strength of connections between family factors and health variables can occur. In the case of the association between family status and health variables, a decline in correlations has already been observed.

Although the family is the basic context in which health behaviors are learned, further research studies should well consider the interactions between the family and other social contexts, like peer or work relationships. Influences from the side of these other persons and systems could, for instance, be one possible explanation for the fact that until now studies have found only rather weak links between 'family health cognitions' and health-related behaviors.

All in all, knowledge of the frames for a healthy development from and in families is very relevant for different branches of psychology. The application of this knowledge is especially important for family interventions, such as prevention programs, but also for family politics.

See also: Childhood Health; Family and Gender; Family and Psychiatric Illness/Family Therapy; Family, Anthropology of; Family Systems in Europe; Gender Role Stress and Health; Public Health; Social Support and Stress; Stress and Health Research

Bibliography

- Akamatsu T J 1992 Family health psychology: Defining a new subdiscipline. In: Akamatsu T J, Parris Stephens M A, Hobfoll S E, Crowther J H (eds.) *Family Health Psychology*. Hemisphere, Washington, DC, pp. 239–50
- Aldous J, Klein D M 1988 The linkages between family development and family stress. In: Klein D M, Aldous J (eds.) *Social Stress and Family Development*. Guilford Press, New York, pp. 3–19
- Antonovsky A 1979 *Health, Stress and Coping*. Jossey-Bass, San Francisco, CA

- Barnett R C, Marshall N L 1991 The relationship between women's work and family roles and their subjective well-being and psychological distress. In: Frankenhaeuser M, Lundberg U, Chesney M (eds.) *Women, Work and Health. Stress and Opportunities*. Plenum Press, New York, pp. 111–36
- Booth A, Johnson D R 1994 Declining health and marital quality. *Journal of Marriage and the Family* **56**: 218–23
- Cotten S R 1999 Marital status and mental health revisited: Examining the importance of risk factors and resources. *Family Relations* **48**: 225–33
- Gove W R, Hughes M, Style C B 1983 Does marriage have positive effects on the psychological well-being of the individual? *Journal of Health and Social Behavior* **24**: 122–31
- Heaven P, Searight H R, Chastain J, Skitka L J 1996 The relationship between perceived family health and personality functioning among Australian adolescents. *The American Journal of Family Therapy* **24**: 358–66
- Kluwer E S, Heesink J A M, van de Vliert F 1997 The marital dynamics of conflict over the division of labor. *Journal of Marriage and the Family* **59**: 635–53
- Olson D H 1989 Circumplex model and family health. In: Ramsey C N Jr (ed.) *Family Systems in Medicine*. Guilford Press, New York, pp. 75–94
- Olson D H, Stewart K L 1991 Family systems and health behaviors. In: Schroeder H E (ed.) *New Directions in Health Psychology Assessment*. Hemisphere, New York, pp. 27–64
- Pruchno R, Burant C, Peters N D 1994 Family mental health: Marital and parent-child consensus as predictors. *Journal of Marriage and the Family* **56**: 747–58
- Repetti R L 1989 Effects of daily workload on subsequent behavior during marital interaction: The roles of social withdrawal and spouse support. *Journal of Personality and Social Psychology* **57**: 651–9
- Ross C E, Mirowsky J, Goldsteen K 1990 The impact of the family on health: The decade in review. *Journal of Marriage and the Family* **52**: 1059–78
- Satir V 1983 *Conjoint Family Therapy*. Science and Behavior Books, Palo Alto, CA
- Schneewind K A 1995 Impact of family processes on health beliefs. In: Bandura A (ed.) *Self-efficacy in Changing Societies*. Cambridge University Press, New York
- Schneewind K A 1991 *Familienpsychologie (Family psychology)*, 2nd edn. Kohlhammer, Stuttgart, Germany
- Schneewind K A, Ruppert S 1998 *Personality and Family Development: An Intergenerational Longitudinal Comparison*. Lawrence Erlbaum Associates, Mahwah, NJ
- Schwarzer R (ed.) 1997 *Gesundheitspsychologie. Ein Lehrbuch (Health Psychology: A Textbook)* 2nd rev. edn. Hogrefe Verlag für Psychologie, Göttingen, Germany
- Wickrama K A, Conger R D, Lorenz F O 1995 Work, marriage, lifestyle, and changes in men's physical health. *Journal of Behavioral Medicine* **18**: 97–111

J. Weiß

Family Law

1. Introduction

Geertz described law as much a part of the culture of a society as its poetry, painting, or music (Geertz 1983). This profound and illuminating approach to

the patchwork of norms, rules, and institutions by which we regulate our daily lives is nowhere more appropriate than when we consider the regulation of family matters. Here law is at its most social as we move forward to new ways of organizing our households and relationships in a multicultural context. The study and practice of family law has attracted socially minded lawyers and has provided a seedbed for the development of the sociological approach to law now known as sociology of law.

In this short article, I will first lay out the various ways of conceptualizing the role of law in family life as the framework within which the individual, the family, and the state define their mutual obligations, and second examine the implications of these models for the framing and operation of the legal assumptions and instruments which regulate family matters.

As Aubert has pointed out, both law and social science are concerned with social interaction: with rights and expectations, with obligations and sanctions, and with predictability (Aubert 1963). But the same terms are used for different purposes. The lawyer uses a normative framework; she gives directions to her client and expects action to follow. The social scientist, on the other hand, observes and describes, seeking to identify and to explain patterns of behavior. The lawyer interested in family work is often concerned with how people behave in family settings, as well as with the strict applications of legal rules. The social scientist who looks at family law is often policy-aware and interested in the practical implications of the work, as well as in the theoretical analysis. The study and development of family law in many different societies, including Japan, Italy, France, the UK, the USA, Australia, and Poland has been imbued with collaboration between social scientists and lawyers (Murayama 1999, Ronfani 1991, Bastard and Cardia Voneche 1995, Commaille and Singley 1997, Maclean and Eekelaar 1998, Weitzman 1991, Funder 1996, Maclean and Kurczewski 1994). The result is a rich understanding of law in family life.

2. *Families, Politics and the Law*

What do we want from law in the context of family life—as opposed to business life, or economic life, or international relations?

The way we model our answer will depend on our view of the family, of the law, and of the state. Let us take three examples of the interweaving of politics, family life, and the law: an old fashioned liberal democracy, an old democracy emerging from totalitarian rule, and a new democracy emerging into a multicultural society. From these examples, we shall derive our models.

In the predominantly secular UK, discussion of the boundaries between the obligations of individuals to

each other as individual citizens or as members of family groups (however these are defined) usually is couched in economic rather than in moral terms. In times of economic stringency, there is a tendency to retrench public spending and to stress the obligations of families to support their members, and also to extend the ties which carry a duty to support. For example, the obligations of parents who do not share a household with their children will be highlighted and enforced. Such moves are countered by the claims of weaker family members, usually women and children but also the older generation, to establish their right to call on the public purse as individuals, irrespective of their family relationships. In the UK, family forms have changed and diversified as a result, in part, of the increased opportunities to form separate households when housing became more widely available in the late 1960s and early 1970s. We have experienced a divorce explosion and a rise in cohabitation. More older people are forming separate households, more women are working despite limited child care, and more men are unemployed.

'Supporting families' has become a political slogan attracting universal support in the UK, but for a variety of reasons. Family life is seen as an answer to the need for increased social control, particularly of young delinquents, and also as the alternative to public provision of shelter and support. The conservative administration of the Thatcher era was active in passing legislation to increase the visibility and enforceability of private obligations as an adjunct to the rolling back of the welfare state. Thus, the UK saw the paradox of a government committed to 'privatizing' the family and to withdrawing from intervention in family matters but compelled to construct a regulatory framework in order to do so. The Child Support Act of 1989 stands at the pinnacle of this policy. This Act attempted to reverse the widely held assumption that when a couple separate and the breadwinner re-partners and produces a second family, their resources should be mainly available to the household in which they now live and not be directed back to the first set of children. The Act firmly stated that first children must have first call on their absent parent's resources. To achieve this change, the government, which was pledged to reducing the role of the 'nanny state,' was required to take draconian powers of investigation and enforcement.

On the other hand, in Poland between 1945 and 1990, the family had a unique role as the only institution supported both by the ruling Communist Party and by the Catholic Church, for different reasons but with a similar purpose. The family was seen as the institution through which moral values are best transmitted and reinforced, and as a powerful agent of social control. In times of intellectual and spiritual alienation, the family became of primary importance as the place where trust predominated over suspicion and where communication could be frank and open.

When there were shortages, the family functioned as an effective searching, queuing, and allocating organization, providing specialist's support for the group in bringing home goods in short supply, or in helping with access to jobs and housing. Following this period of intense interdependency, it has been suggested that family members may wish to withdraw from this level of closeness and move towards the pure relationships of coupledness, parenthood, and friendship defined by Giddens (Reith Lectures 1999). But if the public welfare provision of care for children and care for the sick and elderly break down, then the family may be needed to replace these functions. Despite the emphasis on individual rights and liberties in the free market after the end of authoritarian rule in Poland, family values remain strong and are reinforced by the Catholic Church, whose wish to restrict divorce and abortion has had a profound impact on family law since 1990. Divorce has been removed to the higher level of the courts, which means a long wait and increased costs to the petitioner. And at the same time, the long struggle in the *Sejm* to restrict the right to abortion succeeded in January 1993 by a narrow margin.

For our third example, let us look to South Africa. Those who prepared the new Rainbow Constitution, which signals the passing of the apartheid regime, have struggled with the question of whether to constitutionalize the rights of the individual to marry the person of his choice and to found a family. The Constitutional court acknowledges that a number of international instruments clearly establish the right to form and live in a family. But although many constitutions expressly defend the family (e.g., Germany, Pakistan, China), many do not (e.g., Austria, India, Singapore, Taiwan). The South African government decided not to constitutionalize family matters, stating that the very nature of the family, particularly in multicultural and multi-faith societies, is so diverse that it is best to leave the fundamental rights and liberties in relation to family life to the basic principles of freedom, security, and choice. 'Then through legislation, the development of precedent in courts, social custom and practice, the different forms of family life will manifest themselves and be appropriately protected. The minute you constitutionalize the family, the courts are obliged to establish a prototype of what is meant by a family, and families take such diverse forms in South Africa that this could impose a straight jacket on future developments.' (Eekelaar and Nhlapo 1999).

3. Modeling Family Law

We have looked at three very different examples of the interrelationship between the political situation, the institution of the family, and family law. Now we will try to draw on these variations to develop a set of models which incorporate the key elements and

illuminate our own choices. We can categorize the family law system according to the level of intervention accorded to the state in family life.

3.1 The Interventionist Model

In societies where the state has a strong and coherent message to communicate, based on a political or religious ideology or world view, we tend to find an interventionist model of family law. The function of the family may be to raise good socialist citizens, as in the former communist countries of Eastern Europe, or to raise good catholic citizens, as in Southern Europe. In the former, divorce and abortion were freely available, gender equality was well developed, citizens were equal in the eyes of the state, and labor productivity required women to be available for work. In the latter, divorce and abortion were unavailable. If the family is seen as a box with walls of varying degrees of porosity, in interventionist societies the box is very transparent and porous, allowing the regulatory system to enter. In this model, the family is less able to function as an institution of civil society but, instead, serves the dominant ideology, whether secular or sacred. Such a system has difficulty providing for a multicultural society and traditionally obscures ethnic differences. Such systems ironed out the differences in family life between the Roma, Turks, Pomacks, and Bulgarians in Bulgaria (Maclean and Kurczewski 1997).

3.2 The Residualist or Minimalist Model

The residualist model lies at the other extreme of our continuum. In this model, the family is seen as the desirable or competent source of nurture, caring, social control, and perhaps moral education. In such a model, the walls of the family box are dense to opaque, and the state does not easily penetrate the private family world. The function of the law is to do no more than protect the weaker members of the family group from violence or neglect. This may be left to the criminal law rather than requiring a special form of intervention through family law. Such a model is not acceptable to feminists, as nonintervention in effect favors the status quo and does not redress the power imbalance characteristic of patriarchal societies (see *Gender and the Law*).

3.3 The Facilitative Model

There is, however, a middle position which may be termed the Facilitative model of family law. Under this approach, family law is required to respect the human rights outlined above and (Eekelaar and Nhlapo 1999) to be based on the underlying principles

of freedom, security, and choice, to enable each individual to develop to the fullest and to accommodate the varieties of faith, ethnicity, and sexual orientation. Such a model would encompass acceptance of cohabitation, same sex marriage, assisted reproduction, abortion, divorce, and remarriage, and maybe polygamy and polygyny.

4. *The Way Ahead*

Whichever model of family law underpins the regulatory system of a particular country or state, there do seem to be some common directions for change. In the 1980s and 1990s, writers on family law identified what became known as the rolling back of family law, that is, the increased reliance on individual decision-making and the removal of a number of matters related to personal relationships from legal intervention. It became less common for legal rules to regulate choice of marriage partner (i.e., the old restrictions from the Judaeo-Christian tradition on marrying the dead wife's sister which conflict with the Moslem requirement to take on responsibility for a dead brother's wife). In many societies, no-fault divorce became a matter of personal choice. But although the couple relationship is coming to be seen more as a matter of choice, and to lie outside the boundaries of legal regulation, the relationship between parents and children becomes more exposed and subject to increasing legal intervention. Where children live in households with parents, the need for protective legislation remains, but the need for regulation of how the household sustains itself is minimal (see *Children and the Law*).

But as couple relationships become divorced from the parenting function, the need for legal regulation of the obligations of parents increases. If two people produce a child, and either never live together or separate, and either or both go on to have other children with other partners, the potentially conflicting obligations between this network of parents and children requires a regulatory structure. We referred earlier to the difficulty in the UK when the Child Support Act in 1991 required a change in the distribution of the assets of the absent parent, requiring him to give priority to the children born first, and not to children currently sharing the new household. This was a hard switch to make, and it aroused great public outcry.

But there is a further aspect of the parent-child relationship which may also give rise to legal regulation in the future. With an aging population, and the rolling back of welfare provisions, we may be looking to adult children to offer more support to their elderly relatives (see *Law and Aging*). After the establishment on a firmer footing of child support, we may need to consider legislating for parent support. Such a development manifests a view of family law as no longer

concerned with moral codes of behavior or political ideologies, but as trying through basic human rights principles to deal as best it can with multicultural and multifaith societies. Family law is developing to deal with personal obligations such as material obligations to support and enable.

At the same time, as the thrust of legislative intent focuses on obligations, rather than on rights, so too there is a move away from lawyers and courts to enforce these obligations. We see a resurgence of interest in paralegal activity and in finding alternative methods of dispute resolution. Mediation and conciliation, which require the individuals concerned to take responsibility for arrangements made in family matters, are expected to reduce levels of hostility and to produce longer lasting settlement. Empowerment of the individual is thought to be well served by these means, together with improved ability to communicate, which will be increasingly important as parents continue to take long-term responsibility for their children while no longer sharing a household.

Finally, as family arrangements become more attuned to individual choices in societies with a more tolerant view of different ways of arranging personal life, the ability to negotiate and to adjust becomes more important than the ability to fight for a particular right. If we continue to move in this direction, then it is probable that family law will become less concerned with adjudication or settlement of disputes, and more concerned with the business of managing change. A separation or divorce will involve less acrimony, and there will be clear directives on how to allocate responsibility for children. Legal intervention in families may come to resemble the technical financial advice of the accountant, rather than the fighting spirit of the old divorce lawyer 'taking the husband for all he's got' or 'making sure she doesn't get away with a penny.' The family law of the past was prescriptive, derived from the allocation of property between marrying groups. The family law of the future will be grounded in human rights and will facilitate the maximization of individual choices about personal living arrangements. It will ensure protection for the vulnerable family member, who is likely to be defined by age and relationship to the labor market, rather than by gender.

See also: Children and the Law; Family, Anthropology of; Family as Institution; Family Bargaining; Family Processes; Gender and the Law; Law and Aging; Law and Society: Sociolegal Studies; Motherhood: Economic Aspects; Property: Legal Aspects of Intergenerational Transmission

Bibliography

Aubert V (ed.) 1963 *Sociology of Law*. Penguin, Harmondsworth, UK

- Bastard B, Cardia Voneche L 1995 Professional intervention in divorce. *International Journal of Law and the Family* 9: 275–85
- Commaille J, Singley F (eds.) 1997 *The European Family*. Kluwer, Dordrecht
- Eekelaar J, Nhlapo T (eds.) 1999 *The Changing Family*. Hart, Oxford, UK
- Funder K 1996 *Remaking Families*. AIFS, Melbourne, Australia
- Maclean M, Eekelaar J 1998 *Parental Obligation*. Hart, Oxford, UK
- Maclean M, Kurczewski J (eds.) 1994 *Families, Politics and the Law*. Oxford University Press, Oxford
- Maclean M, Kurczewski J (eds.) 1997 *Family Law and Family Policy in the New Europe*. Hart, Oxford, UK
- Murayama M 1999 Does a lawyer make a difference. *International Journal of Law Policy and the Family* 13: 52–77
- Naumova S 1997 Ethnicity and Family. In: Maclean M, Kurczewski J (eds.) *Family Law and Family Policy in the New Europe*. Hart, Oxford, UK
- Ronfani P 1991 Towards a Dejuridification and a Dejurisdictionalisation of the Family in Family Law and Social Policy. Onati Proceedings (13)
- Weitzman L J, Maclean M 1991 *The Economic Consequences of Divorce*. Oxford University Press, Oxford

M. Maclean

Family Planning Programs: Development and Outcomes

1. Demographic Momentum

One of the most important developments of the twentieth century was a huge increase in global population and a massive change in the distribution and age structure of the population. In the second half of the century, differences in birth rates were an important factor dividing the world into developed and developing nations. For a long time there has been compelling evidence that many couples all over the world want fewer children. In most Developing nations with high birth rates, national family planning programs have been established. Some of these programs have been highly successful, but others failed due to controversy, lack of political will, or inappropriate technical advice.

In the twenty-first century the world will divide along a new socio-economic fault line, also largely driven by differences in birth rates. Countries that achieve replacement level fertility by about 2010 are likely to progress economically. Those that do not will face continued rapid, and in some countries unsustainable, population growth. In much of Africa, for example, half the present population is below the mean age of marriage, and even if the fresh cohorts of young people are able to have smaller families, there will be so many new parents that the population will continue to grow. This phenomenon of population, or demographic, momentum occurred in China in the

1960s and 1970s. As a result of rapid population growth, many more Chinese brides were entering the fertile years than were leaving them by death or reaching the menopause. In 1979 China instituted a one-child policy, but even today the country still has one million more births than deaths each month—and therefore a net growth to China's population of one million each month. Since the one-child policy began, China has added more people than live in the whole of the USA.

In order to overcome past failures in family planning, countries that are being overtaken by a new wave of demographic momentum will need to adopt bold new policies. This will be difficult.

2. Why Do People Have Children?

Unlike most other animals, human beings have frequent sex; most of it unassociated with ovulation and often taking place thousands of times more frequently than is required for fertilization. This behavior probably evolved as a social bond linking parents in the arduous task of bringing up a family. In preliterate societies, pregnancies were naturally spaced by hidden physiological mechanisms that suppress ovulation during breastfeeding. In such communities, women averaged 4–6 live-born children in a lifetime, of which approximately half died before they could reproduce. In the modern world, family size has increased because (a) women often only breastfeed for a few months or not at all, (b) puberty occurs several years earlier than it used to, and (c) infant mortality has plummeted. These are the factors that drove the twentieth century's 'population explosion,' and these factors will continue to operate in the twenty-first century.

The simple fact that human beings have sexual intercourse much more frequently than most animals is obvious, yet often overlooked by governments when they are setting family planning policies, or specialists analyzing demographic changes. Many policy-makers continue to focus on factors that correlate with family size, such as education or income, instead of those that cause changes in family size, which, in addition to the age when intercourse begins and patterns of breastfeeding, are the use of contraception and abortion.

Economists often speak of people making rational choices over the costs and benefits of a particular act, such as buying a washing machine. However, as a result of frequent sexual relations between couples, human fertility is in effect perpetually turned on, and such rational choices are not possible. We buy a washing machine when we want one, but if this worked like human fertility, the washing machine would arrive whether it was ordered or not. In the modern world it is possible to have a small family only if the couple have access to contraceptives and abortion. Tragically, as the twenty-first century begins, hundreds of millions of people still find contraceptives geographically, socially, or financially inaccessible, and for even more

women, abortion is available only in dangerous and commonly financially or sexually exploitative circumstances.

3. Changing Attitudes

In 1900 the world was deeply hostile to the public support of fertility regulation. In the US and parts of Europe, the sale of contraceptives was illegal, while in the UK sales were 'under the counter.' Throughout Europe, North and South America, and in all the European colonies, abortion was illegal. Yet doctors and clergymen, who often opposed contraception in public, were beginning to limit the number of children they had. Condoms, spermicides, intrauterine devices (IUDs), and voluntary sterilization were all known in 1900, but were of poor quality and were difficult to acquire. The theory behind the Pill was described in the 1920s, but research in contraception was actively discouraged. When Margaret Sanger opened a family planning clinic in New York in 1916, it was closed by the police. When Marie Stopes started family planning clinics in England in 1921, she had to face stiff criticism and even arson.

The pivotal point in the development of family planning programs came near the middle of the twentieth century. The 1950s saw the introduction of oral contraceptives, the more widespread use of IUDs, and some Scandinavian and eastern European countries introduced liberal abortion laws. Family planning clinics run by non-governmental agencies began to be important in a few developed nations. In 1965, the US Supreme Court struck down the last of the anti-contraceptive laws passed almost 100 years earlier. In 1966 the British reformed their abortion law, and in 1973 the USA Supreme Court (*Roe v. Wade*) made safe abortion legal throughout the country.

In developing nations, non-governmental organizations also took the lead. In 1952 the International Planned Parenthood Foundation (IPPF) was founded to tie these efforts together. In the 1960s, US foundations and the Swedish International Development Agency (SIDA) began to offer money and know-how to help family planning programs in developing nations. The US Agency for International Development (USAID) followed soon afterwards, and finally a number of developed nations joined in offering support.

But there were also problems. Much of the academic community clung too long to outdated paradigms and until 1959 the US National Institutes of Health (NIH) was forbidden to conduct contraceptive research. The Vatican State, which is part of the UN system, successfully blocked the World Health organization (WHO) from offering governments help in family planning until the 1960s. The British representative at the WHO Assembly was ordered by the Foreign Office to not discuss family planning, but eventually India,

Trinidad and Tobago, and the Netherlands (with British support) did succeed in getting a resolution permitting WHO assistance to respond to requests for family planning technical assistance.

In 1968, many Catholics were stunned when Pope Paul published the Encyclical *Humanae vitae* and upheld the 1600-year-old tradition of St Augustine that sexual intercourse was sinful unless it was open to the possibility of procreation. (In a clear demonstration of the fact that theological opinions on reproduction are not necessarily constant, *Humanae vitae* permitted the use of periodic abstinence, or the safe period, which St Augustine in the fourth century had explicitly condemned as immoral.) Today, almost every developed nation, with the exception of the Republic of Ireland and Malta, offers its citizens access to modern family planning. Pills, condoms, safe and effective IUDs, increasing access to voluntary sterilization for either sex, and above all, access to safe abortion has revolutionized the lives of women in developed nations. For the first time in history, they could fit childbearing into a variety of chosen lifestyles and no longer had to spend every month worrying whether they were going to see their next menstrual period. Nearly every developing country has also established some sort of national family planning program. However, the hostility and controversy that had characterized the slow and painful process of family planning in the first half of the twentieth century has not gone away, and many family planning programs in developing nations are deeply flawed and fail to offer the choices required to truly control family size.

4. Family Planning Programs

The development of international family planning movement can be divided conveniently into three phases, each one symbolized by an international population conference organized by the UN.

In the 1950s and 1960s, a few timid, and sometimes frankly unrealistic, efforts at family planning began. In 1951 the WHO responded to a request by the Indian government by advising use of periodic abstinence. (At that time, the Vatican City, as a member of the WHO governing body vetoed more practical methods.) India adopted a formal governmental population family planning program in 1952, but efforts were made to use IUDs on a large scale failed because the health personnel or facilities to support the method were not available. Very different programs began in South Korea (1963) and Taiwan (1959). Here, several methods including voluntary sterilization were offered and safe abortion, although initially illegal, was widely available. New choices were added to the menu of contraceptives as couples showed they wanted them.

These initial efforts were largely brought about by a small number of charismatic, hard-working individuals. They probably had a powerful impact because

they were responding to a deep and long-hidden desire for fertility regulation in groups all around the world. In the US William Draper Jr. led the Population Crisis Committee, badgered Congress, and played a key role in putting the UNFPA together. Reimert Ravenholt became an extraordinarily effective leader of the early USAID programs. Alan Guttmacher led Planned Parenthood/World Population in the US and changed the face of domestic and international family planning. Ottesen Jensen in Sweden and Madam Kato in Japan moved mountains of opposition. Fernando Tamayo, Jae Mo Yang, Dapne Chung, Lady Rama Rao, L. P. Chow and others appreciated the damage rapid population growth was doing to their societies and economies, and launched family planning programs in their respective developing nations. Those most closely involved in putting programs together, whether from developed or developing nations, were universally impressed by the unmet demand for family planning and the opportunity that family planning provided to meet individual, as well as community needs. Many of these early pioneers were also particularly skilled at turning the controversy, which associated their efforts to their own advantage.

The United Nations Fund for Population Activities (UNFPA) was launched in 1969, and the first UN population conference was held in Bucharest, Romania in 1974. Demographers and economists warned against the dangers of rapid population growth. The Bucharest meeting was dominated by the developed nations who emphasized the need for family planning programs. But a number of developing nations, including India, expressed resentment at what they perceived to be the demographic emphasis of the meeting.

However, when it comes to sex and reproduction, people do not always say what they mean, or mean what they say. Although India and China adopted the rhetoric that 'development is the best contraceptive,' almost as soon as the meeting ended China introduced the one-child family and India embarked on a top-down family planning campaign that ended with episodes of coercive sterilization, which brought down the government of Indira Gandhi. The problem in China was that Chairman Mao had denied there was such a thing as a population problem, and this delayed the introduction of a realistic family planning program. As noted, China ended up with an enormous imbalance of young people. No one in China liked the one-child policy but neither has anyone ever come up with a plausible alternative strategy. Had China started a voluntary family planning program just a few years earlier, there is every reason to believe it would have been as successful as Korea, which saw the TFR fall from 6.0 in 1960 to 1.7 in 1990. Indeed, during the 1970s a number of countries, such as Colombia and Thailand, followed the example of Taiwan and South Korea. They forged ahead offering several methods of contraception, they developed non-medical channels

of distribution and they confronted the public health issues of unsafe abortion. Whereas it had taken the USA almost 60 years (from 1842 to 1900) for the total fertility rate (TFR) to fall from 6.0 to 3.5, Taiwan, Columbia, and Thailand all made the same transition in a few years. In 1984 the world community met for a second population conference, this time in Mexico City. Whereas the US had been criticized in 1974 for emphasizing population problems, in 1984 the Reagan administration asserted that population growth was not a serious problem and followed this up with policies in the US. It was now the turn of the developing nations to call for more support of family planning programs.

At the Cairo International Conference on Population and Development in 1994, yet another set of political forces surfaced. Population issues had fractured into five different schools of thought. Women's advocates focused on the needs of the individual, emphasized broader aspects of reproductive health as well as access to family planning, and downplayed demographic forces, often portraying family planning programs as coercive. Those emphasizing socio-economic development focused on the inequalities between North and South and the need to redistribute the world's wealth more equally. Voices from the South harked back to the Bucharest conference, and again used the slogan that 'development is the best contraceptive' to diminish any role for purely population issues. The Vatican continued its hard-line stand against any form of 'artificial contraception' and condemned the use of condoms, even by monogamous couples where a man might have acquired HIV through a blood transfusion. The group who had launched family planning in the 1960s were reduced in numbers, but they maintained a traditional concern about rapid population growth and emphasized that if the unmet need for family planning was met in a respectful way then fertility would fall rapidly. Most groups at Cairo (other than the Vatican) considered the meeting a considerable success because consensus appeared to have been achieved on so broad a range of issues.

The program of action that came out of Cairo underscored the heavy burden of social and cultural injustice that falls on many women in the world. The conference looked at sexually transmitted diseases and AIDS, it condemned female genital mutilation, and it set family planning in a broad context of reproductive health care. With the passage of time, however, the Cairo Program appears more blurred and less satisfactory. One of the weakest parts of the Cairo document were the targets set for finance, which were much too low. A broad agenda not backed up by the necessary budget is a dangerous situation. It is a sad fact that since the Cairo conference, the sum total of human suffering in the reproductive health field has significantly increased. The WHO has raised the estimates of women dying from pregnancy, childbirth,

and abortion from 500,000 to 575,000 a year—more than one a minute. In the countries most heavily affected by HIV/AIDS, reproductive health has deteriorated more in the past five years than in any similar period during the twentieth century. (See *HIV and Fertility; Mortality and the HIV/AIDS Epidemic*.)

5. *Meeting Individual Family Planning Needs*

A great deal is known about how family planning choices can be made available in a poor country. Experience suggests that a 15 percent rise in contraceptive prevalence is associated with approximately one birth less per family. The wider the choice of contraceptive methods and the greater the variety of distribution channels the more rapidly contraceptive prevalence rises. In many countries prevalence has been going up at more than 2 percent per annum. Family planning is a choice and not a therapy, and therefore contraceptives can be distributed responsibly within a community, often outside any clinic-based health system. Unfortunately, the expansion of family planning also continues to be limited by restrictive medical policies. For example, injectable contraceptives could perfectly well be distributed at a subsidized cost by traditional practitioners, injection doctors, and others who provide medicines in rural and shantytown communities. The subsidized sale of contraceptives has proved an effective way of distributing pills and condoms and is almost always more cost-effective than distribution through medical clinics.

Family planning methods vary in their cost-effectiveness. Condoms are relatively expensive, while voluntary sterilization and intrauterine devices often cost the least per couple-year of protection. Probably the most realistic and acceptable way of satisfying the community's need for family planning would be as follows. The government health services would focus on IUDs and voluntary sterilization. Pills, condoms, and injectables would be available in government clinics, but the primary mode of distribution would be through the commercial sector or through subsidized marketing, involving a variety of outlets, from kiosks on the pavement, through chemical sellers and pharmacies, to traditional practitioners and finally trained physicians. Private physicians could offer sterilization and IUDs. The consumers would pay what they can afford and the government would make up the shortfall. Such a scheme has worked exceedingly well in South Korea and Taiwan and needs to be repeated in countries like India or Nigeria.

6. *Abortion*

The Cairo conference was the first UN meeting to mention the health consequences of unsafe abortion. It concluded that 'prevention of unwanted pregnancies

must always be given the highest priority and every attempt should be made to eliminate the need for abortion.' These are goals that most reasonable people support, but it will take a long time to achieve them given the current low levels of funding.

No country has achieved a low birth rate without a considerable number of abortions—whether legal or illegal. Conversely, wherever women have access to safe abortion, fertility has declined to a lower level than where such a choice is not available. Indeed, nearly all countries that have access to safe abortion have reached or passed below replacement-level fertility. Technically, abortion in the first 3 months of pregnancy can be safely preformed in any clean room. Manual vacuum aspiration (MVA) using hand-held plastic equipment costing less than \$20 has proved an appropriate technology in locations from rural Vietnam to urban San Diego, California. The technique can be used to treat patients with botched abortions as well as to terminate an early pregnancy. In Bangladesh, MVA is available for menstrual regulation, approved by the government as a family planning method through approximately 12,000 providers, and it is one reason why the TFR has fallen from almost 7 in 1960 to 3.3 in 1998.

Abortion is an ethical problem, and all modern democratic societies that separate church and state have to find ways to accommodate a variety of sincerely held views on this critically important topic. Some groups may have to learn the hard lesson that toleration is preferable to confrontation.

7. *HIV/AIDS*

By the early decades of the twenty-first century, AIDS will have killed 30 to 40 million people—more than died as civilians and combatants in World War II. Current UN population projections assume that the global population in the twenty-first century will be 400 million less than it would otherwise have been as a result of the spread of HIV/AIDS. This does not imply that 400 million people will die from this disease, but that the approximately 30 million projected to die early in the next millennium will not themselves have children, and the downstream effects of these deaths will be considerable. In most of the worst hit countries of Africa AIDS will slow population growth but it will not halt it.

8. *The Future*

The developing world is approximately halfway through the demographic transition that began in the 1960s. Family planning programs have played a significant, (and some would argue, a major) role in this important change. As the twenty-first century opens, the largest cohort of young women in history is

entering the fertile years. Contraceptive prevalence is rising. Unfortunately, at this time of unprecedented need, attention is being taken away from family planning. There is an increasing demand for subsidized contraceptives to meet the needs of many millions of very poor people, but there is no clear strategy in the donor community to provide the essential commodities required to fulfill this need.

In 1998 the UN population division issued new projections for global population. Demographic projections are 'what if' statements. The UN projections assume a continued fall in fertility, and in regions such as Africa, these are based on the assumption that the needed contraceptives and services will be available. This may not happen, in which case there will either be a 4-fold rise in abortions, or else millions of unintended pregnancies, in which case even the high UN projections will be exceeded. Of the small amount transferred from rich nations to the poor in overseas aid, only between 1 and 2 percent goes to international family planning. With the added burden of AIDS and lack of a coordinated global policy to ensure that essential commodities are available, the world faces a genuine crisis in family planning and reproductive health.

9. Summary

A great deal is known about how to provide effective family planning choices. When people are offered several methods of contraception and access to safe abortion, fertility always falls. However, family planning has a controversial history. Currently, poor countries sometimes lack the will to make realistic services available and rich countries are reluctant to put enough money into needed programs. As a result the world looks destined to become increasingly divided between high fertility and low fertility countries.

See also: Economic Development and Women; Family Planning Programs: Feminist Perspectives; Family Planning Programs: Quality of Care; Family Size Preferences; Family Theory: Economics of Childbearing; Family Theory: Role of Changing Values; Fertility Control: Overview; Fertility: Institutional and Political Approaches; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations; HIV and Fertility; Land Rights and Gender; Mortality and the HIV/AIDS Epidemic; Population Dynamics: Momentum of Population Growth; Population, Economic Development, and Poverty; Population Policies, Politics of; Population Policy: International; Poverty and Gender in Developing Nations; Reproductive Rights in Developing Nations; Rural Industrialization in Developing Nations and Gender.

Bibliography

- Campbell M 1998 Schools of thought: An analysis of interest groups influential in international population policy. *Population and Environment* **19**: 487–512
- Cleland J 1994 Different pathways to the demographic transition. In: Graham-Smith F (ed.) *Population: The Complex Reality*. The Royal Society, London, pp. 229–48
- Kulczycki A, Potts M, Rosenfield 1996 Abortion and fertility regulation. *The Lancet* **347**: 1663–8
- Potts M 1997 Sex and the birth rate: Human biology, demographic change, and access to fertility regulation methods. *Population and Development Review* **23**: 1–39
- Potts M, Walsh J 1999 Making Cairo work. *The Lancet* **353**: 315–8
- Potts M 2000 The unmet need for family planning. *Scientific American* (January) **282**: 88–93

M. Potts

Family Planning Programs: Feminist Perspectives

1. Population Paradigms

Since the 1960s and even earlier in some cases, governments and private organizations in many poorer countries have developed and implemented family planning programs. By 1985, 37 governments had established national family planning programs, and an additional 33 countries had significant non-governmental programs with the direct or indirect support of governments (Bongaarts et al. 1990). These efforts were largely driven by fears of rapid population growth, and the reduction of fertility was their highest priority. By the 1990s, however, critics of these programs succeeded in causing, what many have called, a revolution in the rationale for these programs. The human right to autonomy in matters of sexuality and reproduction was proposed by feminists, as the cornerstone for these efforts. As a result of this critique, 180 countries achieved consensus on a new approach to population policies and family planning programs at the United Nations International Conference on Population and Development (ICPD), held in Cairo in 1994.

1.1 Family Planning Programs and the Erstwhile Population Approach

To fully understand the significance of the Cairo accord and the influence feminist thinking had on it, it is necessary to look at the philosophies and practices it replaced. The 1960s, with the advent of new con-

traceptive technologies and an increase in fears of a population explosion, saw a significant increase in support for family programs in both industrialized and Third World nations. In 1967, the United Nations Fund for Population Activities (later renamed the United Nations Population Fund) was established to support population and family planning programs. At the UN population conference in Bucharest in 1974, many policy makers—mainly from Northern nations—viewed unchecked population growth as the root cause of poverty, environmental degradation, and social unrest. Many developing countries resisted this line of thinking, suggesting instead that economic development was a more straightforward solution to societal ills, than family planning. While women's issues gained some attention, they were not seen to be central. Meanwhile, on the ground, family planning programs continued, largely unaffected by the debate at the international forum (Demeny 1985).

At the 1984 Mexico City conference, demographic concerns continued to be the driving force in discussions. However, in a reversal of its previous policy, the powerful US delegation now asserted that population growth had a neutral, not harmful, effect on economic development. Furthermore, anti-abortion politics gained center stage as the US launched an attack on abortion rights (Demeny 1985). Still, family planning programs on the ground continued to be guided by demographic goals and to use mechanistic measures such as reaching target numbers of 'contraceptive acceptors.' These programs' principal strategy was the distribution of contraception, chiefly to married women. Perhaps called population programs more accurately than family planning programs, these efforts usually did not address other health needs, even reproductive health needs.

2. The Feminist Critique of Population and Family Planning

As these programs proliferated throughout the 1970s and 1980s, women around the world began focusing on the negative effects of family planning programs that were based on demographics. While not denying the benefits of family planning, feminists rejected the depiction of women as targets of programs designed to limit births (Pitanguy 1995). The feminist critique concentrated on several aspects of program practices, including their vulnerability to coercion, their low quality, and their sole concentration on fertility regulation to the exclusion of other women's health needs.

The vulnerability to coercion was a major element of the feminist critique. In some countries, health care providers were under pressure to achieve targets or be subject to penalties, and workers pressed people into accepting contraception or sterilization against their

will. With success of family planning programs measured in terms of numbers of acceptors, some agencies and governments devised incentive and disincentive schemes to induce couples to participate in family planning programs. For some time, there was a debate about the fine line separating incentives from coercion. Feminists argued that, while milder than sheer force, the incentive practice is coercive; incentives are aimed primarily at the poor, for whom an economic inducement can be a vital necessity, not a free choice (Hartmann 1995). The Cairo Programme of Action later disavowed incentives.

Furthermore, women—and women were the objects of these campaigns in much greater numbers than men—were given little contraceptive choice. The focus was on sterilization or other modern, provider-controlled, methods of contraception such as the injectable Depo-Provera. The risks of these methods, feminists argued, were largely ignored and even exacerbated as a result of poor screening procedures and the lack of follow-up examinations. Safer methods, such as the condom and diaphragm were neglected, even though barrier methods have the ability to help prevent the spread of sexually transmitted diseases. Indeed, apart from fertility regulation, family planning programs provided little else in the way of services. Furthermore, male contraceptive methods were ignored in the research sphere and male participation was neglected in family planning programs, except for sterilization campaigns in some countries (Dixon-Mueller 1993).

Feminists insisted that family planning programs be comprehensive in nature with the limiting of births being only one item in a wide range of women's reproductive health needs. Other items included skilled care during pregnancy; delivery and postpartum care; prevention of infertility; counseling for sexual dysfunction; safe abortion; prevention and management of sexually transmitted diseases; reproductive tract infections, and reproductive system cancers; and treatment of postmenopausal health problems (HERA n.d.).

Furthermore, feminists demanded that women's health advocates should be represented at all levels of policy and decision making. They also maintained that men's responsibilities in the reproductive realm should not be ignored (Women's Declaration on Population Policies 1992).

Perhaps most central to the feminist critique is the assertion that women have the ability to make appropriate reproductive choices. In the feminist view, reproductive choices belong at the individual level, not with the health practitioner and especially not with the state. Noncoercive family planning programs aim to prevent unwanted fertility—as defined by couples, not by the government or other outside parties—through the provision of competent services that respect dignity and privacy. The essence of the feminists' perspective is the view that reproductive rights are human rights.

Another fundamental premise of the feminist perspective on family planning programs, sees women's empowerment as a desirable goal in its own right, rather than a means to a demographic end. Feminists also emphasize gender equity, including men's responsibilities in sexual and reproductive matters (García-Moreno and Claro 1994).

Articulating these views, women in many countries around the globe began advocating wholesale transformation of the approach to population concerns and the basis for family planning programs. By the time of the 1994 ICPD, an international women's health movement was organized and had developed the expertise and the professional skills needed for influencing policy makers and shaping the paradigm shift on population and family planning (Klugman 2000). 'The Cairo Programme of Action sums up the agenda of priorities which women from all over the world, through their leadership networks, built gradually during the years of preparation for the conference' (Berquó 1997).

3. *The Cairo Programme of Action and the 'Feminist Imprint'*

The Cairo conference ushered in a new era of population policy and family planning programming. Its resulting Programme of Action reflects a clear 'feminist imprint' (Petchesky 1995). The conference document presents a new perspective, in which high rates of population growth are understood as an interdependent and aggravating factor—rather than the cause—of problems such as poverty and environmental degradation. Even more critically, it places women's wellbeing at the center of population policy and points to the human rights of individuals to determine and plan family size. Unlike documents from previous population conferences, the ICPD Programme of Action includes chapters on 'Reproductive Rights and Reproductive Health,' and 'Gender Equality, Equity and the Empowerment of Women.'

Echoing the feminist critique, the comprehensive nature of reproductive health is elucidated in section 7.2 of the Programme of Action, as follows:

Reproductive health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity, in all matters relating to the reproductive system and to its functions and processes. Reproductive health therefore implies that people are able to have a satisfying and safe sex life and that they have the capability to reproduce and the freedom to decide if, when and how often to do so. Implicit in this last condition are the right of men and women to be informed and to have access to safe, effective, affordable and acceptable methods of family planning of their choice, as well as other methods of their choice for regulation of fertility which are not against the law, and the right of access to

appropriate health-care services that will enable women to go safely through pregnancy and childbirth and provide couples with the best chance of having a healthy infant.

While previous population conferences had always seen women's development as a tool for achieving lowered fertility rates, the Cairo Programme of Action places women's empowerment as a worthy goal in its own right. 'The empowerment and autonomy of women,' the agreement reads, 'and the improvement of their political, social, economic and health status is a highly important end in itself' (*Programme of Action*: 4.1). In another clear echo of the feminist critique, the document goes far in its call for gender equality (*Programme of Action*: 4.1):

The power relations that impede women's attainment of healthy and fulfilling lives operate at many levels of society, from the most personal to the highly public. Achieving change requires policy and programme actions that will improve women's access to secure livelihoods and economic resources, alleviate their extreme responsibilities with regard to housework, remove legal impediments to their participation in public life, and raise social awareness through effective programmes of education and mass communication.

4. *Critiquing the Critique*

The feminist critique has its detractors and the outcome of the Cairo Conference was not universally applauded. Vocal critics were mainly religious fundamentalists, especially the leadership of the Roman Catholic Church (Shannon 1994). A much quieter reaction was registered among some demographers and others whose concerns about high rates of population growth led to their dismay at what they saw as the dismantling of population policies (Dixon-Mueller and Gemain 2000). For example, Malcom Potts maintained that the 'flaw in the Cairo process was to move from a well-tested middle ground of supplying family planning to adopting a much broader agenda' (Potts 1995). Jason Finkle and Alison McIntosh claimed that women's groups 'co-opted' the ICPD. The conference, they maintained, was 'not a step forward for the population movement, which for many years has given primacy to efforts to limit population growth' (Finkle and McIntosh 1996).

Even some supporters of the reproductive health paradigm adopted in Cairo found some faults in its processes and outcomes. In a review of responses 'that tend to temper the initial feeling of euphoria surrounding the triumph of the feminist agenda,' Huda Zurayk writes that while a large number of women from Third World countries were present in the ICPD process, 'the strings were definitely being pulled by the well-organized Western feminist groups' (Zurayk 1997). Others, however, point to the significant con-

tribution of Third World—particularly Brazilian, Indian, Egyptian, Nigerian, Mexican, and South African—women to ICPD discourse.

5. Beyond Cairo: The Feminist Critique Moves Forward

After the Cairo Conference, the feminist critique expanded to include a vision of policies, governing not only family planning programs but also the broader panoply of issues related to reproduction and sexuality. One year after the ICPD, in 1995, the United Nations held another of its major international conferences, this time the focus was on women and the venue was Beijing. Again the ensuing conference document, the Beijing Platform for Action, reflected feminist input. While refraining from using the phrase 'sexual rights,' the agreement nonetheless endorsed the basic tenets of such rights, and placed them within a human rights framework. 'The human rights of women include their right to have control over and decide freely and responsibly on matters related to their sexuality, including sexual and reproductive health, free of coercion, discrimination and violence,' reads paragraph 96 of the *Plan of Action*.

There is little question that feminist analysis has had significant influence on policy worldwide, principally through the actions of women's nongovernmental organizations. Feminist scholars are now pressing for a similar impact on the discipline of demography, where gender issues have remained marginal. While ideological resistance presents a considerable obstacle, 'intellectually, the incorporation of gender issues in demography promises to enhance the science by offering a deeper understanding of social process and the application of multilevel methodologies' (Presser 2000).

Feminists are now focusing on implementation of the ICPD. Even with the advances in international agreements, many observers have pointed out that change on the level of rhetoric is one thing, and change in the everyday lives of women is another. Studies in many countries show varying levels of success in making reality conform to aspirations. Feminists continue to point to reluctant donor nations, recalcitrant cultural conventions, and wayward practitioners as some of the many obstacles to realizing quality family planning programs that are based on the expressed needs and desires of individual women and men.

See also: Economic Development and Women; Family Health; Family Planning Programs: Development and Outcomes; Family Planning Programs: Quality of Care; Fertility Control: Overview; Fertility: Institutional and Political Approaches; Gender and Health Care; Gender and Reproductive Health; Household Production; Land Rights and Gender; Population

Policies, Politics of; Population Policy: International; Poverty and Gender in Developing Nations; Reproductive Rights in Developing Nations; Rural Industrialization in Developing Nations and Gender

Bibliography

- Berquó E 1997 The ICPD programme of action and reproductive health policy development in Brazil. In: *Proceedings of the International Population Conference*. International Union for the Scientific Study of Population, Liège, Belgium
- Bongaarts J, Parker M W, Phillips J F 1990 The demographic impact of family planning programs. *Studies in Family Planning* 21(6): 299–310
- Correa S, Petchesky R 1994 Reproductive and sexual rights: A feminist perspective. In: Sen G, Germain A, Chen L (eds.) *Population Policies Reconsidered: Health, Empowerment and Rights*. Harvard University Press, Boston
- Demeny P 1985 Bucharest, Mexico City, and beyond. *Population and Development Review* 11(1): 99–106
- Dixon-Mueller R 1993 *Population Policy and Women's Rights: Transforming Reproductive Choice*. Praeger, Westport, CT
- Dixon-Mueller R, Germain A 2000 Reproductive health and the demographic imagination. In: Presser H, Sen G (eds.) *Women's Empowerment and Demographic Processes: Moving Beyond Cairo*. Oxford University Press, Oxford, UK
- Finkle J, McIntosh A 1996 Cairo revisited: Some thoughts on the implications of the ICPD. *Health Transition Review* 6: 110–4
- Garcia-Moreno C, Claro A 1994 Challenges from the Women's Health Movement: Women's rights versus population control. In: Sen G, Germain A, Chen L (eds.) *Population Policies Reconsidered: Health, Empowerment, and Rights*. Harvard University Press, Boston
- Hartmann B 1995 *Reproductive Rights and Wrongs: The Global Politics of Population Control*. South End, Boston
- HERA: Health, Empowerment, Rights and Accountability. *Women's Sexual and Reproductive Rights and Health Action Sheets*. HERA, New York
- Jain A, Bruce J 1994 Quality: The key to success. In: Mazur L (ed.) *Beyond the Numbers*. Island, Washington, DC
- Klugman B 2000 Empowering women through the policy process: The making of health policy in South Africa. In: Presser H, Sen G (eds.) *Women's Empowerment and Demographic Processes: Moving Beyond Cairo*. Oxford University Press, Oxford, UK
- Petchesky R P 1995 From population control to reproductive rights: Feminist fault lines. *Reproductive Health Matters* 6: 152–61
- Pitanguy J 1994 Feminist politics and reproductive rights: The case of Brazil. In: Sen G, Snow R (eds.) *Power and Decision: The Social Control of Reproduction*. Harvard University Press, Boston
- Pitanguy J 1995 From Mexico to Beijing: A new paradigm. *Health and Human Rights* 1(4): 454–60
- Potts M 1995 Cairo's skewed consensus. *Population Today* 23(11)
- Presser H 2000 Demography, feminism, and the science-policy nexus. In: Presser H, Sen G (eds.) *Women's Empowerment and Demographic Processes: Moving Beyond Cairo*. Oxford University Press, Oxford, UK
- Shannon D 1994 All roads led to Cairo. *Conscience* 15(4): 3–10
- Women's Declaration on Population Policies 1992 In: Mazur L (ed.) *Beyond the Numbers*. Island, Washington, DC

Zurayk H 1997 Reproductive health in population policy: A review and a look ahead. In: *International Population Conference*. International Union for the Scientific Study of Population (IUSSP), Liège, Belgium

C. Barroso

Family Planning Programs: Quality of Care

The term 'quality of care' was introduced into the family planning lexicon around 1990. Since then, the concept has been widely accepted as an important dimension of family planning programs by managers, donors, health activists, and researchers. However, the actual progress in improving quality of care rendered by these programs has been limited for various reasons: lack of demonstrated large-scale experience in improving quality in different settings; limited number of studies to document the effects of improved quality; misperception of high cost; and lack of indicators to measure and monitor quality on a regular basis. A number of field studies are currently underway which will remedy this situation. This article reviews the concept of quality of care and the progress made in the field of family planning over the past decade.

1. The Concept of 'Quality of Care'

The concept of quality has been difficult to pin down precisely. Bruce (1990) defined quality as an important but neglected 'dimension' that all family planning services have. Whether the quality of a given service is good or bad is a matter of judgment. Bruce identified six elements in the quality of care framework as 'critical' aspects of services from the perspective and experience of clients. These are: choice of methods; information given to users; technical competence of providers; interpersonal relations between providers and clients; follow-up or continuity mechanisms; and appropriate constellation of services. While the use of these elements will assess from client's perspectives, their actual articulation in the framework are not based on any systematic or random surveys of clients' attitudes toward and expectations from services. Furthermore, clients may not be able to assess the adequacy of elements such as the technical competence of providers or what information they should have received at the time of interaction. The main purpose of this framework has been to draw attention to a neglected dimension and to generate interest among managers and donors toward improving the quality of care.

One criticism of Bruce's framework has been that clients themselves may not necessarily define these

elements as critical in their decision to use family planning services. In this sense, quality needs to be distinguished from other concepts, such as availability, accessibility, and cost of services—all of which may also influence an individual's decision concerning the use of services. Sometimes there is a tendency to discuss all these factors together; but that is not useful for understanding and improving any particular dimension of services. While availability simply refers to the physical availability in an area, accessibility to a service delivery point (SDP) is usually defined and measured in terms of the available mode of transportation and the time taken in reaching that SDP. The cost of utilizing services is typically defined and measured by financial cost involved in getting services once the client reaches that SDP. At times, total cost also includes other costs that the client may bear—social, psychological, and the opportunity costs. Quality, on the other hand, reflects what happens to the client once they reach the SDP.

Often the two terms—'quality of services' and 'quality of care'—are used interchangeably. Nevertheless, a distinction must be made between them. The quality of services refers to attributes of the family planning services, e.g., the number and types of contraceptive methods, staff, and equipment available at an SDP. In comparison, quality of care refers to the way clients are treated by the system providing these services. For example, choice can be offered to clients only if the SDP stocks various methods; on the other hand, availability of methods at an SDP does not guarantee that a choice will be offered to a particular client. In this sense, quality of care places additional emphasis on the interpersonal interactions between providers and clients.

1.1 Improving Quality

A distinction must be made between determinants, indicators, and impact of quality. The determinants of quality include factors such as national population policies, resources allocated to services, and other social and cultural factors. The impact of quality, by contrast, can be considered in terms of improvements in clients' knowledge and behavior with respect to contraception, fertility, and health. As mentioned above, quality is a dimension that all programs have. In this sense, there is always room for improvement, irrespective of the current level of quality. There is no real maximum or upper limit. As far as the minimum level of quality is concerned, such standards should be set locally and explicitly through a process of discussion, which should seek to address the following four interrelated questions:

1. What quality of care does a program *intend* to offer?
2. What quality of care is a program *prepared* to offer?

3. What quality of care is *actually offered* to clients?

4. What quality of care is *received* by the clients?

The process of quality improvement can start with any question. The process is illustrated here by considering the element of choice. The first question deals with the explicit statement at the policy level about clients' right to choose among appropriate methods. The statement then has to be translated into practice through asking the second question, i.e., by considering the type and number of methods that are approved by the program in a country and by considering how appropriate the logistic system to deliver these methods to SDPs is. The next step in the process will be to ascertain whether the intended methods are actually available at an SDP and whether the providers are prepared to offer these methods, i.e., whether they have the requisite training and equipment. This type of assessment at the SDP level will inform the program managers about the extent to which the SDPs are prepared to offer a choice.

Next comes the assessment of the interactions between providers and clients, and this can be done from two perspectives, namely (a) whether the provider has actually offered a choice to clients, and (b) whether the client has received the choice. A particular program may, for example, decide to include three methods, improve the logistic system to get these methods and other required equipments to SDPs, and train providers to offer these methods, i.e., the program intends and is prepared to offer three methods. When a client walks into the SDP, the provider may even offer a choice between two methods; but the client's experience may be that she was not offered the third method of her choice even though all the procedures were in place and even though she had no medical contraindication.

Thus, it is important to think of quality at three levels—policy, SDP, and client. Quality at the policy level and quality of services at the SDP level can be considered as determinants of quality of care actually received by clients. What a particular client receives is the most important basis for assessing, monitoring, and improving quality of care. The identification of gaps between intention, preparedness, services offered, and care received through self-assessment and through systematic data collection is important for an eventual improvement of the quality of care received by clients over time.

The management literature suggests that the commitment of top management is critical in improving quality of care. But how one can generate commitment to quality of care among top management of family planning programs remains an unresolved issue. Unlike in the private sector, a commitment to improve quality in public sector programs with no competition cannot be generated by the motivation of maximizing profit or by the fear of going out of business if consumers do not like the product. Nevertheless, a

combination of various approaches can make it happen.

One way to generate commitment to quality of care is through an organized movement of consumers to demand services of good quality. Another way could be for international donors and institutions to lobby for good quality of care from the consumer's-right perspective. A third possibility is for the top management to realize that improvements in quality will help them to achieve their stated goals of providing services in the first place. A fourth way is for the research community to generate empirical evidence to demonstrate the effect of improved quality of clients' knowledge and their reproductive behavior and health.

1.2 Monitoring Quality

There is considerable interest in creating indicators to measure and monitor quality. This interest stems from the observation that what gets measured and monitored also gets improved. Various efforts have been made to create indicators of quality. A breakthrough came with the methodological invention—Situation Analyses—to collect data from SDPs by actually visiting them. The technique usually involves collection of data about services on the day of a team's visit to a SDP through four questionnaires. The inventory questionnaire collects data about equipment and commodities available at the SDP. Information about provider's background, training, experience, attitudes, and knowledge is collected by interviewing providers working at the SDP. Information about quality provided and received is collected by observing client-provider interactions and by interviewing clients after they finish receiving services. These surveys conducted throughout many developing countries, especially in Africa, have provided important and useful information about the service environments. Surveys conducted over time on a representative sample of SDPs are also useful in measuring, monitoring, and improving quality. However, there is no agreed and tested set of indicators of quality. Additional efforts continue to be made.

Regular monitoring of commodities, equipment, and staff is equivalent to monitoring quality of services, which is sometimes included in what is called 'quality assurance.' Quality assurance is, however, not the same as quality of care, and it may be more difficult to measure and monitor the latter regularly. The problem is created by the small number of clients who use the SDP on the day it is visited by the data collection team. In some surveys, many SDPs visited did not have a client, and quite a few had just one or two clients. Biases are also introduced simply by the presence of an 'observer' at the time of interaction. Finally, it is difficult to assess the effects of variations among SDPs and providers on the overall quality of

care indicators developed on the basis of only a few observations. Some of these issues are also being addressed by ongoing research.

2. Impact of Quality

Clients deserve services of good quality, and they deserve to be treated with dignity and respect while accessing these services. From this perspective, there is no need to demonstrate the effect of improving quality on the reproductive behavior and health of clients. However, since family planning programs in developing countries have also been justified and evaluated in terms of their contribution to fertility decline, an assessment of the effect of improved quality on some indicators of fertility becomes critical in the process of improving quality.

Women who are very motivated (or desperately want) to control their fertility would do so even without the public sector services or even if these services were of poor quality. At the other extreme, women who do not want to regulate their fertility would not use services provided even if they were of the best quality and were available next door at no cost. Improved quality is hypothesized to affect the behavior of women in between these two extremes—those who are ambivalent about their fertility intentions, and those who do not go to services or discontinue the use of contraception because they feel that they receive poor quality or bad treatment. Improvements in quality of care thus can be expected to contribute to the reduction of unintended (unplanned and unwanted) pregnancies and not so much to the reduction in intended pregnancies.

Empirical studies to demonstrate the effect of improved quality on women's reproductive behavior are limited. It has been suggested that a focus on decreasing discontinuation of contraceptives is a better strategy than a focus on attracting new users and not taking good care of their needs. There is also some indication that improved quality in terms of providing a choice among methods increases contraceptive use by attracting new users, as well as by improving continuity of use. Moreover, improving the contents of information exchange between clients and providers improves the initial selection of the method and also improves continuity of contraceptive use. Additional studies are underway in Pakistan, the Philippines, Senegal, Zambia, and other countries, and they will shed further light on the effect of improved quality on clients' knowledge and behavior.

3. Future Directions

Quality was a neglected dimension of family planning services until about a decade ago. Clients have the right to, and they deserve, services of good quality. Therefore, program managers need to pay attention to

improving quality of services. Empirical studies to demonstrate the effect of improved quality on clients' knowledge and behavior can only provide additional justification for improving quality; they cannot replace the main rationale of improving quality—clients' right to good quality.

The original framework of quality assessment was developed for contraceptive services provided through the public sector and through its fixed SDPs. Moreover, the framework is primarily applicable to first-time users of these services. This does not mean that services provided through the private sectors are of good quality or are monitored regularly. This also does not mean that repeat users of a SDP receive services of good quality. There are some efforts to extend the framework to other services and also to incorporate the experiences of repeat users. Such efforts need to be continued. Above all, there is a continued need to initiate processes in developing countries to improve quality of care received by clients.

It is possible to improve quality. It is true that improvements in certain elements (e.g., improving choice of contraceptives by adding new methods) will imply additional costs to the program. It is also true that poor quality resulting in the underutilization of services is costly. Moreover, an improvement in certain other elements is unlikely to cost much. This is especially true for items such as treating clients with respect and dignity and providing them with adequate and accurate information.

Providers often go through initial and refresher training programs. It should not be costly to modify the content and philosophy of these training programs so that providers learn how to solicit information from clients about their reproductive intentions, prior contraceptive experience, and to learn about other pertinent social factors which might influence the selection and use of a particular contraceptive method. It should also be possible to modify training programs so that providers learn how to offer clients adequate and accurate information about how to use the method selected and its adverse effects, how to manage those effects, when to come back for a checkup or a new supply, protection against transmission of disease, and the possibility of switching between methods or sources of supply whenever the method or source initially selected does not remain suitable for the client. Such a transformation, however, would require a shift from using family planning programs as a tool to reduce aggregate fertility and population growth to helping or empowering clients to achieve their own stated reproductive goals in a healthful manner.

See also: Family Planning Programs: Development and Outcomes; Family Planning Programs: Feminist Perspectives; Fertility Control: Overview; Fertility: Institutional and Political Approaches; Fertility: Proximate Determinants; Regionalism and Cultural Expression

Bibliography

- Bruce J 1990 Fundamental elements of the quality of care: A simple framework. *Studies in Family Planning* 21(2): 61–91
- Fisher A A, Mensch B, Miller R, Askew I, Jain A K, Ndeti C, Ndhlovu L, Tapsoba P 1992 *Guidelines and Instruments for a Family Planning Situation Analysis Study*. The Population Council, New York
- Jain A K 1989 Fertility reduction and the quality of family planning services. *Studies in Family Planning* 20(1): 1–16
- Jain A K (ed.) 1992 *Managing Quality of Care in Population Programs*. Kumarian Press, West Hartford, CT
- Jain A K 1999 Should eliminating unmet need for contraception continue to be a program priority? *International Family Planning Perspectives* 25: S39 plus suppl. S
- Jain A K, Bruce J, Mensch B 1992 Setting standards of quality in family planning programs. *Studies in Family Planning* 23(6): 392–5
- Mensch B, Arends-Kuenning M, Jain A K 1996 The impact of quality of family planning services on contraceptive use in Peru. *Studies in Family Planning* 27(2): 59–75
- Mensch B S, Arends-Kuenning M, Jain A K, Garate M R 1997 Avoiding unintended pregnancy in Peru: Does the quality of family planning services matter? *International Family Planning Perspectives* 23: 21–27
- Miller A, Ndhlovu L, Gachara M M, Fisher A 1991 The situation analysis study of the family planning program in Kenya. *Studies in Family Planning* 22(3): 131–43

A. K. Jain

Family Processes

Although there is no univocal or consensual definition of the meaning of family processes, it most often refers to the interactions by which families make up and maintain their unity, manage their conflicts, achieve their socioeconomic integration, and model the social and emotional personality of their members (Sussman et al. 1999). These processes have been analyzed from various theoretical perspectives: structurofunctionalism, symbolic interactionism, development framework, system theory, conflictualism, rational choice, ecological analysis, and so forth (Klein and White 1996). This article will briefly review some of the major results and issues raised by these analyses.

1. The Constitution of Couples and Family

Several trends characterize the process of constitution of couples and family in modern societies. First, there is the widespread custom of romantic marriage, where the partners themselves—rather than parents or tradition—choose their spouse and their civil status.

However, various authors have shown that this individualization of the process of mate selection is accompanied by strong homogamy: people often

choose their partner within their own social, cultural, ethnic, or religious background. The consequence of this homogamy—which is not so obvious in the middle of the social pyramid than in its extremes—is that love marriage does not disrupt social divisions, but conforms to and possibly reinforces them, as in the case of arranged marriages (Aldous 1996). It should, however, be noted that initial homogamy often alters during the family life course because of the disparity of social mobility between the spouses.

Former linear and standardized sequences of becoming part of a couple (i.e., courtship, engagement, marriage), punctuated by strong rites of passage, are often replaced by recursive and individualized paths, which are not culturally well-established. This is true for both the life course of one particular couple and the frequent succession of several marital stories: Family careers diversify and jostle together. Far from signifying an obsolescence of marital union, this individualization of the process seems to meet the couple's increasing concern for personal identity and stability within a mobile, polyreferential, anonymous, and segmented society. In the same way, different forms of today's family ties—cohabitation, one-parent families, step-families, and so forth—can be viewed simultaneously as much in terms of their appropriateness to a complex society as of a factor of stress or deficiency for the person (cf. Kaufmann 1993, Singly 1997).

Although it is individualized, the process of constitution of the couple is also shaped by omnipresent scientific and popular models of 'the' good marriage, models which partly replace the role of juridical norms and religious beliefs.

The transition from couple to family has formed the subject of more contradictory theoretical modeling. Economic theories, like Becker's (1981), model this transition as a balance between the important emotional meaning of the child for potential parents from all social backgrounds, and the cost it involves (loss of profit, restriction to autonomy, contribution and support to education, etc.). Therefore, the number of children born in a family, or their 'quality' (i.e., their social endowment), mostly depends on the parent's material resources. Without denying the growth—within secularized societies—of the importance of intentionality, of evaluation in terms of costs and profits, and concern for family planning, sociological theories are reluctant to limit the meaning of the child to its relational dimension. They add to it a statutory motivation for parents (the child as a resource of prestige, as a means to gain power, as a resource against the outside world) and a motivation regarding the child's needs (care, protection, attendance, etc.). According to the spouses' social assets, these different types of motivation can either be in competition with each other (the attraction for the relationship can be balanced by the fear of losing one's self-identity, professional power, and daily activities) or in con-

vergence. The situation of competition occurs more frequently as the cultural level of the wife increases. This often leads to a delay in producing a child, even when strongly desired, or even to this being renounced it completely by the couple (cf. Teachman et al. 1999).

2. Types of Functioning

Once the families are constituted, how do they work? Early analysis on this theme (e.g., Burgess et al. 1949) have contrasted traditional families (institution)—strongly institutionalized, with clear differentiation of roles between spouses, hierarchical organization of the relationships, and with instrumental goals (economic survival, social integration) prevailing within the couple—with modern families (companionship), marked by the predominance of expressive goals, of equality as the norm, where institutionalization and the differentiation of family member's social identities are softened.

But the growing variety of structural types of families has led several researchers to also emphasize the diversity of the types of functioning of contemporary families. Different typologies have been put forward.

First, focusing on primary goals, Farber contrasts the 'child-oriented families'—where emphasis is placed on children's social integration, on the differentiation of domestic tasks, and where constraint of socialization prevails on the couple's personal interests—with the 'home-oriented families,' where the emphasis is placed on comfort and mutual support in the family, and a flexibility of regulations, and to 'parent-oriented families,' more concerned with the parent's socioprofessional integration, who place parental constraints over those of children, and promote a certain lack of differentiation of roles.

Second, and more centered on modes of cohesion, the typologies of Roussel (1989) contrast 'association families' (strong emphasis on the autonomy of the spouses, a little differentiation of roles, essentially individualized goals) with 'enmeshed families' (emphasis on consensus and full sharing in the couple, predominance of emotional goals of the couple, relative flexibility of roles), and 'alliance families' (strong institutionalization, predominance of the family's social integration as a goal, strong differentiation of roles).

Lastly, other authors have focused on types of normative regulations. For Kantor and Lehr, 'open families' are characterized by organization standards (roles, behavior, etc.) that evolve with contextual changes, whereas 'random families' show almost no intersituational coherence in their norms of functioning, and 'closed families' are characterized by a rigidity of their rules and incapacity to adapt to the evolution of the context and to take into account external

information. A similar interpretation can be found by Reiss, who contrasts 'consensus-sensitive families' with 'environment-sensitive families.' Olson's circumplex model, crossing dimensions of cohesion and regulation, seeks to show that extreme types of functioning (i.e., too rigid or too enmeshed) are more at risk than balanced types. These types of functioning vary according to the social status of the spouses: an increase in economic and cultural assets is linked to a more distinct emphasis on individual autonomy, a flexibility of regulation, and wide receptiveness of the group towards relationships and external information (see Kellerhals et al. 1994 for a synthesis).

3. Distribution of Roles, Power, Justice

Within the study of these modes of functioning, the question of role and power distribution has obviously occupied a central position. In the 1950s, the dominant idea was that small groups—that is, the nuclear family—generally engender a differentiation between instrumental roles (contribution of external resources, choice of the group's orientation) and expressive roles (organization of the internal production and the management of relationships), that brings about a higher efficiency of effort and turns out to be essential to the success of the socialization process (Parsons and Bales 1955). However, under the triple influence of feminism, the diversification of lifestyles, and the revisiting of ethnological analysis, the universality of this model has been contested and its functional character challenged by research showing that within a romantic marriage, the sharing of the expressive function and the equality of involvement is a determining factor of relational and educational success (Aldous 1977).

This 'functionalist consensus' was then overtaken by a rational utilitarian perspective—the resource theory (Blood and Wolfe 1960)—that understands the exercise of power in the family as a reflection of the inequality of socioeconomic resources between the spouses. According to this theory, the actor with higher outside resources exercises—despite possible egalitarian norms—a greater internal decision-making power; moreover, the comparative significance of the spouses' resources dictates their respective contribution to educational and domestic tasks. Criticism of this theory has been mostly concerned with the fact that (a) resources, far from being only economic, can also be charismatic, can concern expert knowledge, or be normative; (b) the weight of the latter vary according to the cultural context; and (c) the power of the other depends on the ego's alternatives. This triple determination can explain the permanency of sexism or domestic violence, as much as the unequal distribution of tasks in contemporary families. However, contemporary formulations of the problem emphasize the fact that by increasing the wife's potential in-

dependence and economic power, professional integration of women also increases her motivation and ability to speak for herself and require some consideration, while these elements will in turn reinforce the actor's power in negotiations (cf. Szinovacz 1987, Gelles 1995). The mechanistic exchange model is thus taken over by a sociocognitive formulation, where processes by which identity is constituted play an important part.

This transition from one paradigm to another has had repercussions particularly on the theorization of feelings of justice (fairness) in the family. The extent of inequality suffered by couples on a long-term basis tends to show that the simple equity theory—according to which feelings of injustice (unfairness) appear when the balance between one actor's contributions and rewards is different from another—is not always appropriate here. In fact, criteria of fairness vary according to status (social rank, age, sex), to the primary goals of the group, and to its type of cohesion. Three very different concepts of justice can thus be drawn (Kellerhals et al. 1997). The 'voluntarist' concept considers that justice comes when all concerned parties have been duly referred to and have been able to make an explicit or implicit agreement concerning the distribution of tasks or rewards in the family. Agreement between each other's will creates fairness. The 'finalist' concept defines the occurrence of fairness when the distribution of rights and duties brings positive outcomes to the individual or the group, with no primary consideration for merit or agreement. Finally, the 'comparative' concept considers that fairness occurs when people with the same status are treated in a similar way.

As we can see, these concepts of justice enable individuals to allow or make sense of very strong objective inequalities. They result from a collective construction of meaning, rather than from a static comparison between debts and bonds. This collective construction varies according to the type of cohesion of the group (normative families with a strong sense of unity tend to favor the 'comparative' concept, whereas more associative and negotiator families prefer the 'voluntarist' concept), and to social status ('comparative' concepts are more common at the bottom of social hierarchy, whereas 'voluntarist' ideas prevail at the top). Lastly, the spouses' concepts are not necessarily identical, so much so that the fairness/unfairness debate often seems to refer more to the confrontation of ideals, rather than to the objective disparity between contribution and retribution.

4. *The Educational Relationship*

Every family has to comply with an educational task that usually aims at providing children with social competencies (cognitive skills, social skills, and sense of autonomy/belongingness) essential to their social

integration. However, in doing so, the parents' educational strategies vary greatly. They have often been characterized by their degree of permissiveness or control, on the one hand, and by their degree of warmth, commitment, and support (vs. lack of concern or hostility), on the other. It is on that basis that Baumrind, for instance, distinguishes between an authoritarian style (weak support and nurturance, quite strict and punitive authority, little rational communication), as opposed to a permissive style (high support and weak control, either rational or punitive), to a harmonious style where parents try to have influence on their children by egalitarian negotiation, and an authoritative style (high degree of support and control) which favors rational communication and the child's autonomy. Many researches show that an average degree of control and support are favorable to a good acquisition of social competencies and good self-esteem, whereas too much authority and punishment—or conversely too much protection—lead to blocks or rejections (cf. Rollins and Thomas 1979).

The parent's social status (education, occupation, income) affects educational strategies. Kohn (1977), for instance, shows that self-control, autonomy, and curiosity are given more value at the top of the hierarchy, whereas immediate conformity tends to be emphasized at the bottom. Concerning educational methods, it has been found that coercive methods are slightly more used in poorer backgrounds, with emphasis on motivation increasing along the social scale (Gecas 1979). According to Kohn (1977), these disparities reflect the occupational situation of the parents: The more their occupation necessitates initiatives, autonomy, and symbolization, the more their educational methods will value these elements, and vice versa.

Subsequent studies have improved the rather mechanistic aspect of these analyses by giving more importance to a mutual, joint, and socially constructed character of socialization. First, observation of infants has shown how their reaction can modify the mother's attitude. Moreover, the mutual ability of children and adults to establish synchronies, turn-taking, or behavior state matching proved to be an important factor of the attachment to, and the transmission of, norms (Peterson and Rollins 1987). Lastly, sociocognitive aspects, expectations relating to the 'good child,' and common knowledge relating to interpretations of deviant behavior model the adult's response to the child's actions and doings. This diversity of perspectives leads naturally to viewing the educational process as a site of interinfluence of the family microsystem, the institutional mesosystem, and the society macrosystem (Bronfenbrenner 1986). This 'ecological' perspective has notably shown how much the quality of the relationship between husband and wife affects the mother's relationship with her child (second-order effect). In the same way, it has also been proved that

educational style is an integral part of the family's type of cohesion: 'associative' families are likely to insist on autonomy and creativity, whereas merged families concentrate mostly on mimicry and conformity (Kellerhals et al. 1994). Then again, mechanistic models have been taken over by more symbolic interactionist perspectives.

5. Families and Kinship Network

Because families have so many tasks to accomplish, there is a question as to whether they generally face these problems on their own or not. In this regard, the idea of the modern family withdrawing into itself has not stood up to several studies on the theme of family solidarity which show that the kin group's support in the family management—purchase of property, child care, moral support, and so forth—is often essential, due to its flexibility, versatility, and speed (Pitrou 1992). However, these kinship ties are often characterized by a poor number of effective links, a clear primacy of vertical links (filiation) over horizontal links (sister/brother link), the predominance of the mother's kin, and by a much higher involvement of women. Kinship solidarity is, therefore, much affected by geographical distance, divorce, and professional occupations of both spouses, three points which limit availability.

Moreover, because kinship relationships often rely on affinity rather than on duty or status, and because the number of recomposed families following a divorce is multiplying, the kinship tie looks increasingly like a network (weak bond connections and absence of an important center point) rather than like a group concentrically focused on forebears and endowed with a strong collective consciousness (cf. Coenen-Huther et al. 1994).

6. Adjustment and Divorce

The study of family processes logically leads to research concerning reasons for the success or failure of the family faced with its tasks; in other words, concerning factors of adjustment. But what criteria of adjustment can be used?

Measuring marital and family adjustment in regard to the propensity for divorce creates confusion between the question of relational difficulties and economic or cultural obstacles preventing separation. Situations where marital stability coexist with disappointment often occur. However, it is important to try to define what are the factors of divorce. Before and during the 1960s, marital separation was usually interpreted with deficit models. Poor adjustment has therefore been viewed as a result of deficiency in the socialization to the other: couples getting married and having children very young, and also those whose

parents have failed to pass on stable or harmonious models of marital union, have a higher rate of divorce than others. It has also been shown (Cherlin and Furstenberg 1994) that probability of divorce is higher in second unions, where there is often a lack of clear cultural models to guide relationships between generations.

Poor adjustment has also been related to a deficiency of symbolic or economic resources. The emphasis on happiness, communication, and consumption often found in contemporary family's plans demands financial resources that are not always available. Consequently, it is understandable that couples with low income and poor education separate more often, the distance between objectives and means being so large.

Poor adjustment has finally been viewed as the result of pronounced cultural differences between spouses, which finds expression in a lack of shared codes and values. A higher rate of divorce has therefore been noticed among mixed religion or nationality couples, or among heterogamous couples in age or social level (cf. Faust and McKibben 1999). However, once again, we should be aware that the social meaning of difference is probably even more important than the difference itself: hypergamous couples (where the man occupies a higher position than the woman does), for instance, are far more stable than hypogamous couples (where the situation is reversed), who are not so well accepted socially. But these different deficit models fail to explain the majority of contemporary divorces. More global, but also less verifiable, interpretations have been suggested, like Berger's, for whom too much expectation regarding the couple and the family explains the extent of disillusion. Other analyses (Popenoe, Bellah, Roussel) insist on the incompatibility of individualistic concern for personal genuineness with constraints related to community commitment; others also have placed the accent on the existing hiatus between egalitarian models of marital relationship and very unequal marital practices. In all these approaches, divorce appears to be more like an intrinsic property of contemporary couples, rather than linked to clearly identifiable individual deficiencies (Roussel 1989).

Given the biases of the 'divorce' indicator, there have been attempts—since initial studies by Burgess and Cotrell—to measure marital adjustment with a great many multidimensional scales (implying, e.g., degree of satisfaction, cohesion, consensus, and emotional expression) and to examine factors that affect this marital quality. These are due to the marriage's previous history (homogamy, socialization to appropriate role models), to the spouses' current resources (integration and community support, economic level, educational level), and to the nature of their interaction (quality of communication, appropriateness to roles) (Gelles 1995). One can believe that these different 'resources' make it possible to avoid extreme

forms of cohesion (enmeshment or disengagement) and of normative regulation (rigid or chaotic) which themselves limit the family's ability to cope with stress, to adapt to external changes, and to manage disruptions (cf. Olson et al. 1983, Boss 1988).

See also: Divorce and Gender; Divorce, Sociology of; Domestic Violence: Sociological Perspectives; Families and Households, Behavioral Demography of; Families as Educational Settings; Family and Gender; Family as Institution; Family Bargaining; Fatherhood; Generations, Relations Between; Kinship in Anthropology; Marriage; Marriage and the Dual-career Family: Cultural Concerns; Motherhood: Economic Aspects; Motherhood: Social and Cultural Aspects; Socialization, Sociology of; Status and Role, Social Psychology of

Bibliography

- Aldous J 1977 Family interaction patterns. *Annual Review of Sociology* 3: 105–35
- Aldous J 1996 *Family Careers*. Sage, Thousand Oaks, CA
- Becker G 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA
- Blood R O, Wolfe D H 1960 *Husbands and Wives: The Dynamics of Family Life*. Free Press, New York
- Boss P G 1988 *Family Stress Management*. Sage, Newbury Park, CA
- Bronfenbrenner U 1986 Ecology of the family as a context for human development: Research perspectives. *Development Psychology* 22(6): 723–42
- Burgess E W, Locke H J, Thomes M 1949 *The Family: From Institution to Companionship*. American Book, New York
- Cherlin A J, Furstenberg F F 1994 Stepfamilies in the United States: A reconsideration. *Annual Review of Sociology* 20: 359–81
- Coenen-Huther J, Kellerhals J, von Allmen M 1994 *Les réseaux de solidarité dans la famille*. Réalités Sociales, Lausanne, Switzerland
- Faust K A, McKibben J N 1999 Marital dissolution: Divorce, separation, annulment and widowhood. In: Sussman M B, Steinmetz S K, Peterson G W (eds.) *Handbook of Marriage and the Family*, 2nd edn. Plenum, New York
- Gecas V 1979 The influence of social class on socialization. In: Burr W R et al. (eds.) *Contemporary Theories about the Family*. Free Press, New York, Vol. 1, pp. 365–404
- Gelles R J 1995 *Contemporary Families*. Sage, Thousand Oaks, CA
- Kaufmann J C 1993 *Sociologie du couple*. Presses Universitaires de France, Paris
- Kellerhals J, Modak M, Perrenoud D 1997 *Le sentiment de justice dans les relations sociales*. Presses Universitaires de France, Paris
- Kellerhals J, Troutot P Y, Lazega E 1994 *Microsociologie de la famille*. Presses Universitaires de France, Paris
- Klein D M, White J M 1996 *Family Theories—An Introduction*. Sage, Thousand Oaks, CA
- Kohn M L 1977 *Class and Conformity: A Study in Values*. University of Chicago Press, Chicago

- Olson D H, Hamilton I, McCubbin 1983 *Families: What Make them Work?* Sage, Beverly Hills, CA
- Parsons T, Bales R F 1955 *Family: Socialization and Interaction Process*. Free Press, New York
- Peterson G W, Rollins B C 1987 Parent-child socialization. In: Sussman M B, Steinmetz S K, Peterson G W (eds.) *Handbook of Marriage and the Family*. Plenum, New York, pp. 471–507
- Pitrou A 1992 *Vivre sans famille? Les solidarités familiales dans le monde d'aujourd'hui*. Privat, Toulouse, France
- Rollins B C, Thomas D L 1979 Parental support, power and control technics in the socialization of children. In: Burr W R et al. (eds.) *Contemporary Theories about the Family*. Free Press, New York, Vol. 1, pp. 317–64
- Roussel L 1989 *La famille incertaine*. Odile Jacob, Paris
- Singly F de 1997 *Le soi, le couple et la famille*. Nathan, Paris
- Sussman M B, Steinmetz S K, Peterson G W (eds.) 1999 *Handbook of Marriage and the Family*, 2nd edn. Plenum, New York
- Szinovacz M E 1987 Family power. In: Sussman M B, Steinmetz S K, Peterson G W (eds.) *Handbook of Marriage and the Family*. Plenum, New York
- Teachman J D, Polonko K A, Scanlon J 1999 Demography and families. In: Sussman M B, Steinmetz M K, Peterson G W (eds.) *Handbook of Marriage and the Family*, 2nd edn. Plenum, New York

J. Kellerhals

Family Size Preferences

Preferred or desired family size is the number of children wanted in one's lifetime. Family size preferences are viewed as measures of the demand for children which, in combination with the supply of children and contraception determines the number of children born. Family size preferences may change over the life course, especially in response to child-bearing and childrearing experience.

Family size preferences or desires are conceptually and empirically distinct from family size ideals, norms, intentions, and expectations. Fertility ideals and norms represent what is desirable for a group of people or a typical group member, rather than what is desirable for every person in the group. Family size intentions or expectations reflect not only family size preferences but also constraints on one's ability to achieve desired or preferred goals.

1. Conceptualization and Measurement

1.1 Desired, Intended, Expected, and Ideal Family Size

McClelland assessed the construct desired family size as a measure of the demand for children, that is, 'the

number of children parents would have if there were no subjective or economic problems involved in regulating fertility' (1983, p. 288). He concluded that family size desires do, indeed, represent the theoretical construct, demand for children.

Family size intentions and expectations reflect not only family size preferences but also perceptions of the likelihood that preferences can be translated into outcomes. The gap between intended or expected family size and desired family size may reflect influence of partner's desires, perceived fecundity and fertility control, or other known constraints on fertility behavior or outcomes.

Questions about a general ideal family size produce responses that fit the character of a social norm rather than that of a personal preference. For some individuals—those who have internalized societal ideals or who believe their own situation is typical—ideals may be identical to desires (Mason 1983).

1.2 Measures of Family Size Preferences

Two types of questions are most commonly used to measure family size preferences. The first asks respondents how many children they would want to have in their lifetime, if they could choose the exact number. For childless respondents, this question measures desires prospectively; for parents, the question includes a retrospective assessment of previous births with a prospective assessment of potential future births. Retrospective assessments of birth wantedness are subject to social desirability bias or rationalization and therefore produce upwardly biased estimates of desired family size. Another version of this question attempts to give respondents permission to report previous births as unwanted by asking something like: 'If you could start life over again, how many children would you want to have?' Such questions may also place conditions on the question, specifying that the respondent's life would be in other respects about the same as it had been so far. It is not clear that these two forms of the lifetime question produce different responses.

The second type of question asks respondents how many more children they would like to have; for childless respondents, the answer would presumably be the same as for the first type of question, but for parents, retrospective assessments of previous births are excluded. To obtain an estimate of desired family size, some analysts simply add the respondent's current number of children to the additional number desired. This strategy assumes that all previous births were wanted, producing an even stronger upward bias than does the first type of measure. Several methods have been proposed for estimating the number of unwanted previous births from parity-specific values of additional desired children, and thereby to produce a more accurate measure of desired family size

(Bongaarts 1998). Unfortunately, these methods do not provide estimated family size preferences for individuals.

An alternative measure of family size preferences asks respondents to make choices between pairs of family sizes, producing a scale reflecting desires for large versus small families. Scales have also been developed to incorporate preferences for sex of children in family size preferences and vice-versa (Coombs 1979). The interdependent IN and IS scales reflect the fact that family size desires may incorporate desires for particular numbers of boys or for particular numbers of girls or some balance between the two.

1.3 Non-numerical Responses

In high-fertility societies, small but significant proportions of respondents do not provide a numerical response to questions about desired family size, reporting instead that the number of children they have is 'up to God' or that they will accept whatever number of children they have. These responses raise the question of whether individuals form preferences about things they believe they cannot or should not try to influence. Such respondents do respond to questions about family size ideals, preferences for one family size over another, and unwanted or mistimed previous births (McClelland 1983, Riley et al. 1993).

2. Her, His, and Their Family Size Desires and Intentions

Most data on family size desires or intentions come from interviews with women. The assumptions underlying women-only fertility surveys are that (a) women and men share the same conditions of life and will therefore want and intend the same number of children, (b) because women bear children and are primary childrearsers, their desires and intentions are those that determine fertility, and (c) women's desires and intentions will incorporate those of their partners, reflecting resolution of any initial disagreements. These assumptions appear to be tenable in the aggregate; when data are collected from women and men, their respective family size desires and intentions are more similar than different (Mason and Taj 1987). On the other hand, data collected from couples consistently show that substantial proportions of couples disagree about desires or intentions for further childbearing or completed family size (Becker 1996).

3. Family Size Preferences, Intentions, and Births

The relationship between family size desires and births depends on relationships between preferences and intentions, between intentions and behavior, and

between behavior and outcomes. Couple disagreement is a primary source of gaps between desires and intentions, that is, the gap between individual desires and intentions to have a child is explained in large part by a spouse's different preference (Thomson 1997). Another key source of gaps between desires and intentions is the perception of control over fertility. Those who believe themselves to be infecund may desire more children than they intend; those who do not perceive themselves to be able to control fertility may intend more children than they desire.

Family size preferences or intentions may also change over the life course and are particularly likely to be influenced by experiences with previous births (Namboodiri 1983). In spite of these expected gaps, measures of family size desires, particularly the desire for additional children, have moderately strong effects on fertility behavior and outcomes (McClelland 1983). Lee (1980) has demonstrated that a 'moving target' model of desired and achieved fertility produces even stronger correspondence between family size desires and period fertility.

Davidson and Jaccard (1979) identified many of the key conditions under which a strong effect of fertility desires or intentions on births should be expected. First, the strength of the effect depends on the extent to which the ultimate goal (birth or no birth) is achieved through a sequence of behaviors and events, including sexual intercourse, contraception, conception, and abortion. Fertility desires and intentions should have stronger effects on the behaviors at the beginning of the sequence than on the ultimate outcome. Second, the effect will be stronger over shorter versus longer periods of time because the opportunity for new experiences that change desires or intentions is reduced. Third, effects of desires or intentions on behavior and outcomes should be stronger for individuals who place a high value on cognitive consistency, who believe they have control over their actions, and who generally engage in planful behavior than for those who do not. Finally, desires and intentions and behaviors and outcomes should be measured at the same level of specificity in order to expect a strong relationship; desired or ideal family size should not, for example, be expected to have strong effects on contraceptive use or even the timing of a particular birth.

The relationship between childbearing desires or intentions and births depends to a considerable extent on couple agreement. Couple disagreement leads to birth rates between those for couples who both desire or intend a child and those who do not. This means that both partners' desires have at least some influence on births (e.g., Thomson 1997, Thomson and Hoem 1998, Dodoo 1998).

Estimated relationships between fertility desires and intentions or between intentions and behavior also depend on measurement quality. The IN-scale discussed above provides more accurate measures of

preferences and therefore has larger estimated effects on fertility behavior than does a single most preferred number of children (Coombs 1979). Measures of the desire for children that incorporate dimensions of attitude strength have stronger effects on childbearing intentions and behavior than do yes-no-don't know responses to direct questions about desire for children (e.g., Davidson and Jaccard 1979). Similarly, intentions measured with some degree of certainty or uncertainty are better predictors of outcomes than a simple dichotomy (Thomson 1997).

See also: Children, Value of; Demographic Transition, Second; Family as Institution; Family Processes; Family Systems and the Preferred Sex of Children; Family Theory and the Realities of Childbearing Behavior; Family Theory: Economics of Childbearing; Fertility Change: Quantum and Tempo; Fertility Control: Overview; Fertility of Single and Cohabiting Women; Fertility: Proximate Determinants; Reproductive Rights in Affluent Nations

Bibliography

- Becker S 1996 Couples and reproductive health: A review of couple studies. *Studies in Family Planning* 27: 291–306
- Bongaarts J 1998 Fertility and reproductive preferences in post-transitional societies. Policy Research Division Working Paper No. 114, Population Council, New York
- Coombs L C 1979 Reproductive goals and achieved fertility: A fifteen-year perspective. *Demography* 16: 523–34
- Davidson A R, Jaccard J J 1979 Variables that moderate the attitude-behavior relation: Results of a longitudinal survey. *Journal of Personality and Social Psychology* 37: 1364–76
- Dodoo F N 1998 Men matter: Additive and interactive gendered preferences and reproductive behavior in Kenya. *Demography* 35: 229–42
- Lee R D 1980 Aiming at a moving target: Period fertility and changing reproductive goals. *Population Studies* 34: 205–26
- Mason K O 1983 Fertility norms. In: Bulatao R A, Lee R D (eds.) *Determinants of Fertility in Developing Countries, Vol. 1, Supply and Demand for Children*. Academic Press, New York
- Mason K O, Taj A M 1987 Differences between women's and men's reproductive goals in developing countries. *Population and Development Review* 13: 611–38
- McClelland G H 1983 Family-size desires as measures of demand. In: Bulatao R A, Lee R D (eds.) *Determinants of Fertility in Developing Countries, Volume 1, Supply and Demand for Children*. Academic Press, New York
- Namboodiri N K 1983 Sequential fertility decision making and the life course. In: Bulatao R A, Lee R D (eds.) *Determinants of Fertility in Developing Countries, Volume 1, Supply and Demand for Children*. Academic Press, New York
- Riley A P, Hermalin A I, Rosero-Bixby L 1993 A new look at the determinants of nonnumeric response to desired family size: The case of Costa Rica. *Demography* 30: 159–74
- Thomson E 1997 Couple childbearing intentions and births. *Demography* 34: 343–54

Thomson E, Hoem J M 1998 Couple childbearing plans and births in Sweden. *Demography* 35: 315–22

E. Thomson

Family Systems and the Preferred Sex of Children

1. Introduction

Most human societies exhibit some degree of gender inequality, but there is enormous variation in its nature and manifestations. In some societies, women face only mild or moderate constraints on functioning as socially and economically autonomous persons in the public life of the community; in others, women are largely subordinate to males. This has implications for any old-age support parents can expect from daughters, which, in turn, affects the aspirations of couples regarding the sex of their children. In most regions of the world, parents express a preference for a gender-balanced family—with, perhaps, mild preferences for children of a particular gender. However, in some regions parents exhibit a significant degree of son preference. These preferences manifest themselves in decision making at the margin in many ways—ranging from investing more in educating sons to the extreme outcome of excess mortality among daughters.

To understand why some societies have strong son preference while others do not, we need to examine the gap between sons' and daughters' ability to contribute to the physical, emotional, and financial well-being of their parental household—as this is the unit primarily responsible for making and implementing decisions about childbearing and childrearing. This gap is determined largely by kinship systems, which specify clearly defined roles for different categories of family members. In most societies, it is possible for women to maintain mutually supportive relations with their parental household even after marriage, making for little difference in the value of girls and boys to their parents. However, in other societies women's links with their parental household are largely severed at marriage and sons are enjoined to care for their parents—making it clearly more attractive for parents to rear sons instead of daughters. Gender-appropriate roles are also decided, to a lesser extent, by other norms—such as religious beliefs and injunctions, which can, for example, discourage women from participating in the labor market and thus reduce their ability to help their parents financially.

In this entry we review the broad features of family systems in different cultural regions of the developing world to describe how far daughters are in a position to contribute to parental well-being—both in absolute

terms, and relative to what is expected from sons. We then explore how these sex preferences are reflected quantitatively in survey data on family-building desires and in measures of excess female child mortality.

2. Family Systems

The strongest son preference is evident in societies which are rigidly patrilineal and patrilocal. Patrilineality implies that group membership and productive assets are passed on through the male line. Patrilocality implies that it is normative for a woman to take up residence with her husband's family after marriage—often outside her parental village. In these societies, in which both inheritance and marriage rules are heavily weighted in favor of men, women have little economic independence or autonomy. Furthermore, since daughters move away from their birth homes after marriage, parents can have little expectation of subsequent physical or emotional sustenance from them since they are not living in close proximity. Even if women are educated and employed, the fruits of their labor are monopolized by their husband's family, and parents cannot claim a share of their daughters' earnings. On the other hand, a son is expected to provide his parents with economic security and insurance. He is also expected to bring in a daughter-in-law who will take care of his parents' physical needs in their old age. In these societies, women also gain in prestige and status if they bear sons to continue the family line, so both women and men have a strong interest in producing sons.

In some patrilineal societies, there are additional norms which increase the burden to parents of raising daughters. Religious beliefs and social customs such as dowry and 'purdah' serve as further instruments of patriarchal control, and reinforce son preference. Furthermore, family honor often depends on the purity of patriarchal descent, which requires ensuring women's sexual purity. This often entails secluding women from outside activities—which implies reduced income-earning opportunities—and marrying them off at young ages, before they have had a chance to develop into independent people.

In contrast, where kinship rules and norms make fewer distinctions between males and females, women have greater voice in the household and in public spaces, and face fewer constraints in becoming independent economic and social actors. In these societies, women are not constrained from providing old-age support to parents, and are valued accordingly.

2.1 South and East Asia

Much of South Asia displays the patterns of patrilineal and patrilocal kinship systems. This is especially

strong in the northern part of the subcontinent. The widespread practice of dowry payments constitutes a drain on family resources, further raising the relative cost of a daughter (Miller 1981). In Bangladesh, as in much of the rest of the subcontinent, daughters are perceived as burdens due to the custom of dowry and their limited income opportunities, while sons are preferred due to their higher earning potential (Muhuri and Preston 1991).

Much of East Asia is also characterized by rigidly patrilineal kinship organization, although in recent decades women have become educated along with men and have been active in the industrial labor force. Patrilineality is most pronounced in South Korea and amongst the majority Han population of China (Goodkind 1996). Moreover, it is believed that only sons can give their parents the crucial support they need in their afterlife, to prevent their being hungry spirits for eternity. This puts additional pressure on people to have sons even if pension plans and other mechanisms for providing for one's old age are in place, as they cannot provide for the afterlife—this can only be done by sons. (Japan is an exception to the East Asian pattern, practicing a very flexible form of patrilineal kinship organization which permits daughters under many circumstances effectively to substitute for sons; Nakane 1967.)

However, in Kerala and Sri Lanka, kinship systems allow women greater flexibility to interact with their parents after marriage than in the north (Dyson and Moore 1983). Village endogamy and cross-cousin marriages are favored, allowing women to live closer to their parents. Women sometimes transfer and inherit property as well. Thus there is considerable scope for mutually supportive relations between a married woman and her parental household.

2.2 Southeast Asia

Most Southeast Asian societies have bilateral kinship systems, which allow people to maintain strong links with both their family of birth and their family of marriage. Reciprocal obligations and expectations are more flexible than in strongly patrilineal societies. In much of Southeast Asia—as in Thailand, Indonesia, and the Philippines, for example—women have traditionally had considerable social and economic autonomy, playing a significant role in marketing and other independent income-generating activities. This makes it possible for them to provide old-age support for their parents should it be needed. In some cases, daughters are the main source of support—in Northern Thailand, for example, the norm is for newly married couples to live with the wife's parents for some years, and for the last daughter to stay on and inherit the land and house after the parents' death (Limanonda 1995). Although Islam has been actively

practiced in several of these societies for some centuries, this had relatively little impact on traditional kinship systems, as documented for Malaysia (Kling 1995).

On the other hand, the Chinese populations of Southeast Asia—including those in Malaysia and Singapore—display patrilineal and patriarchal kinship systems (Wongboonsin and Ruffolo 1994). Vietnam is an interesting amalgam of kinship systems, with strongly patrilineal values stemming from the incursion of Confucian ideologies from the north, superimposed on more bilateral Southeast Asian kinship systems, resulting in moderate levels of son preference (Goodkind 1994). As in China, sons alone can perform traditional rights of ancestor worship.

2.3 Middle East and North Africa

North Africa and the Middle East are characterized by patrilineal systems in that the family line is traced through males, but people are permitted to marry within their lineage, including cousins. This means that women often live in close proximity to their own parents, and can be a source of physical and emotional support to them. Economic reasons for son preference also do not appear to be strong—in Egypt, for example, parents value the labor of both sons and daughters, marriage expenses may actually be higher for males, and daughters can remain close to their parents after marriage (Makinson 1986). There are significant groups of educated and employed women in most of these countries, especially amongst the middle classes (Rostam-Kayali 1999).

However, the extent to which daughters can provide material support is constrained by the norms of Islam and the Arab family, whereby the husbands are the dominant decision makers in the household and most women are discouraged from active participation in the world outside the home (Obermeyer 1995). Normative pressures for bearing sons is strong, and women derive social status while men continue their name and lineage through sons. Overall, women are not as well placed as in Southeast Asia to contribute to their parental households, but not nearly as constrained from doing so as in East Asia and the northern parts of South Asia—making for moderate rather than strong son preference in the region.

2.4 Sub-Saharan Africa

Sub-Saharan Africa includes societies which are patrilineal, bilateral, and matrilineal. Despite these apparent differences, they have a great deal in common in terms of gender relations. Men tend to be dominant in the household—even in several of the matrilineal societies, property is passed through the female line,

but from men to men (Kevane and Gray 1996). However, conjugal bonds in sub-Saharan Africa are weaker than in South Asia, and women have a higher degree of economic independence (Cain 1984). They have freedom to remarry upon divorce or widowhood (Bledsoe 1990). Marriage often involves bride-price from the groom to the kin of the prospective bride, rather than dowry (Oppong 1987).

Women typically function easily as independent economic actors who can operate comfortably in public arenas—whether as self-employed market women, or as managers of their own farm plots. In several West African societies, women are very active traders and are completely accustomed to being the mainstay of the household, through marital union and dissolution. There is substantial flexibility in terms of those whom people can choose to live with and support, and parents have little reason to be desperate about old-age support if they do not have sons. Hence, one is unlikely to see strong preferences for children of a particular sex on the part of parents.

2.5 Latin America

Central and South American societies are formed from the confluence of many different ethnic groups, each with their own customs and social organization. For all their differences, however, these societies derive from cultures which are not heavily patrilineal or gender differentiated. The dominant culture is that of those of European origin, mixed with those of sub-Saharan African origin, neither of which place rigid constraints on women's participation in public life and interaction with their parental household. The Latin American family structure allows women a high degree of female autonomy. The conjugal bond is weaker than in countries of South Asia, the Middle East, and North Africa (Cain 1984). Childbearing outside formal marriage or stable unions is common, and women may spend much of their reproductive lives in consensual unions with men. Economic dependence of women on men is inconsistent with such family patterns. The result is that there is a strong tradition of male machismo, but there are few rigid rules constraining women from participating actively in social and economic life, or from helping their parents. Thus, there is little reason to expect strong gender preferences among parents—and whether or not they receive support from daughters is a matter of choice and circumstance, rather than one of custom and ability.

3. Quantitative Evidence of Son Preference

Son preference is manifested in quantitative data in various ways. In this section, we focus on two measures—survey data on desire for additional chil-

dren by number and sex of living children, and the sex ratios of children across developing countries—to examine the correspondence between family systems and these measures of revealed son preference.

As an indicator of sex preference among currently married nonpregnant women, the Demographic and Health Surveys (DHS) asked respondents whether or not they wanted another child, by sex distribution of their living children. The results from the latest available country studies are reproduced in Table 1, for women with two and three living children of the same sex. (This indicator is limited in that it focuses on a selected sample of women with a specific number of children rather than on the entire population of women of childbearing age, but is suggestive, nonetheless.) Son preference is indicated if, for a given number of children, a lower percentage of women wanted an additional child if all their children were sons, as compared with if they were all daughters. The number of children at which such preferences manifest themselves depends on the overall level of fertility in the country—it is only when people approach their desired family size that they are compelled to make decisions on additional childbearing to achieve their desired sex composition. Thus, where overall levels of fertility are high—as in some countries of South Asia, the Middle East, and North Africa, for example—gender preferences manifest themselves at higher numbers of children than where fertility levels are low.

Another way to measure sex preference is to examine the sex ratios of children aged below five years (Fig. 1). This reflects the extent to which the stated preference for sons actually translates into significant levels of excess female child mortality. Excess female child mortality can result from discrimination before birth (sex-selective abortion), at birth (female infanticide), and during early childhood (most commonly through neglect in providing timely health care). These are all motivated by the desire to rear more sons than daughters, and which method is used depends largely on the technological choices available.

There is no single 'expected' ratio of boys to girls, as this depends on the level of mortality. The observed sex ratio at birth is around 1.05 on average, and excess male mortality in early infancy is the biological norm (Waldron 1998). A recent study of 82 developing countries that had populations in excess of a million in 1990 confirms that mortality among male infants (aged below a year) is higher than that of females, the median sex ratio of infant mortality being 1.18—although actual figures vary around this median, with lower ratios in societies with higher life expectancy (United Nations Secretariat 1998). In the age-group one to four, the median sex ratio of child mortality is about 1. Based on this information, sex ratios among children aged zero to 4 can be expected to be below 1.05, *on average*. We have used a cut-off ratio of 1.045 (see Fig. 1), and sex ratios above this expected value

Table 1

Percentage of currently married nonpregnant women who want another child, by number and sex of living children

	Two Children			Three Children		
	No sons	2 sons	2 sons/No sons	No sons	3 sons	3 sons/No sons
South Asia						
Bangladesh	65	39	0.60	48	20	0.42
India	57	22	0.39	50	11	0.22
Nepal	77	25	0.32	66	10	0.15
Pakistan	77	61	0.79	79	33	0.42
Sri Lanka	47	36	0.77	18	16	0.89
Unweighted mean			0.57			0.42
Southeast Asia						
Indonesia	47	50	1.06	35	26	0.74
Philippines	44	46	1.05	24	30	1.25
Thailand	33	26	0.79	19	20	1.05
Unweighted mean			0.97			1.02
Middle East and North Africa						
Egypt	48	38	0.79	27	18	0.67
Jordan	75	66	0.88	76	55	0.72
Morocco	81	68	0.84	72	38	0.53
Tunisia	66	58	0.88	56	33	0.59
Turkey	25	13	0.52	18	4	0.22
Unweighted mean			0.78			0.55
sub-Saharan Africa						
Botswana	75	75	1.00	64	52	0.81
Burkina Faso	89	91	1.02	76	83	1.09
Burundi	92	91	0.99	89	78	0.88
Cameroon	79	83	1.05	84	73	0.87
Central African Republic	76	73	0.96	72	68	0.94
Ghana	81	80	0.99	60	52	0.87
Kenya	68	69	1.01	47	45	0.96
Liberia	85	87	1.02	95	82	0.86
Madagascar	83	86	1.04	65	70	1.08
Malawi	75	69	0.92	62	68	1.10
Mali	79	82	1.04	79	70	0.89
Namibia	52	65	1.25	46	61	1.33
Niger	88	92	1.05	75	82	1.09
Nigeria	79	83	1.05	82	76	0.93
Rwanda	91	80	0.88	72	71	0.99
Senegal	92	88	0.96	86	81	0.94
Sudan	86	86	1.00	81	77	0.95
Togo	95	91	0.96	83	85	1.02
Uganda	96	95	0.99	88	99	1.13
Zambia	90	90	1.00	87	89	1.02
Zimbabwe	82	79	0.96	63	73	1.16
Unweighted mean			1.01			1.00
Latin America						
Bolivia	29	28	0.97	19	16	0.84
Brazil	31	31	1.00	15	15	1.00
Colombia	26	36	1.38	8	13	1.63
Dominican Republic	42	49	1.17	15	13	0.87
Ecuador	39	36	0.92	26	23	0.88
Guatemala	56	58	1.04	43	41	0.95
Mexico	37	37	1.00	24	12	0.50
Paraguay	57	63	1.11	47	53	1.13
Peru	23	27	1.17	14	14	1.00
Trinidad and Tobago	36	44	1.22	24	29	1.21
Unweighted mean			1.10			1.00

Source: Arnold 1997

For countries with results from multiple years, the latest year is reported

Map 1: Juvenile Sex ratios in Africa, Asia, and Latin America, 2000

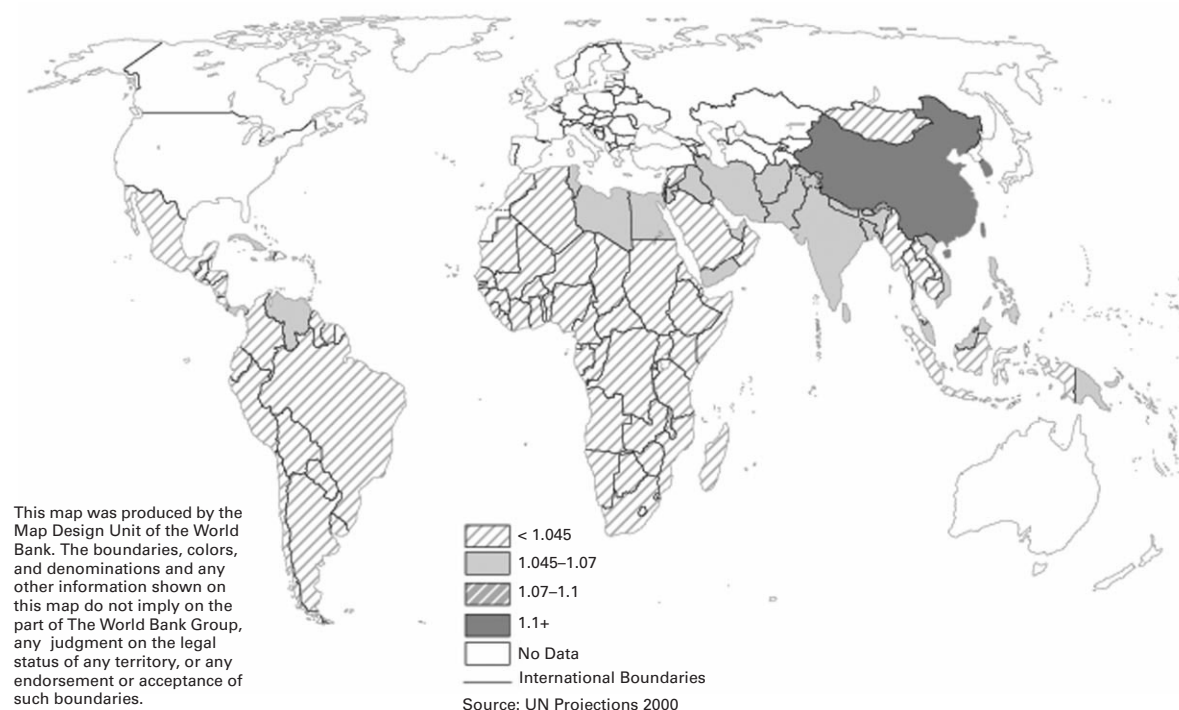


Figure 1
Juvenile sex ratios in Africa, Asia, and Latin America, 2000

can be viewed as indicative of excess female child mortality. The higher the observed sex ratio above this cut-off, the stronger the manifestation of son preference.

The data presented in Table 1 and Fig. 1 are consistent with the broad regional differences in family systems outlined above. In Latin America, sub-Saharan Africa, and Southeast Asia, where there is a higher degree of female autonomy, there is little consistent evidence of son preference in either juvenile sex ratios or desire for additional children. In Latin America, there are actually instances of a mild preference for daughters, in countries such as Colombia, Trinidad and Tobago, and Paraguay (Table 1).

On the other hand, in regions of the world where patrilineal and patrilocal family systems prevail, the data reveal a distinct preference for sons. Son preference in the desire for additional children is evident in all countries in the Middle East and North Africa, and even more strongly in South Asia, at a parity of 2 (Table 1). (Since the DHS surveys were not conducted in East Asia, these observations are missing from the table.) These preferences are strengthened at a parity of 3—reflecting that in high-fertility regions, son preferences are more evident at higher parities. (Only

Sri Lanka in South Asia displays a different pattern; see Table 1.) These regions, along with China and South Korea in East Asia, have distinctly higher juvenile sex ratios than average as well—suggesting excess female mortality (Fig. 1).

Estimates indicate that in China and South Korea, around 7 percent of girls were ‘missing’ in 1990 on account of excess prenatal or postnatal mortality (Das Gupta et al. 1997). Globally, estimates indicate that there are between 60 and 100 million fewer women than there would be in the absence of gender discrimination—mostly in South and East Asia (Coale 1991, Klasen 1994, Sen 1992). (These estimates are based on comparing the actual sex ratio of the population against the sex ratio that would theoretically exist in the absence of gender discrimination.)

Since China and India are large countries with considerable cultural heterogeneity, it is instructive to look at province-level data (Fig. 2). This shows patterns which accord further with the description of family systems above. In India, the northwest shows sex ratios as excessively masculine as South Korea and the mainstream Han-populated provinces of China. South India shows less discrimination against girls, as do the interior provinces of China, where minority

Map 2: Juvenile Sex ratios in Provinces of China and States of India, 1990-91

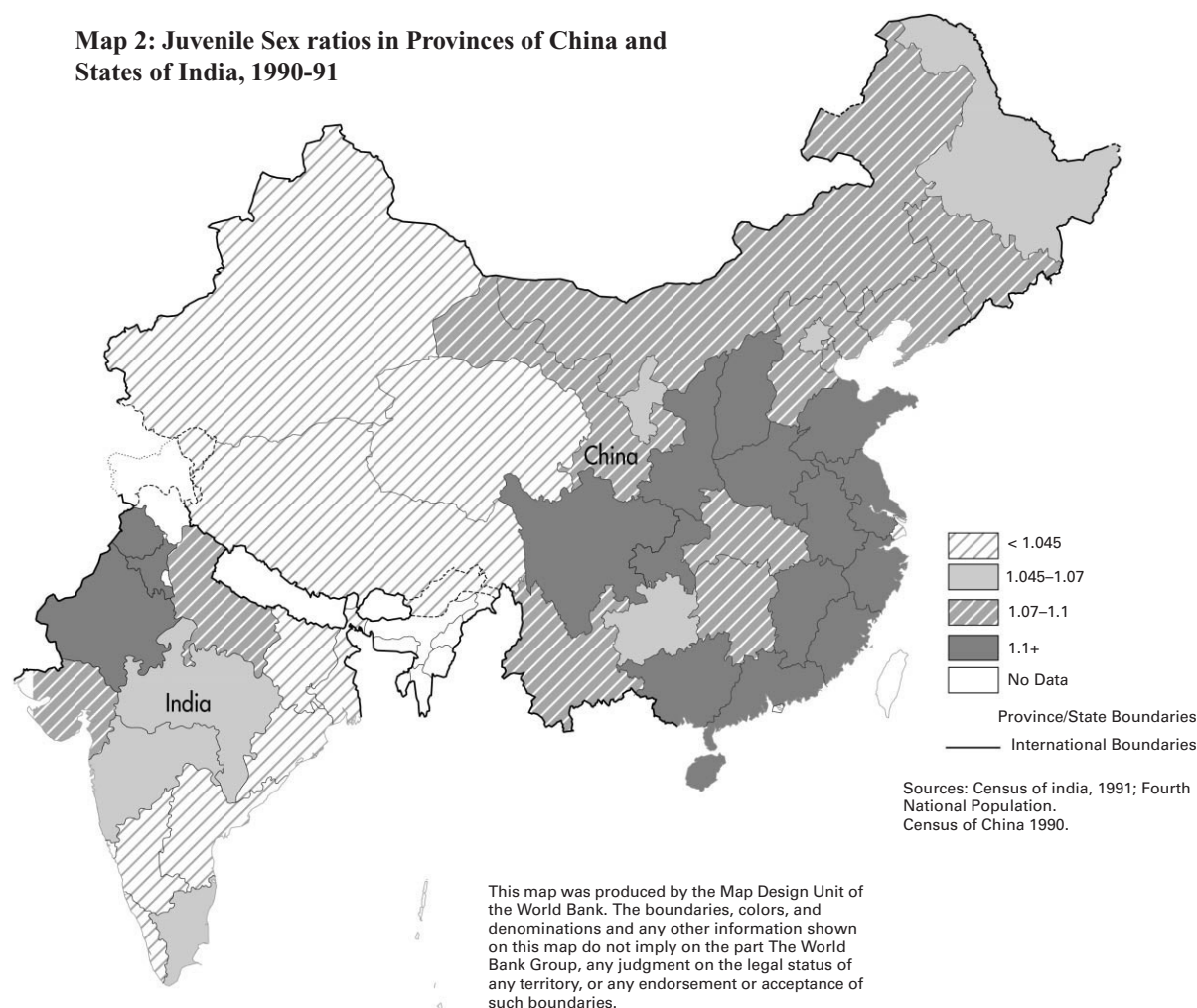


Figure 2
Juvenile sex ratios in provinces of China and states of India, 1990-1

ethnic groups are clustered. Thus the strongest evidence of excess female child mortality is from the regions of East and South Asia characterized by rigidly patrilineal kinship systems. Similar consistency between kinship systems and son preference is evident in the data from Malaysia, where the Malays with their Southeast Asian patterns of bilateral kinship show little sign of excess female child mortality, while the Chinese and Indians show it clearly (Goodkind 1996).

4. Other Factors Affecting Son Preference

Although son preference is culturally determined, the extent to which the preferences within a given society are translated into actual discrimination against

daughters is influenced by the pressures operating on parents at a given point in time. Such pressures can take many forms. For example, in times of adverse income shocks, investment in girls may suffer more than investment in boys. (See, for example, Behrman and Knowles 1998 for Vietnam, which suggests that the income elasticity of investment in daughters is higher than that of investing in sons. This implies that if household income falls, discrimination against daughters will rise.) It is not that parents are unwilling to invest in children, but that they would prefer to invest less in daughters who will be subsequently lost to them.

External crises, such as famine and war, increase the pressures on parents to discriminate against daughters, as the difficulties of nurturing all household members

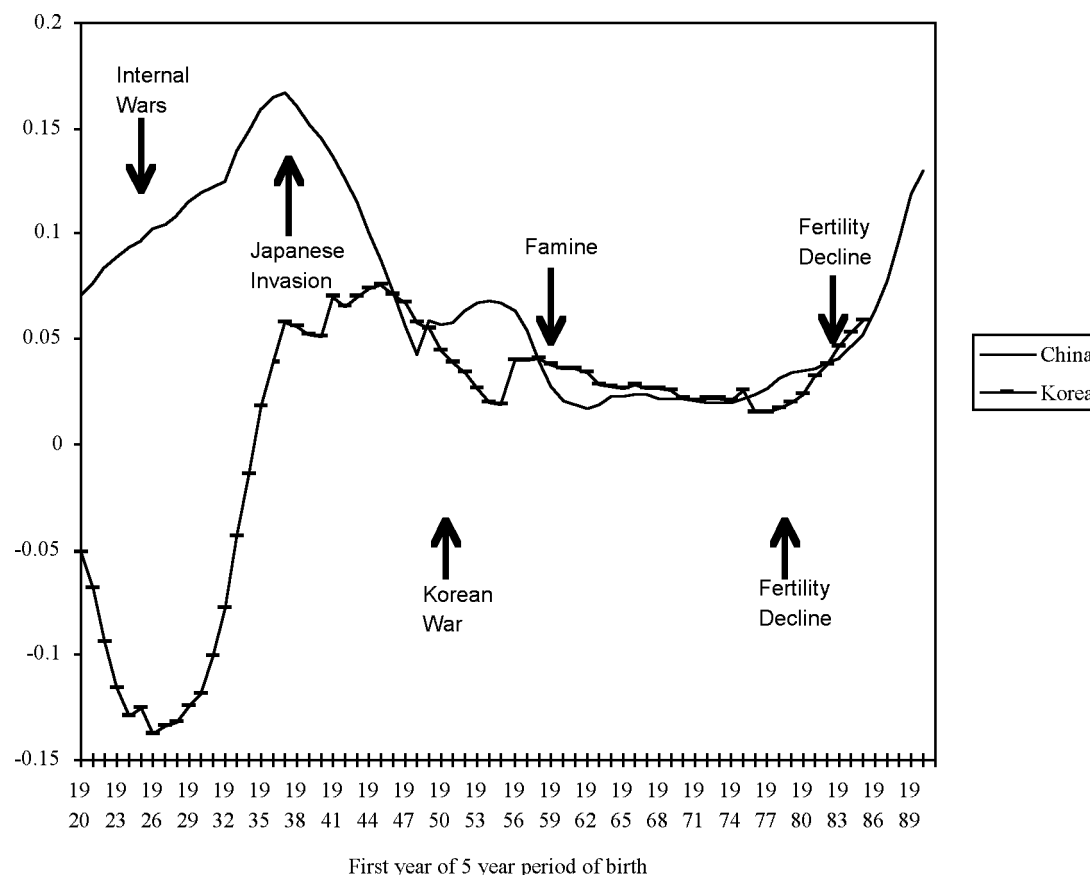


Figure 3

Excess sex ratios at census time by five-year birth cohorts in China and South Korea, 1920–95

Note: The excess ratios peak amongst cohorts born just before a war or famine, because those who were young girls at the time of the crisis experienced the maximum excess mortality

Source: Das Gupta and Li 1999

rise and harsher choices have to be made. Fig. 3 shows how the extent of actual discrimination against girls was affected by wars and famines in China and South Korea during this century (Das Gupta and Li 1999). (China experienced many severe crises during the twentieth century—during the first half of the century, civil war and invasion in China led to an enormous amount of disruption, which was followed by the famine of 1959–61.) The South Korean data show a massive deficit of males associated with the loss of young men during the Korean War in the early 1950s, followed by a rise in sex ratios as fertility declined.

Fertility decline also adds to the pressures to discriminate against daughters. When parents desire a smaller number of children, this further limits the number of daughters they would like to rear. If, for example, they want to have no more than two children, and the first is a girl, they are under severe pressure to ensure that the next is *not* a girl. Thus child sex ratios

were observed to become more masculine when fertility declined significantly in India, China, and South Korea (Das Gupta et al. 1997).

5. Conclusions

This entry provides a broad overview of family systems in the developing world, and how they influence sex preference for children. In general, evidence for son preference is strongest where women are not in a position to contribute to the well-being of the family.

See also: Family and Gender; Family, Anthropology of; Family Size Preferences; Family Theory: Competing Perspectives in Social Demography; Family Theory: Economics of Childbearing; Fertility Control: Overview; Fertility: Proximate Determinants; Gender-

related Development; Infant and Child Mortality in the Less Developed World; Parenting: Attitudes and Beliefs; Sex Preferences in Western Societies

Bibliography

- Behrman J R, Knowles J C 1998. Household income and child schooling in Vietnam. *World Bank Economic Review (International)* **13**: 211–56
- Bledsoe C 1990 Transformation in sub-Saharan African marriage and fertility. *Annals of the American Academy of Political and Social Science* **510**: 115–25
- Cain M 1984 *Women's Status and Fertility in Developing Countries: Son Preference and Economic Security*. World Bank, Washington, DC
- Coale A J 1991 Excess female mortality and the balance of the sexes: An estimate of the number of 'missing females'. *Population and Development Review* **17**: 517–23
- Das Gupta M, Li S 1999 Gender bias and the marriage squeeze in China, South Korea and India 1920–1990: The effects of war, famine and fertility decline. *Development and Change* **30**: 619–52
- Das Gupta M, Jiang Z, Xie Z, Li B, Nam-Hoon C, Woojin C, Mari Bhat P N 1997 *Gender Bias in China, South Korea and India: Causes and Policy Implications*. Harvard University Press, Cambridge, MA
- Dyson T, Moore M 1983 On Kinship structure, female autonomy and demographic behavior in India. *Population and Development Review* **9**: 35–60
- Goodkind D M 1994 Sex preference for children in Vietnam. Paper presented at the International Symposium on Issues Related to Sex Preference for Children in the Rapidly Changing Demographic Dynamics in Asia, Seoul, Republic of Korea, 21–4 November
- Goodkind D M 1996 On substituting sex preference strategies in East Asia: Does prenatal sex selection reduce postnatal discrimination? *Population and Development Review* **22**: 111–25
- Kevane M, Gray L 1996 'A woman's field is made at night': Gendered land rights in and norms in Burkina Faso. *Feminist Economics* **5**: 1–26
- Klasen S 1994 'Missing Women' reconsidered. *World Development* **22**: 106–71
- Kling Z 1995 The Malay family: Beliefs and realities. *Journal of Comparative Family Studies* **26**: 43–66
- Limanonda B 1995 Families in Thailand: Beliefs and realities. *Journal of Comparative Family Studies* **26**: 67–82
- Makinson C 1986 *Sex Differentials in Infant and Child Mortality in Egypt*. Ph.D. thesis, Princeton University
- Miller B D 1981 *The Endangered Sex: Neglect of Female Children in Rural North India*. Cornell University Press, Ithaca, NY
- Muhuri P K, Preston S H 1991 Effects of family composition on mortality differentials by sex among children in Matlab, Bangladesh. *Population and Development Review* **17**: 415–34
- Nakane C 1967 *Kinship and Economic Organization in Rural Japan*. Athlone Press, London
- Obermeyer C M (ed.) 1995 *Family, Gender and Population in the Middle East*. American University in Cairo Press, Cairo, Egypt
- Oppong C (ed.) 1987 *Sex Roles, Population and Development in West Africa: Policy-related Studies on Work and Demographic Issues*. Heinemann, Portsmouth, NH

- Rostam-Koyali J 1999 The politics of women's rights in the contemporary Muslim world. *Journal of Women's History* **10**: 205–15
- Sen A 1992 More than 100 million women are missing. *New York Review of Books* **December**: 61–6
- United Nations Secretariat 1998 Levels and trends of sex differentials in infant, child, and under-five mortality. In: *Too Young to Die: Genes or Gender*. United Nations, New York, chap. 4, pp. 84–108
- Waldron I 1998 Sex differences in infant and early childhood mortality: Major causes of death and possible biological causes. In: *Too Young to Die: Genes or Gender*. United Nations, New York, chap. 3, pp. 64–83
- Wongboonsin K, Ruffolo V P 1994 The future of Thailand's population policy: Potential directions. Paper presented at the International Symposium on Issues Related to Sex Preference for Children in the Rapidly Changing Demographic Dynamics in Asia, Seoul, Republic of Korea, 21–4 November

A. Basu and M. Das Gupta

Family Systems in Europe

Contributing to a long tradition of scholarship on family life in Europe, a group of scholars in the 1960s redirected research by arguing that a unique system characterized the traditional European family. They criticized what they took to be the received wisdom on European family life in the past, which portrayed the nuclearization of the coresidential family as a recent product of industrialization and urbanization. The final decades of the twentieth century saw increasing criticism of this revisionist perspective, leading to a more nuanced and theoretically sophisticated understanding of the nature of coresidential processes in the European past.

1. A European Family System

The debate over the existence of a European family system has revolved around two related elements—a marriage system and a household system—emphasizing the demographic aspects of each. The former focuses on the age at which men and women marry, as well as the proportions who never marry. The latter concerns the principles by which coresidential units form and dissolve. The two are linked via the principles governing postmarital residence, that is, where newlyweds live following marriage.

These are only two aspects of what could be seen as the larger matrix of a family system. The anthropologist Goody (1983) has called attention to a variety of features that together historically differentiated European family dynamics from those found elsewhere in the world—including a prohibition on poly-

gamy and child adoption, and various restrictions on marriage to kin—differences which he attributes in good part to the influence of the Church. An understanding of family systems more generally would also include consideration of kin ties beyond the household; however, relatively little empirical attention has been focused on such issues historically.

1.1 From Le Play to Laslett

Most prominent among the nineteenth-century progenitors of the social scientific study of European family systems was the Frenchman, Le Play (1895), who combined a series of empirical studies of families across much of Europe with pronouncements about the evolution of family life in the west, warning of an alarming erosion in traditional family forms that was eating at the moral fabric of European society. He contrasted what he took to be the traditional family, consisting of a large, patrilineally extended group of kin under the senior male's autocratic control, with the unstable family that was rapidly spreading throughout the industrializing population of a rapidly changing Europe.

Between the two extremes was the household form that Le Play thought best, the stem family, which mixed a high degree of respect for authority and concern for other kin with an element of independence and room for innovation. The farmstead lay at the center of stem family life, and was passed on, intact, from generation to generation, to a single child, preventing land fragmentation.

Le Play believed that the large patriarchal family had already receded in Europe, remaining primarily among Eastern nomads, Russian peasants, and the Slavs of Central Europe. The epic struggle that gripped nineteenth-century Europe, in Le Play's view, pitted the stem family against the unstable family.

Such evolutionary perspectives as Le Play's were taken up by American sociologists in the middle decades of the twentieth century. Many assumed that large, complex family households had characterized the European past, and were being ineluctably replaced by small, nuclear family households as a result of modernization, however defined (Goode 1963). Challenging these assumptions, Laslett (1965), using British historical evidence, argued that the view of the large, complex household as typical of the European past was a 'myth.' According to Laslett, western Europe was long characterized by neolocality (i.e., couples established their own independent households at marriage) and it was rare for people in the past to live in large, complex family households. The Cambridge Group for the History of Population and Social Structure, which Laslett and Wrigley cofounded at this time, became an influential center for historical household studies, and the 1972 volume edited by Laslett and Wall, *Household and Family in Past Time*,

set the terms of the debate for years to come and spawned hundreds of local-level studies throughout Europe.

1.2 The European Marriage Pattern

Just as Laslett was attacking the myth of the European extended family, Hajnal (1965) advanced the claim that a unique marriage system had characterized the whole of Europe except for the east and southeast for at least two centuries preceding 1940. It had two key features: a high age of marriage and high proportions of people who never married. A line could be drawn, he argued, stretching from St. Petersburg (then Lenin-grad) southwest to Trieste. To its east the system was not found.

Two decades later, Hajnal (1983) partially retreated from his by then famous 'line,' portraying the European marriage pattern as characterizing northwestern Europe and excluding not only eastern Europe, but the south as well. In this reformulation, he emphasized a third characteristic of the European family system, in addition to high marriage age (men 26, women 23) and high proportions (10 per cent or more) never marrying: young people in large numbers left their parental home and spent the years before marriage circulating between houses of nonkin in the capacity of servants. It was this circulation of servants that gave flexibility to the nuclear family system, facilitating the redistribution of a household-based rural labor force in response to each household's labor needs.

1.3 Geographic Diversity

As Hajnal's retreat from his own 'line' suggests, the proponents of a traditional European family system soon found themselves under heavy criticism from scholars who determined that some form of a more complex household system had indeed been typical of the areas they were studying. Both Hajnal and Laslett attempted to salvage their model of the European family system by restricting its territorial reach. It is revealing that French family historians never participated in these efforts to find a single European family system, for they were all too aware of the fact that even within a single country—in this case, their own—more than one family system had existed. In northern France a system much like that described by Laslett and Hajnal appeared, but in the central and southern regions larger, complex family households were common in past centuries. Similarly, work on Spain (Reher 1997), Portugal (Kertzer and Brettell 1987), and Italy (Barbagli 1984) found clear internal divisions: in each country nuclear family systems predominated in the south but more complex household systems were found in parts of the north. Nuclear family systems and neolocality were linked to systems

of partible inheritance or contexts in which there was little of productive value to inherit (as in the case of the proletariat).

Laslett (1983) responded to this new evidence by proposing a new division of traditional Europe into four areas—the West, Middle, Mediterranean, and East—each having a distinctive family system. It turned out that the ‘myth’ that Laslett had punctured, if it was a myth at all, pertained only to one of the four areas of Europe, the West. In this new model, not only did Eastern Europe have low female marriage age and high proportions of joint-family households, but so did Mediterranean Europe.

Laslett continued to hold to Hajnal’s notion of a family system in which neolocality, high female age at marriage, and the prevalence of service as a life stage go together. Yet the Italian case, among others, shows that this is not necessarily true: in Italy in past centuries, it was the central and northern areas having high proportions of joint family households that had high proportions of servants, while in the south, where nuclear family residence prevailed, life-cycle service was rare (Da Molin 1990). Moreover, age of female marriage was typically higher in the center and north than in the south, again going against the Hajnal/Laslett model.

2. *Rethinking European Family Systems*

Laslett’s later work on European family systems rejected crude generalizations about ‘the’ European family. A more fully satisfactory approach should build on these insights, recognizing the role played by political economic, demographic, and cultural forces in local contexts. It should also abandon the notion that northwestern Europe is the kernel of Europe and the lingering attempts to rescue a quintessential European family system.

2.1 *Political Economy*

Family historians frequently have pointed to two major, related economic pressures shaping family systems: the household as production unit, and inheritance norms. The traditional model of rural life in preindustrial times depicts a peasant society in which the household was the unit of production. Where this is the case, the composition of the household, and the timing of marriages, are shaped by labor needs and relations between peasants and the land they farm (or ownership of other productive resources, such as livestock). A dramatic case comes from central Italy, where sharecroppers typically lived in large joint family households. Landowners sought to maximize the number of adult laborers on each plot of land, to maximize their income (classically one-half of the produce). In the Alpine area in the north of Italy,

likewise, Viazzo (1989) describes a system in which complex family households prevailed, allowing peasants to diversify their holdings and their risks, combining herding with farming scattered plots of land at different altitudes.

Political forces entered into these dynamics both indirectly, via political support for various land-holding systems, and directly, as when systems of household taxation or military conscription influenced peasants to live in more extended or more nuclear households. In the case of serf systems of Eastern Europe, controls on movement out of estates had a similar influence (Czap 1982).

Inheritance systems are tied closely to family systems. In Spain, France, and elsewhere, areas practicing impartible inheritance tended to have stem family systems. On the other hand, where peasants did not own their own land, the link did not operate in this way. Italian sharecropping households practiced partible inheritance, but the pressures to have large numbers of coresiding adults resulted in joint family households.

2.2 *Demographic Constraints*

A continuing theme in debates about European family systems concerns the constraints placed on family dynamics by the brute facts of demography. In 1965, Levy argued that the great majority of people in the past lived in relatively small nuclear family households regardless of family formation norms, because high mortality rates interfered with the formation of extended family households. In critiquing Laslett’s original thesis of the prevalence of the nuclear household system in Europe, Berkner (1972) took up this idea, arguing that census data showing that the great majority of households at any one time are nuclear does not in itself provide evidence of a nuclear family system. Rather, he argued, under preindustrial demographic conditions, people following extended family norms would spend most of their lives living in nuclear households.

Given the workings of such demographic constraints, inheritance was typically a much more messy process than a focus on norms would suggest. A study of a nineteenth-century Austrian community, in which impartible inheritance was the rule, for example, found that large numbers of households heads died before they had a son old enough to take over. When their widows remarried, tensions between stepfather and sons resulted (Sieder and Mitterauer 1983). Yet work in sharecropping Italy and on the Russian serf estates shows that such demographic constraints were much less evident in joint family systems: where all sons are supposed to bring their brides into the household at marriage a nuclear phase is much less common, as many other combinations of coresiding kin are possible (Kertzer 1989, 1991, Czap 1982). Indeed, there is

evidence that in the nineteenth century the proportion of people living in large, complex family households was growing in several areas as a result of a combination of demographic and economic factors (Ruggles 1987). In central Italy, with a largely fixed number of sharecropper farms available, a growing rural population, and the primary alternative—agricultural wage labor—providing a much poorer living, there was increasing pressure on the children of sharecropper families not to strike off on their own at marriage, and in this area the nineteenth century saw the historical zenith in the proportion of people living in joint family households (Barbagli 1984).

3. Culture

Understanding European family systems requires not only attention to historically changing political economic and demographic forces, but to cultural norms regarding proper family life as well. For example, each society has norms about the proper timing for a woman's marriage. While in some settings marriage age was affected by changing economic conditions (i.e., marriage being postponed during hard economic times) (Wrigley and Schofield 1981), there is also much evidence of the resistance of cultural norms to economic pressures. One manifestation of this is the number of studies showing a very similar female marriage age across different economic segments of local stratified populations (Knodel 1988, pp. 130–6).

The great bulk of the work on European household systems—following Laslett and Hajnal—focuses on the cultural norm of postmarital residence as the key (identifying neolocality with nuclear family household systems). However, it is becoming clear that another cultural norm must be entered into the model: the fate of older people as they become frail or widowed (Kertzer 1995). Even where newlyweds routinely establish their own household, a pure nuclear family system does not necessarily result. Recent studies show that in the very heartland of Hajnal and Laslett's northwestern European family system, large numbers of older people did live in extended family households. Wall (1995), for example, in a sample of English communities for 1599–1796, found that half of the men aged over 65 and over a third of the women lived with at least one child. While some of these adult children were not married, many of them were.

Finally, while most work on European family systems has focused on household units and on marriage, any satisfactory understanding of the role of family must consider relations that extend beyond single households. Reher (1997, p. 114), for example, found that in the Spanish community of Cuenca in the middle of the nineteenth century, while neolocality prevailed about three-quarters of the newlyweds set up their households in the same immediate vicinity of one of their parental households.

In short, no single family system—in the sense intended by Hajnal or Laslett—characterized 'traditional' Europe, nor even western Europe. While broad general differences among family systems in different parts of Europe can be distinguished, attempts to divide the continent into distinct family zones are misguided. In those areas where complex family households were commonly found, significant sections of the population lived in simpler households, while even in the heart of northwestern Europe, the epicenter of the presumed nuclear family system, more complex household dynamics were often at work.

See also: Families and Households, Behavioral Demography of; Families and Households, Formal Demography of; Family and Kinship, History of; Historical Demography; Western European Studies: Gender and History; Western European Studies: Society

Bibliography

- Barbagli M 1984 *Sotto lo stesso tetto*. Il Mulino, Bologna, Italy
- Berkner L 1972 The stem family and the developmental cycle of the peasant household. *American Historical Review* 77: 398–418
- Czap P 1982 The perennial multiple family household, Mishino, Russia. *Journal of Family History* 7: 5–26
- Da Molin G 1990 *La famiglia nel passato*. Caccuci Editore, Bari
- Goode W J 1963 Industrialization and family change. In: Hoselitz B F, Moore W E (eds.) *Industrialization and Society*: UNESCO, New York
- Goody J 1983 *The Development of the Family and Marriage in Europe*. Cambridge University Press, Cambridge, UK
- Hajnal J 1965 European marriage patterns in perspective. In: Glass D V, Eversley D (eds.) *Population in History*. Edward Arnold, London
- Hajnal J 1983 Two kinds of pre-industrial household formation system. In: Wall R, Robin R, Laslett P (eds.) *Family Forms in Historic Europe*. Cambridge University Press, Cambridge, UK
- Kertzer D I 1989 The joint family household revisited: Demographic constraints and household complexity in the European past. *Journal of Family History* 14: 1–15
- Kertzer D I 1991 Household history and sociological theory. *Annual Review of Sociology* 17: 155–79
- Kertzer D I 1995 Toward a historical demography of aging. In: Kertzer D I, Laslett P (eds.) *Aging in the Past*. University of California Press, Berkeley, CA
- Kertzer D I, Brettell C B 1987 Advances in Italian and Iberian family history. *Journal of Family History* 12: 87–120
- Knodel J 1988 *Demographic Behavior in the Past*. Cambridge University Press, Cambridge, UK
- Laslett P 1965 *The World We Have Lost*. Methuen, London
- Laslett P 1983 Family and household as work group and kin group: Areas of traditional Europe compared. In: Wall R, Robin R, Laslett P (eds.) *Family Forms in Historic Europe*. Cambridge University Press, Cambridge, UK
- Le Play F 1895 *L'Organisation de la famille selon le vrai modèle signalé par l'histoire*, 4th edn. Alfred Mame, Paris
- Marion L Jr 1965 *Aspects of the Analysis of Family Structure*. Princeton University Press, Princeton, NJ

- Reher D S 1997 *Perspectives on the Family in Spain Past and Present*. Clarendon Press, Oxford, UK
- Ruggles S 1987 *Prolonged Connections: The Rise of the Extended Family in Nineteenth-century England and America*. University of Wisconsin Press, Madison, WI
- Sieder R, Mitterauer M 1983 The reconstruction of the family life course. In: Wall R, Robin J, Laslett P (eds.) *Family Forms in Historic Europe*. Cambridge University Press, Cambridge, UK
- Viazzo P P 1989 *Upland Communities: Environment, Population and Social Structure in the Alps since the Sixteenth Century*. Cambridge University Press, Cambridge, UK
- Wall R 1995 Elderly persons and members of their households in England and Wales from preindustrial times to the present. In: Kertzer D, Laslett P (eds.) *Aging in the Past*. University of California Press, Berkeley, CA
- Wrigley E A, Schofield R S 1981 *The Population History of England 1541–1871*. Edward Arnold, London

D. I. Kertzer

Family Theory and the Realities of Childbearing Behavior

Childbearing takes place within the socially defined context of family life, but there have been substantial changes in family life in recent decades and it differs between societies and population subgroups. While childbearing patterns have also changed substantially in developed societies since the 1950s, there has been a tendency for trends to be broadly similar (Fig. 1(a)). The annual level of fertility in the post-1945 period rose to a peak around the late 1950 to mid-1960s, followed by a sharp decline in the 1970s and a slight rise and relative stabilization in the recent period, although some countries such as those in Southern Europe have continued to decline to levels of fertility just above one child per woman in the late 1990s (Coleman 1996). The average age at childbearing fell until the late 1960s or early 1980s and then rose substantially (Fig. 1(b)).

Annual levels and mean age at childbearing are closely related since a move towards earlier childbearing will increase the annual fertility rate even if the women having their children in that period do not ultimately have any more children by the end of their reproductive lifetime than other cohorts of women. Most of the large-scale movements in fertility in the postwar period are attributable to shifts in the age at childbearing rather than to changes in the overall propensity to have a given number of children: 'most of the baby boom would have occurred without any change in the numbers of births per woman, and most of the decline since the baby boom as well' (Ryder 1980, pp. 39–40). While the level of childbearing in a given year may reflect contemporary factors such as income or employment, some women may be having

children then because they have postponed births from an earlier period or wish to bring their births forward for particular reasons. The fact that married women are much more likely to have a child than those not in a partnership, with cohabiting women generally in an intermediate position, means that earlier experiences of entry into and exit out of partnerships also need to be taken into account.

The main alternative theories that have been put forward to explain childbearing patterns will be reviewed briefly, before considering the extent to which they can be empirically validated.

1. Explanations of Childbearing

1.1 Proximate Determinants of Fertility

Fertility is often usefully analyzed in terms of a proximate determinants framework (Bongaarts and Potter 1983) which for developed countries means concentrating on three main factors; the proportion of women in a sexual union, their use of contraception, and the level of induced abortion. These factors biologically determine childbearing, but women's choices about partnership and family limitation may also be consequences of their fertility decisions. One reason why women may enter a partnership is in order to have children, but it is unlikely to be the only one. On the other hand, use of contraception and abortion on 'social' grounds are determined by fertility preferences. Over time, there has been a change from a period when methods of family limitation were generally ineffective, difficult to obtain, and/or disapproved of, to one when methods are safe, easy to use, relatively inexpensive, reliable, and acceptable. This change occurred with the introduction of the contraceptive pill and IUD around 1960 in many developed countries, and the subsequent introduction of laws permitting abortion on wider grounds than previously. Studies in the US showed that the level of 'unwanted' fertility in the 1960s was about 0.5 of a child per woman (Westoff and Ryder 1977), similar to the difference in cohort fertility between women completing their childbearing at that time and today. The availability and acceptability of efficient contraceptive methods has had a major contribution to the change in timing which is the major determinant of overall period change. The effect of access to improved contraception not only affects childbearing, but permits couples to have a childless relationship (and, in particular, to facilitate the emergence of cohabiting relationships that have low rates of childbearing) and to decide on the pattern of childbearing with greater precision, since, for example, the possibility of an unwanted birth can be eliminated by contraceptive sterilisation. In turn, this has permitted the increase in mean age at childbearing and hence led to changing attitudes about women's roles since a full satisfying

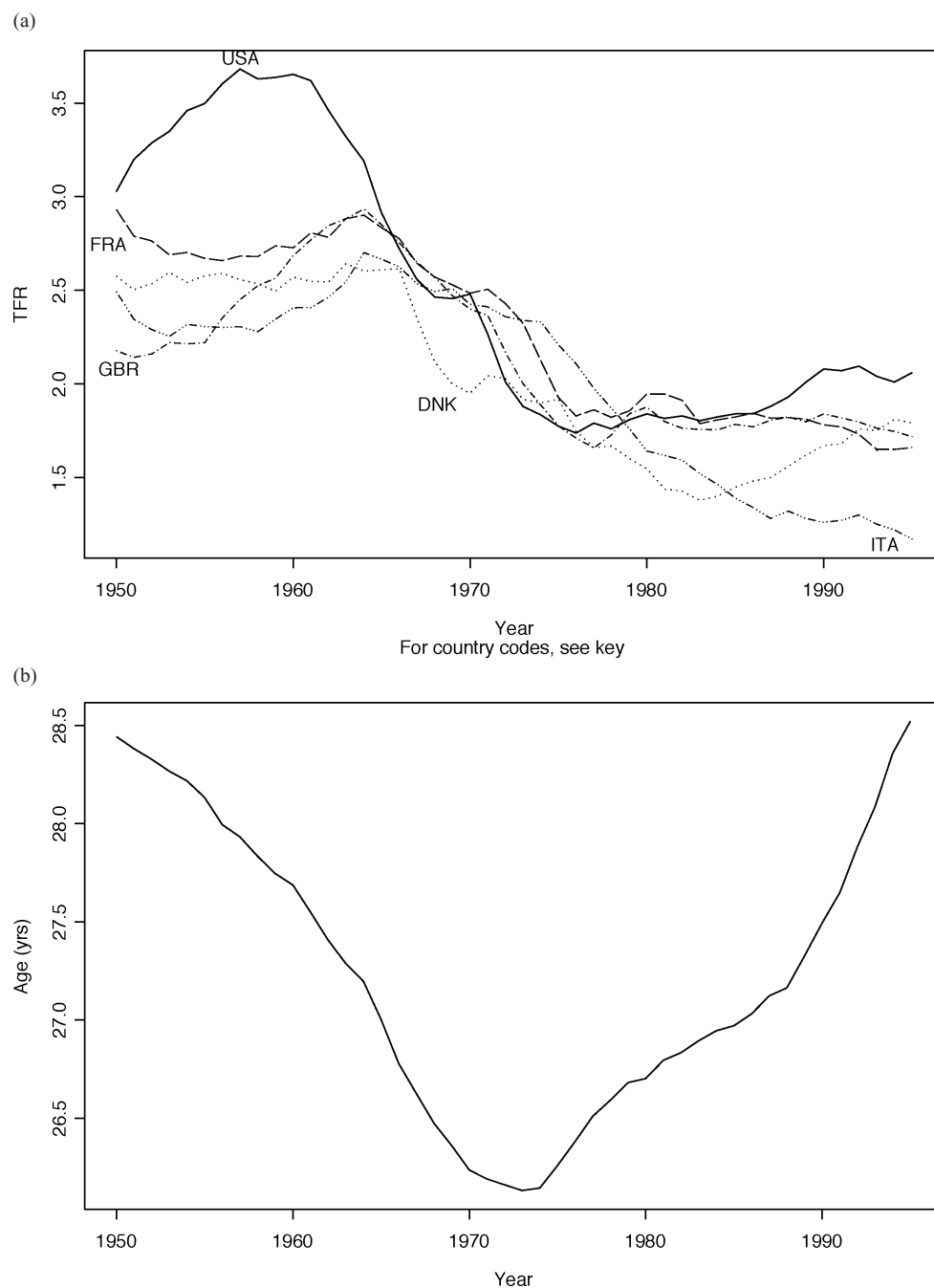


Figure 1
 (a) Total fertility rate, selected OECD countries, 1950–95. (b) Mean age at childbearing, England and Wales, 1950–95. Key for charts: AUS, Australia; AUT, Austria; BEL, Belgium; CAN, Canada; DNK, Denmark; FIN, Finland; FRA, France; DEU, Germany; GRC, Greece; ISL, Iceland; IRL, Ireland; ITA, Italy; JPN, Japan; KOR, Rep. of Korea; LUX, Luxembourg; NLD, Netherlands; NZL, New Zealand; NOR, Norway; PRT, Portugal; ESP, Spain; SWE, Sweden; CHE, Switzerland; GBR, England and Wales; USA, United States (Source: World Bank and National Statistical Offices)

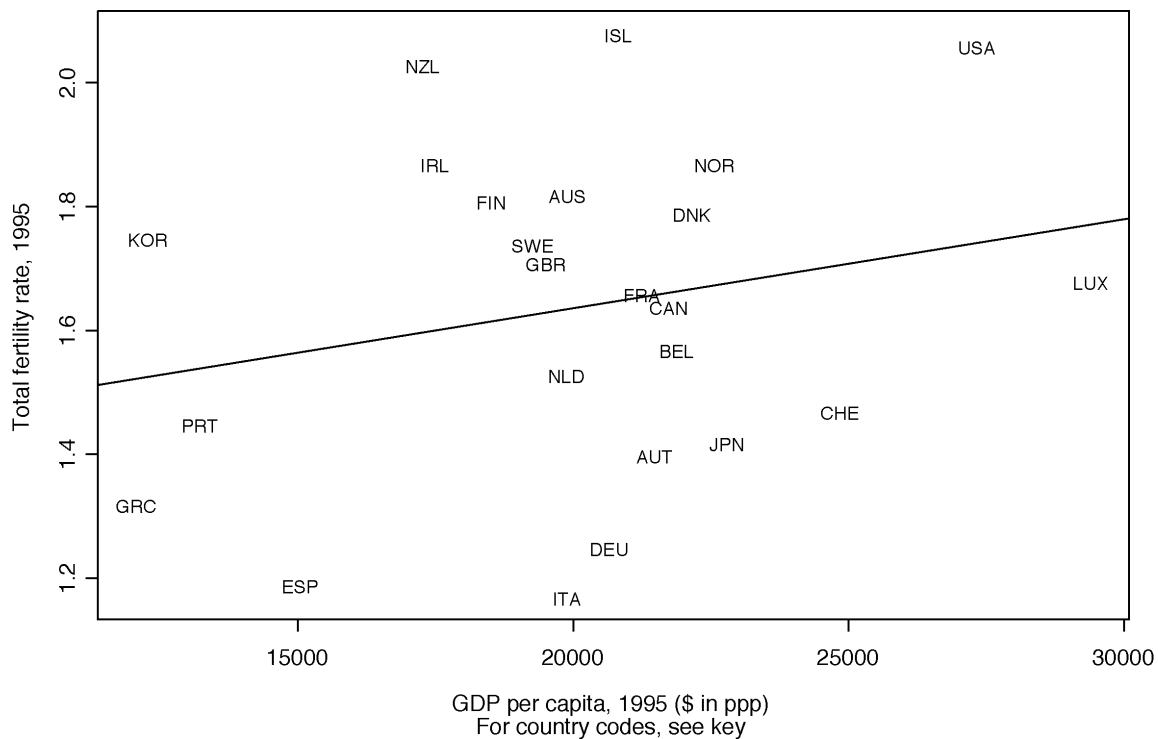


Figure 2
Total fertility rate vs. GDP per capita for a range of countries in 1995

sex life is now compatible with career progression by married women that would have been difficult in the past.

As improved access to methods of family limitation have increased across and within nations, variability in fertility has declined, so that the maximum difference in TFR levels between the OECD member countries in 1995 was 0.9 of a child, compared with a value of 1.9 in 1965 and 1.8 in 1980.

1.2 Economic Approaches

In the past, a positive relationship between fertility and income (in cross-section) and with the business cycle (in time trend) had been accepted. However, in the period since 1945, fertility and real incomes generally have been negatively rather than positively related in developed countries, since fertility is now at an all time low, whereas average real incomes have risen. The current widely used microeconomic model of fertility is the 'new home economics' one particularly associated with the work of Becker (1981) that has been extended to both marriage and marital breakdown.

The model brought together two features within the framework of microeconomic analysis to account for the postwar baby boom and bust: the first was the increasing importance of women's work outside the home and its competition with childrearing for the wife's time. Since women undertake the greater burden of childrearing, the costs of childbearing, especially in terms of earnings foregone while they are looking after children and future labor market prospects, fall most heavily on them. Therefore, the number of children is expected to be strongly negatively related to her earnings, but weakly positively related to her husband's earnings, and the opportunity costs of childbearing are greatest for women in highly paid jobs.

The second element, the demand for children, is determined in part by economic factors and in part by 'tastes' which may differ by religion, race, age, and time period, etc. Decisions concerning childbearing have two major components, those related to quantity and 'quality': 'quality' is simply measured by the amount spent on a child on items such as clothing, education, etc. ('higher quality' does not mean superior, just more expensive). Thus, parents can choose higher 'quality' of a lower quantity of children for a given level of overall expenditure. It is assumed that increased income will be diverted more towards

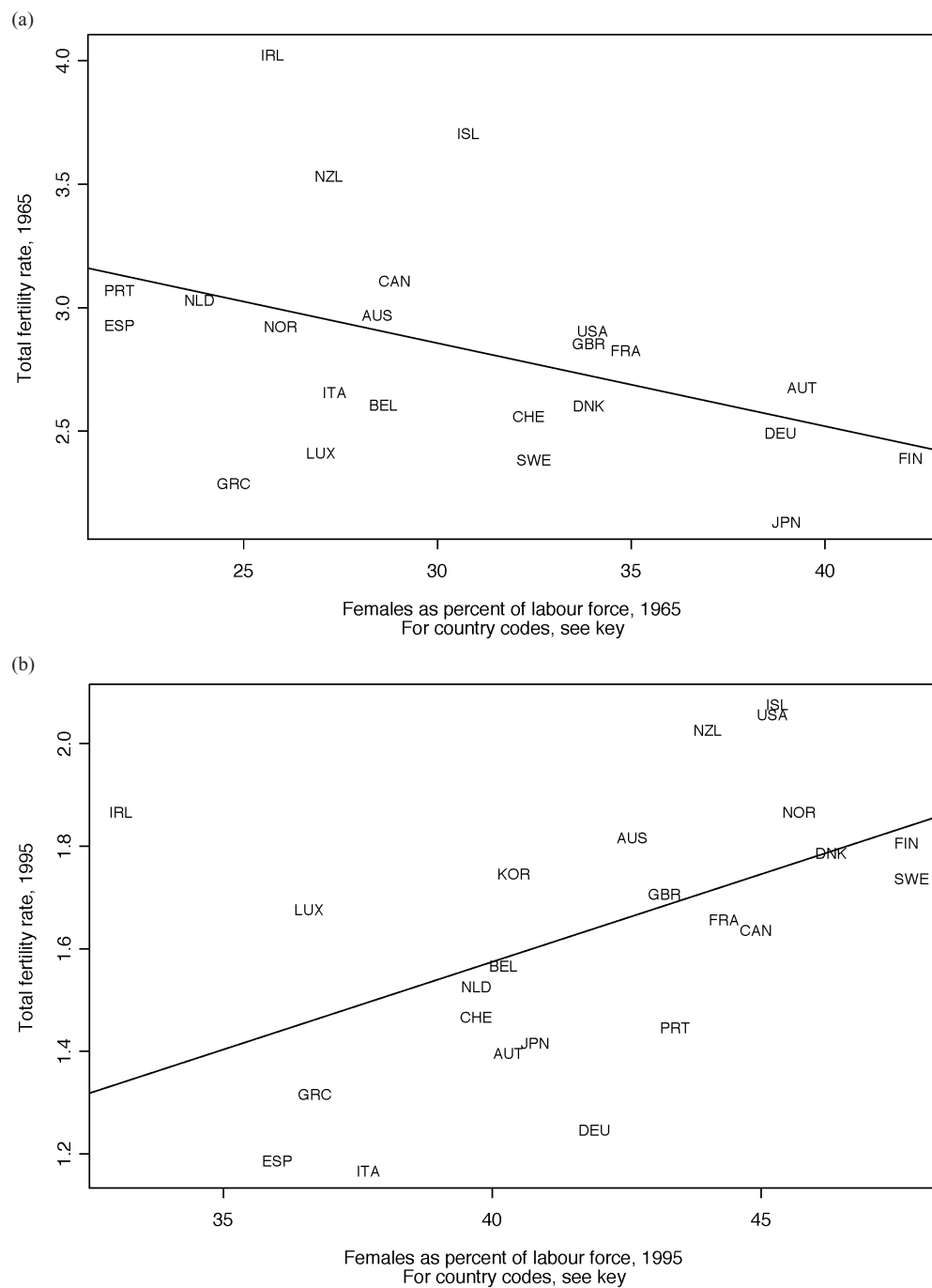


Figure 3
Total fertility rate vs. females as a percentage of the labor force

increases in quality rather than quantity, and it is possible that if the shift towards higher quality is so substantial, increased income may result in fewer

children. At present, the cross-national relationship between average income and fertility is positive but weak in developed countries (Fig. 2).

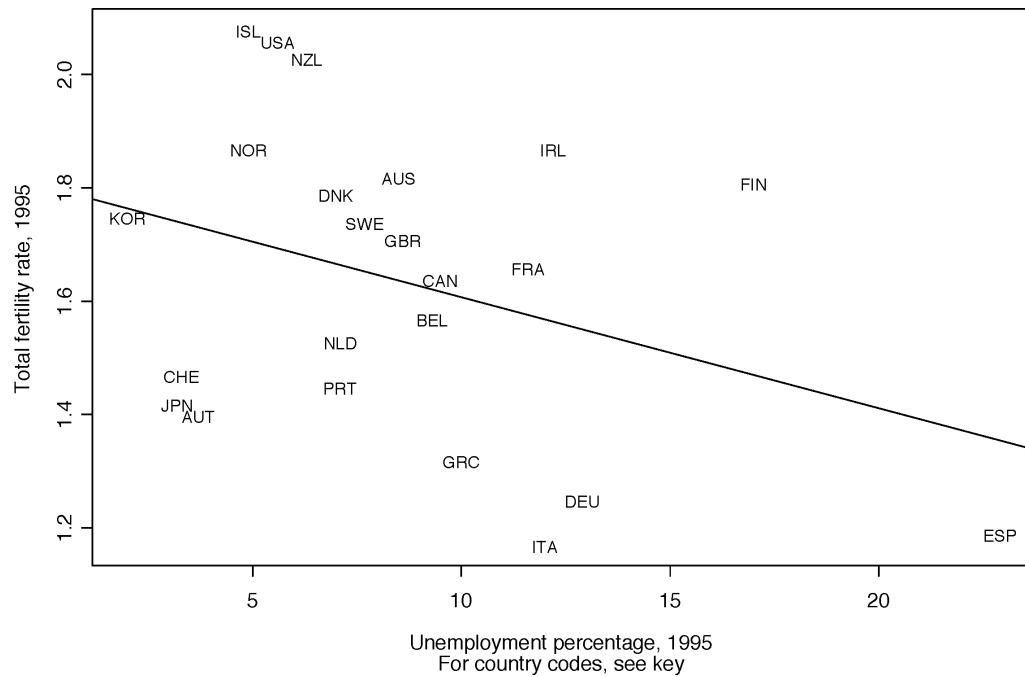


Figure 4
Total fertility rate vs. unemployment percentage, 1995

The relationship between female employment and fertility is expected to be more clearly negative since a basic incompatibility between work outside the home and childrearing is assumed as they are competing demands for the mother's fixed amount of available time. (This approach has been criticized on the grounds that it emphasises the differentiation of gender roles: however, since it reflects how things are, rather than how many think that things ought to be, such criticisms are unfair.) While in the 1960s, there was the expected negative relationship between female employment and fertility across countries, by the 1990s, the relationship had become positive (Fig. 3). In the 1990s, Scandinavian countries have relatively high fertility and female employment, while Mediterranean ones score low on both counts, whereas although the female employment rankings have remained relatively constant since the 1960s, the fertility ones have reversed. Of course, such simple correlations do not establish that the nature of the relationship has switched over time, but it is clear that the other factors that cause the correlation to be in the unexpected direction in the 1990s must also have a very substantial, if not dominant, effect on childbearing.

An alternative economic-demographic model is that of Easterlin (1980). In simplified form, it assumes that individuals' (or couples') expectations are set by the standard of living in the home that they had as children. Those born into large birth cohorts will have

many direct competitors for education, employment, housing, etc., and are likely to be relatively disadvantaged compared with those from smaller cohorts and they will react by marrying later and having fewer children. There is little empirical evidence that large cohorts are disadvantaged and earnings are more likely to be associated with other period and structural factors. This model emphasises men's as well as women's employment situations, but because it appeared to have little explanatory power, the role of male employment has tended to be neglected compared to that of women. However, the declining position of men relative to women in the labor market, and, in particular, the rise in unemployment that particularly affected young men is likely to inhibit leaving home and partnership formation, as well as making young men less attractive lifetime partners to young women (Oppenheimer 1994). Such considerations may be especially important among disadvantaged groups such as young black males in developed countries. Overall, there is a negative relationship between unemployment and fertility in the OECD countries (Fig. 4). Although the unemployed are disproportionately drawn from groups that would in any case be expected to start childbearing early and for this reason, among others, may have higher fertility than those in employment, a high level of unemployment is likely to have the effect of inhibiting partnership formation and childbearing of those in employment.

1.3 'Cultural' Explanations

In a study of the determinants of fertility decline in European historical and contemporary Third World countries, Cleland and Wilson (1987) concluded that 'economic' theories are much less successful than 'cultural' explanations. For example, fertility change was more closely associated with factors such as language and voting patterns than with income or employment (Coale and Watkins 1986).

Attitudes to increased sex role egalitarianism, and permissiveness to divorce and cohabitation have become more positive over time, whereas attitudes to religion, traditional family orientation, and deference to authority have declined. These changes are associated with a move towards greater individualism that has been associated with a range of emerging patterns of family formation and dissolution often referred to as a 'Second Demographic Transition,' based on individual autonomy, female emancipation, and increased consumerism (van de Kaa 1987, Lesthaeghe 1995). Such approaches integrate the interaction of different life events such as partnership, childbearing, and employment within the historical context, both national and individual, including cohort effects.

2. Assessing the Empirical Validity of Alternative Explanations

While swings in annual levels of childbearing have attracted much comment, especially the very sharp and largely unanticipated change between the late 1950s and 1970s, from a long-term perspective, it is the baby-boom period that stands out as the anomaly as fertility has shown a generally declining trend for over a century in many developed countries. In the 1950s the main alternative to remaining at home with one's parents was early marriage (sometimes precipitated by pregnancy), followed by a relatively quick start to childbearing, both because this was expected and the possibilities of delaying births were limited. The relatively high level of fertility at that period reflects the selection by women of a package of husband/home/children rather than solitude/parental residence/work. In addition, couples expected that the marriage would last and that employment would be permanent, so increasing the attractiveness of marriage. More recently, people have had the opportunity of childbearing without partnership and vice versa, and other options such as living alone have become more feasible. Therefore, the nature of the choices made with respect to childbearing have changed substantially over time, which complicates the specification of appropriate empirical models.

There are many levels at which analysis and interpretation may be undertaken: international comparisons; subpopulations within nations; families/households; individuals; and, most generally, combin-

ations of data from different levels. Childbearing is clearly sensitive to short-term factors such as pill scares, which may therefore provide good short-term explanations of fertility change (Murphy 1993). However, establishing the reasons for the long-term changes shown in Fig. 1(a) is more difficult. Some of the most widely cited studies for empirical underpinning of the theory have been based on the macrolevel time-series studies of the effects of income on fertility in developed societies. The problem in using aggregate-level data to draw conclusions about individual-level behavior is widely recognized.

A rational choice model underpins most discussion of childbearing, although this is often less helpful than might be thought since the number of factors that enter into the choice are very broad in both time and space. Childbearing is likely to have substantial effects on the emotional, lifestyle, and financial circumstance of the mother for at least two decades, and for most fathers too. Over time, the balance of advantage will have changed considerably. Some of these aspects may be much easier to measure than others. For example, higher social benefits for unmarried mothers which are relatively easy to measure, would be expected to lead to more births outside marriage, but a more important factor could be a reduction in public disapproval for such childbearing, and the availability of benefits a consequence of changing public attitudes, which may be difficult or impossible to quantify especially if no contemporary data exist.

It may be useful to make a distinction between those factors that have an unambiguous effect on childbearing and those for which the effect is likely to be context-specific. For example, it has been assumed that the effect of women's work outside the home is to reduce their fertility, as seen in the 1960s in Fig. 3(b). However, by the 1990s, the reverse was the case in that those countries with high rates of female employment had higher fertility also. The range of reasons for this could include:

- (a) formerly it was considered inappropriate for women with young children to be in paid employment, whereas now it is not considered harmful to the child (social influence);

- (b) to bring up a child, two incomes are now required whereas one would suffice in the past (reversed female income effect);

- (c) improved childcare facilities and technology makes it more feasible to combine childrearing and work than in the past (institutional and technological change);

- (d) withdrawal from the labor force for childrearing could have harmful long-term economic and career prospects for women, especially in societies with high rates of marital breakdown (external demographic effect);

- (e) societies that have low female employment have high unemployment and poor job security that inhibits family formation (historical macroeconomic factor);

(f) societies that have exhibited substantial fertility declines recently may reflect reduced fertility by older women who have sufficient children and postpone by younger women (transitory effect).

This list is not exhaustive but serves to emphasise the difficulty in establishing empirically the nature of the relationship between childbearing and the single variable of female employment. The statement that if increases in paid work outside them by women of childbearing age makes bringing up children less attractive as an option, with other things being constant, then childbearing will fall is uncontroversial. However, these other relevant factors that are not constant may be moving and be the dominant determinants of change. The dependent variable in the new home economics model is not number of children born, but it is the benefits obtained from their children, yet this cannot be clearly specified and so number of children is often used as the variable to be explained. The biological, cultural, institutional, economic, and historical factors that influence childbearing decisions are complex and it is impossible to obtain comparable data on these different domains, so in practice, it is almost impossible to specify a comprehensive formal model to 'explain' contemporary patterns. It is, therefore, unsurprising that after many decades of investigation, although a number of valuable insights have been made about the determinants of childbearing trends and differentials between subpopulations, full explanations remain elusive.

See also: Demographic Transition, Second; Families and Households, Behavioral Demography of; Families and Households, Formal Demography of; Family as Institution; Family Processes; Family Size Preferences; Family Theory: Competing Perspectives in Social Demography; Family Theory: Complementarity of Economic and Social Explanations; Family Theory: Economics of Childbearing; Fertility Control: Overview; Fertility: Proximate Determinants; Lone Mothers in Affluent Nations

Bibliography

- Becker G S 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA
- Bongaarts J, Potter R G 1983 *Fertility, Biology and Behavior*. Academic Press, New York
- Cleland J, Wilson C 1987 Demand theories of the fertility transition: An iconoclastic view. *Population Studies - J. Demog.* **41**: 5-30
- Coale A J, Watkins S C (eds.) 1986 *The Decline of Fertility in Europe*. Princeton University Press, Princeton, NJ
- Coleman D (ed.) 1996 *Europe's Population in the 1990s*. Oxford University Press, Oxford, UK
- Easterlin R A 1980 *Birth and Fortune: The Impact of Numbers on Personal Welfare*. Basic Books, New York
- van de Kaa D J 1987 *Europe's Second Demographic Transition*. *Population Bulletin* **42**: 1

- Lesthaeghe R 1995 The second demographic transition in western countries: An interpretation. In: Oppenheimer Mason K, Jensen A-M (eds.) *Gender and Family Change in Developed Societies*. Clarendon Press, Oxford, UK, pp. 17-62
- Murphy M 1993 The contraceptive pill and women's employment as factors in fertility change in Britain 1963-1980: A challenge to the conventional view. *Population Studies - J. Demog.* **47**: 221-43
- Oppenheimer V K 1994 Women's rising employment and the future of the family in industrial societies. *Population and Development Review* **20**: 293-342
- Ryder N B 1980 Components of temporal variations in American fertility. In: Hiorns R W (ed.) *Demographic Patterns in Developed Societies*. Taylor & Francis, London, pp. 15-54
- Westoff C F, Ryder N B 1977 *The Contraceptive Revolution*. Princeton University Press, Princeton, NJ

M. Murphy

Family Theory: Competing Perspectives in Social Demography

Two competing perspectives on the effect of postwar changes in men and women's economic behavior on the family have characterized recent demographic research. Each primarily focuses on one of two fundamental theoretical pillars which underlay most demographic research on families in Western societies: (a) the economic implications for family behavior of an independent nuclear family system; and (b) the importance of sex-role specialization in marriage. This article will undertake an examination of each of these perspectives in terms of these theoretical underpinnings, how they are affected by changing economic behavior, and the implications of this for the emerging nature of the family.

1. Overview

A number of major changes in marriage and family behavior have occurred in the West since the early 1970s. Age at marriage and marital instability have risen considerably while fertility has declined substantially (Blossfeld 1995, Mason and Jensen 1995, Oppenheimer 1994). The magnitude of these changes has fostered considerable debate about their determinants and consequences and what these imply for the future of the family. However, in most cases, the debate is not so much about major theoretical differences on the nature of the family but differences in which aspects of two long and commonly accepted underpinnings should be emphasized the most. One is that marriage as an institution, as well as marriages as

functioning social forms, is based on sex-role specialization. Although there are variations in how this specialization argument is conceptualized, the general idea is that the husband's role is largely economic while the wife's lies in the care of the home and in the reproduction and socialization of the children.

The second theoretical underpinning to most discussions is that an independent nuclear family system is a long-standing feature of populations of Western European origin. Such a family system is characterized by husband-wife families which are economically independent and live separately in their own household, although a single aged parent might also have sometimes been included in the past. This is opposed to an extended family system which is usually characterized by joint family households consisting of several nuclear families. In general, historical demographic research has indicated that marriage and fertility behavior have often been quite different in populations with independent nuclear as opposed to joint family systems. In the former case marriage is contingent on the couple's ability to be self-supporting and able to set up a separate household; in the latter, the economic ability of the household as a whole to absorb new members has tended to be more important. Since young people's economic position is often precarious, the requirement of economic independence has historically led to somewhat later marriages and also to periodic fluctuations in marriage timing, depending on variations in economic circumstances. This has often resulted in substantial short- and long-term effects on fertility, especially in periods before the deliberate control of marital fertility became common.

This article focuses on two opposing positions regarding the effect of shifts in the economic behavior of men and women. One emphasizes sex-role specialization, arguing that the rising employment of women has been the major factor leading to the observed changes in marriage and family behavior because it has reduced specialization and hence permitted women to become economically independent of marriage, thereby reducing the gain to marriage. Assuming women's employment is unlikely to decrease, the result will be a major and irreversible transformation of the family in modern industrialized societies.

The opposing position is that the family has not been transformed as much as the women's independence argument maintains nor has women's rising employment had a negative effect on marriage formation, although it undoubtedly mitigates against high fertility. On the other hand, there is some evidence that a deterioration in young men's economic position has had an important effect on marriage and family behavior because it has seriously affected the ability of young couples to set up an independent household. If so, then the assumed irreversibility of present trends is open to question, although it is extremely unlikely that a return to postwar patterns of the 1950s to 1970s will occur.

These two alternatives do not exhaust the many possible reasons that have been proposed for changing family behavior, although they cover the major debates related to the effect of changes in economic behavior. Nevertheless, it is often the case that different explanations may be implicitly, if not explicitly, dependent on one or both of these positions. For example it has been argued that modernization and the rise of individualism is the major force behind the recent demographic changes (see *Demographic Transition, Second*). However, neither men nor women can be autonomous individualists if they are still economically dependent on traditional marital and kinship ties.

2. Independent Nuclear Family System

Given the assumption of sex-role specialization, demographic research on marriage and family behavior in independent nuclear family systems in preindustrial and early industrial populations has tended to focus primarily on the economic position of men, although there has long been a general recognition that women could help facilitate the formation of a marriage via earnings before marriage. In farming communities, marriage was often considerably delayed as the heirs had to wait for the father to die or retire before they inherited the property and with it the means of marrying. Even where leasing rather than owning land was the rule, as in England, there were often extensive delays until a man could save enough to lease a farm and set up a household. The effect of this pattern on marriage timing—and hence on fertility in historically noncontracepting populations—was often substantial with high proportions not marrying until their thirties and relatively substantial proportions never marrying (see *Family Systems in Europe*; Arensberg and Kimball 1968, Goldstone 1986).

With the growth of wage labor in the cities and on the farms as industrialization progressed, young men could much more easily become economically independent at a relatively young age. This promoted a reduction in the age at marriage, although offsetting factors were the need to save up to establish an independent household and prior obligations to contribute to the parental household, especially if one or both parents had died or were incapacitated in some way. In general, economic development also increased the sensitivity of young men's economic position to business cycles and structural changes in the economy. Recessions or depressions reduced marriage propensities, especially among the young, and recoveries led to sharp rises in rates. Moreover, the structural changes in the economy associated with economic development changed the skill levels and types of skills demanded of workers. This has placed an increasing premium on education and technical training in modern industrial

societies which has, in turn, postponed the full-time entry of young people into adult-level employment and, with it, entry into marriage and the achievement of the degree of economic stability usually considered necessary for starting to raise a family.

2.1 Role of Relative Cohort Size

Postwar analyses of the influence of men's economic position on marriage formation and fertility was initially dominated by the work of Richard Easterlin who hypothesized that fluctuations in relative cohort size had a major impact on young men's labor market position, thereby affecting marriage timing and marital fertility (see *Baby Booms and Baby Busts in the Twentieth Century*; Easterlin 1987). Thus, Easterlin argued that the postwar baby boom, such as the US experienced in the early postwar period, was largely caused by the favorable economic position of the small depression cohorts entering the labor market at that time. This encouraged them to marry earlier and have more children than previous cohorts. However, the relatively large cohorts they produced crowded the labor market, negatively affecting their lifetime earnings. This led to postponed marriage and childbearing, lower completed fertility, and greater marital instability. In addition, more wives would be working to compensate for their husband's poorer economic position. The result was smaller birth cohorts in the next generation—the baby bust cohorts—who would be in a much more favorable relative labor-market position. According to the theory, this, in turn, should lead once again to earlier marriage and childbearing, somewhat higher fertility, increasing marital stability, and the slowing of the growth in married women's employment. In short, the theory posited a self-perpetuating cycle.

The relative cohort size argument ran into empirical difficulties as research in the 1980s showed that postwar marriage and fertility fluctuations were largely *period* rather than *cohort* driven, i.e., fluctuations in demographic rates occurred simultaneously for numerous birth cohorts, regardless of cohort size. Moreover, the theory essentially accepted the premise that marriage is based on sex-role specialization. Women's labor-market behavior was considered to vary primarily in response to men's earnings position rather than in response to women's own employment opportunities, independent career goals, or family cycle variations in economic needs. However, women's employment continued to rise rapidly throughout the postwar period, regardless of variations in cohort size. As a result, the focus of research interest began to shift to the familial consequences of the transformation of women's economic roles. Moreover, since the concern with the economic position of young men had become so closely identified with the relative cohort size argument, interest in men's changing economic pos-

ition flagged for a number of years while the research focus shifted to analyzing the effect of women's rising employment.

3. Sex-role Specialization

At the same time that relative cohort size and hence male-oriented explanations of changes in marriage and family behavior began to lose favor, the continuing rapid rise of married women's employment has seemed to provide a ready explanation for the ongoing changes in marriage and family behavior. Hence, the theoretical emphasis shifted from a concern with the ability to set up and maintain an economically independent household to the critical importance of sex-role specialization for family formation and cohesion.

While it has only recently taken center stage as an explanation of changes in marriage and family behavior, a long-standing and influential tradition in the social science literature is the importance of differentiated sex roles for a stable marriage system. In the early postwar period this idea was emphasized by Talcott Parsons (1949) who argued that sex-role segregation was a functional necessity for marital stability and even for the viability of society itself. More recently, this same theme has been elaborated somewhat differently in the economic theory of marriage by Gary Becker (see *Family Theory: Economics of Marriage and Divorce*). In an argument very similar to one Durkheim made over a century ago, Becker maintains that the major gain to marriage lies in the mutual dependence of spouses, each of whom specializes in certain functions—the woman in domestic production (and reproduction), the man in market work. Marriage then involves trading the fruits of these different skills. However, in response to economic growth and the rising wages it produces, women's market work also rises. The result is that women become less specialized and, as their earnings approach those of potential mates, more economically independent, leading, in turn, to a decline in the desirability of marrying or of staying married as well as the desirability of high fertility.

Not all scholars necessarily agree with Becker's argument in its entirety. Nevertheless, an economic independence argument of one sort or another has had wide appeal among sociologists as well as economists and is currently one of the major contenders in trying to explain recent marriage and family trends in the United States (Cherlin 1992, Oppenheimer 1997). One reason is that the model seems to have tremendous face validity. Not only does the notion of specialized sex roles fit in with long-held views of how the traditional family historically functioned, but the rapid changes in marital behavior appeared to have followed very closely upon the rapid rise in married women's employment, especially that of young

women. Moreover, the theory has the elegance of simplicity, yet it can apparently explain a wide variety of complex changes and differentials in marriage and family behavior—from delayed marriage to non-marriage, marital instability, nonmarital cohabitation, female-headed households, declining fertility, and so forth. Nevertheless, the independence hypothesis has also run into a number of serious theoretical and empirical problems.

3.1 Mismatches

One problem the theory has encountered is that there are number of mismatches between the phenomena it claims to explain and the behavioral changes that are actually occurring. First, the independence hypothesis is essentially a theory of *nonmarriage* for it argues that if, by their own endeavors, women can achieve approximately the same income as a prospective spouse there is not much to gain by marrying and specializing in home production. However, so far, while there has been a considerable increase in *delayed* marriage, as yet there is little evidence of a substantial rise in *nonmarriage* (Oppenheimer 1994, Blossfeld 1995). For example, in the United States although white women reaching age 20–24 in 1980 were more likely to be single than earlier cohorts at that age, by the time they were 35–39 in 1995 just under 10 percent were still never married, below that for late nineteenth century cohorts who were presumably living in a ‘traditional’ family system. Hence one reason for the apparent support for the hypothesis is due to a serious confounding of *delayed* marriage with *nonmarriage* in the statistics. Moreover, to date, there has also been little effort to articulate how the independence argument could lead to *delayed* instead of *nonmarriage*.

A major difficulty in assessing the outcome of recent marriage trends lies in the recency of the trends and the necessity of waiting 20 years or more before one can accurately determine whether there actually will be a significant rise in nonmarriage. Efforts to estimate the future proportions never marrying are often misleading because they typically use the current behavior of older people as a guide. However, during a period of rapid timing shifts, the *current* behavior of somewhat older people, who are the relatively sparse and atypical remains of much earlier marrying cohorts, is often a particularly poor guide to the future behavior of the much larger and less select group of younger people who have been delaying marriage in more recent years.

A second kind of mismatch lies in using the family behavior of the 1950s and 1960s in the US (and somewhat later in Europe) as typifying the traditional family patterns when sex-role specialization prevailed; compared to this, the large subsequent changes in marriage and family behavior certainly appear to represent a major retreat from marriage and the family. However, this was the baby boom era when

fertility was unusually high compared to the previous 50 years; age at marriage was unusually low—far lower than at the turn of the century—and, in the US at least, even marital disruption was somewhat less likely than in previous years. In short, this benchmark was highly atypical historically and unrepresentative of the supposedly ‘traditional’ family. Hence using it as a benchmark has led to a considerable exaggeration of the extensiveness of the subsequent break with historically traditional patterns and of the historical origins of many recent trends.

There is still another type of mismatch characteristic of a theory which posits that the stability of the family is founded on women specializing in home production. The proposed model seems strangely out of tune with the needs and characteristics of modern societies. Much of the specialized home production of women in the past was devoted to bearing and rearing children who never survived to adulthood. For women to be equally occupied in contemporary low-mortality societies would result in large families. However, even moderate family sizes in a low-mortality society lead to rapid population growth. Moreover, couples do not just want to produce children *per se*, they want to produce children like themselves—that is, they are interested in social, not just biological reproduction. But the cost of social reproduction is high in a society where increasingly substantial and lengthy investments in human capital for each child are required. Hence, high fertility is usually not a realistic option for most people in modern industrial societies. With low fertility, however, the female specialist would not be doing anything highly productive for much of her life. In short, specialization in home production appears to be a rather anachronistic institutional role for women in industrial societies.

3.2 Empirical Status of the Independence Hypothesis

Much of the empirical work on the women’s independence hypothesis has not supported it. At any given time, empirical analyses show that indicators of women’s economic independence have not generally had a negative effect on marriage formation. For example, under an independence argument, better-educated women should be more economically independent of marriage; however, microlevel regression analyses show that, once school enrollment is taken into account, they have a *higher* rather than a *lower* propensity to marry.

The empirical work on marital dissolution is more mixed. Some studies have found evidence of an independence effect, whereas others have not. There are a number of possible reasons for these conflicting findings. A major one is that the measurement of women’s economic independence is difficult and inconsistent across studies. Some studies use absolute

income while others use some measure of relative income. While the hypothesis implies that the *relative* earnings of wives should be the important factor and hence should be directly modeled, using the ratio of wives' to husbands' income alone can lead to ambiguous results. The wife's earnings in some families may be higher than in others, reflecting a better labor-market position and, consequently, indicating a greater financial independence of her husband. However, the ratio could be higher because the husband's earnings are low compared to other couples, while the wife herself may also be in a very weak labor-market position.

A second problem that is more likely to occur in analyses of marital dissolution than marital formation is the greater difficulty of determining the appropriate causal direction. Wives who are more extensively employed in one year may be more likely to separate the following year—but not necessarily because they are more economically independent. Instead, their marriage may have already been in trouble, causing them to increase their labor-market involvement in anticipation of a marital breakup—the reverse causal direction than that hypothesized.

3.3 Coping with Risk and Change

Although rarely considered in explications of the theory, historical research on independent nuclear families indicates that extreme sex-role specialization in marriage is essentially a high-risk and inflexible family strategy unless accompanied by supplementary support mechanisms (Oppenheimer 1994). Even with such supports, specialization often entailed considerable individual and social costs. An inherent problem is that the temporary or permanent loss of one specialist in a family can mean that functions vital to the well-being of the complementary specialist and children are not being performed. Husbands/fathers can die or become ill or disabled; they can lose their jobs and have difficulty finding another one; they could desert the family for a variety of reasons or become an alcoholic; and so on. The result is that the family is left without its major source of income. Except for employment-related shifts, there are similar problems involving the wife-mother specialist. In that case, there could be no one to take care of the children or the home. Specialization may be a feasible strategy in a large extended family household where no particular individual is indispensable because of the redundancy in personnel that can characterize such a system. However, for independent nuclear families and their individual members, specialization entails considerable risks.

Extreme sex-role specialization is also not a very flexible way of dealing with the varying needs of nuclear families over their developmental cycle. Since individuals' consumption needs and productive capa-

bilities vary markedly by age, a basic feature of nuclear families is that the ratio of consumers to producers, and hence the family's level of living, can vary substantially over the family's developmental cycle. Hence, specialization involves a potentially serious inflexibility in dealing with changes in both a family's internal composition as well as the stresses posed by its environment.

The large literature on family history in Western societies indicates that a variety of strategies were developed to maintain economic stability over the family's developmental cycle and in the event of the temporary or permanent loss of specializing parents (Goldin 1981, Haines 1979, Rowntree 1922, see Oppenheimer 1994 for a review). In the nineteenth and early twentieth centuries these strategies usually involved utilizing the productive labor of children, daughters as well as sons. The evidence also indicates that utilizing the labor of one's children could exact substantial costs. A sufficient number of children old enough to make an economic difference was generally not available until the middle or later stages of the family cycle. Families who temporarily or permanently lost the contribution of the father early in the family cycle were not greatly helped by such a strategy. If the mother was lost when the children were young, the family might break up with children being parceled out among relatives or even going to orphanages. In general, this economic reliance on one's offspring often led to a pattern of 'life cycle poverty' where periods of poverty and comparative plenty alternated over the life cycle of workers. Another well-known disadvantage to the extensive employment of children to supplement their family's income was that it tended to discourage schooling and hence had a negative effect on children's adult socioeconomic status.

Even aside from the drawbacks of using children's work, the economic advantages of this strategy were eventually bound to decline during industrialization as the adult labor force became more skilled and hence the relative economic contribution unskilled children could make declined. This suggests that other equilibrating mechanisms were likely to develop. One mechanism may be governmental social support systems. At the family level, however, from a historical perspective, the rise in married women's employment might be viewed as a functional substitute for the work of their children, facilitating the more extensive schooling of the next generation and thereby fostering upward intergenerational social mobility. In addition, wives' employment is not limited to the later stages of the family's developmental cycle and wives, being adults, and usually having educational attainments roughly similar to their husbands, can command a much higher wage than unskilled children could. Moreover, recent research indicates that wives' earnings are currently playing an important role in offsetting the less favorable earning position of their husbands (Levy 1998). It seems likely, then, that wives'

employment could be a source of cohesion in marriages; in difficult times, a helpmate may be of much greater value than a complete economic dependent.

Even aside from wives employment during episodic periods of an increased need for additional income, wives who work regularly may also increase the interdependence of spouses. Two-earner families where the couple pool their resources are families whose level of living is based on their *combined* income. Even when the income contribution of each spouse is equal, given economies of scale, neither partner could live as well on his/her own or save and invest as much. Hence, instead of the asymmetric dependence of the wife which is characteristic of complete specialization, there is a symmetric dependence where each spouse depends on the other to maintain living standards. What this may also indicate is that the combined income of the two-earner family has come to form the social standard rather than the husband's income alone. To the extent this is the case, it becomes increasingly difficult for single earners and married couples with a more traditional division of labor to achieve the same level of living as the two-earner family. Hence, the mutual dependence of the two-earner family may not only contribute to their own gain to marriage but may also reduce the relative gain to being single and to marriages characterized by a specialized division of labor.

In sum, it is unclear that specialization is the only, or even the optimum, method of creating cohesion in marital relationships in modern industrial societies. The specialization model itself seems highly problematic and women's rising employment may actually be promoting a more collaborative model of marriage.

4. Rising Economic Inequality

Given the empirical and theoretical limitations of the specialization model of marriage, research attention has been returning once again to the role of economic difficulties in setting up and maintaining an independent nuclear family. While the effect of relative cohort size earlier dominated research in this area, the more recent focus is on such issues as changing economic structure, chronically high unemployment rates (particularly among young adults), and rising earnings inequality (again especially among the young but also among the moderately-to-less educated as well). These analyses have also returned to the question of men's changing economic position which had been neglected for quite some time. Research in this area is still in its early stages but some of the major questions to be addressed are: (a) whether, how, and why young men's labor-market position affects cohabitation and marriage and what difference socioeconomic differentials make; (b) and whether economic changes in a society are likely to have had an

important impact on young people's economic position and have thereby influenced changes in their marital behavior.

Virtually all recent research has shown that men who are more likely to be working full-time/year-round, had higher current earnings, and/or greater future career prospects were more likely to marry in any given short time interval (Clarkberg 1999, Huinink 1995, Oppenheimer 1997, Oppenheimer et al. 1997). Some research in this area has also started to look at the role of economic factors in the formation of cohabiting unions and this indicates that here too, some level of economic stability is important, although apparently not as much as for marriage formation (Bracher and Santow 1998, Clarkberg 1999). Oppenheimer (1988) has theoretically and empirically elaborated on these issues in a variety of ways. First, she has conceptualized the transition to work as a career-entry *process* rather than a *step*, which was more commonly the approach in previous life-course analyses. This change in approach emphasizes that embarking on a relatively stable occupational career takes time, during which a young man is in a relatively 'immature' career state. One aspect of career immaturity directly impeding the formation of an independent household is relatively low income; however, another is uncertainties regarding what kind of stable employment may eventually be achieved, when this may occur, what lifestyle the individual may ultimately expect or desire, and so forth. Such uncertainties can impede mate selection and affect the individual's own desirability as a prospective marriage partner. Hence, career immaturity is not only likely to foster delayed marriage for economic reasons but for social-psychological ones as well. Second, viewing the transition to work as a *process* expands the emphasis beyond the effect of career immaturity at any single point in time to the length of time it takes to achieve some career stability and the cumulative impact this will have on marriage timing. The delaying effect of only a modest effect of career instability in any given year, for example, may nevertheless have a large cumulative impact if the instability persists over many years. Moreover, this approach is well suited for examining how the length and difficulty of career development could vary substantially among different groups and over time as well, thereby facilitating analyses of socioeconomic differentials in marriage timing and trends in marriage behavior.

Using 12 years of an American longitudinal survey of young people, Oppenheimer and her colleagues (1997) found that not only did a variety of indicators of career immaturity have a substantial effect in any given year on marriage formation, but that the cumulative impact was substantial when the career-development process was followed over time. Moreover, there were large race-schooling differences in the length and difficulty of the career-entry process which, in turn, produced major differences in marriage timing.

Because the analysis was limited to the young adult life course of young men born between 1957 and 1964, how much of the rise in delayed marriage over the past 25 years can be accounted for by career-development problems has yet to be directly examined. However, there has been a substantial and well-documented rise in inequality in the US since the 1970s, and this has had a particularly severe effect on moderately-to-less educated younger workers, suggesting that career-entry difficulties may be playing an important role in these trends. In Western Europe there has been a substantial increase in persistent unemployment, especially among the young which also indicates that career-entry difficulties may be important in changing European marital behavior (Levy 1998).

See also: Families and Households, Behavioral Demography of; Family Theory and the Realities of Child-bearing Behavior; Family Theory: Complementarity of Economic and Social Explanations; Family Theory: Economics of Marriage and Divorce; Family Theory: Role of Changing Values; Marriage and the Dual-career Family: Cultural Concerns; Reproductive Rights in Affluent Nations

Bibliography

- Arensberg C M, Kimball S T 1968 *Family and Community in Ireland*. Harvard University Press, Cambridge, MA
- Blossfeld H-P (ed.) 1995 *The New Role of Women: Family Formation in Modern Societies*. Westview Press, Boulder, CO
- Bracher M, Santow G 1998 Economic independence and union formation in Sweden. *Population Studies – Journal of Demography* **52**: 275–94
- Cherlin A J 1992 *Marriage, Divorce, and Remarriage*. Harvard University Press, Cambridge, MA
- Clarkberg M 1999 The price of partnering: the role of economic well-being in young adults' first union experiences. *Social Forces* **77**: 945–68
- Easterlin R A 1987 *Birth and Fortune*. University of Chicago Press, Chicago
- Goldin C 1981 Family strategies and the family economy in the late nineteenth century: the role of secondary workers. In: Hershberg T (ed.) *Work, Space, Family and Group Experience in the Nineteenth Century*. Oxford University Press, New York, pp. 277–310
- Goldstone J A 1986 The demographic revolution in England: a re-examination. *Population Studies – Journal of Demography* **49**: 5–33
- Haines M 1979 Industrial work and the family life cycle, 1889–1890. *Research in Economic History* **4**: 449–95
- Huinink J 1995 Education, work, and family patterns of men: the case of West German. In: Blossfeld H-P (ed.) *The New Role of Women: Family Formation in Modern Societies*. Westview Press, Boulder, CO, pp. 247–62
- Levy F 1998 *The New Dollars and Dreams: American Incomes and Economic Change*. Russell Sage, New York
- Mason K O, Jensen A-M 1995 *Gender and Family Change in Industrialized Countries*. Clarendon Press, Oxford, UK
- Oppenheimer V K 1988 A theory of marriage timing. *American Journal of Sociology* **94**: 563–91

- Oppenheimer V K 1994 Women's rising employment and the future of the family in industrial societies. *Population and Development Review* **20**: 293–342
- Oppenheimer V K 1997 Women's employment and the gain to marriage: the specialization and trading model. *Annual Review of Sociology* **23**: 431–53
- Oppenheimer V K, Kalmijn M, Lim N 1997 Men's career development and marriage timing during a period of rising inequality. *Demography* **34**: 311–30
- Parsons T 1949 The social structure of the family. In: Anshen R (ed.) *The Family: Its Function and Destiny*. Harper, New York, pp. 173–201
- Rowntree B S 1922 *Poverty: A Study of Town Life*. Longmans, Green, London

V. K. Oppenheimer

Family Theory: Complementarity of Economic and Social Explanations

1. The Problem

With the growing specialization in the social sciences the various disciplines have often developed their own traditions and paradigms in explaining social phenomena. There is, however, a risk that this may lead to segregated 'communities of belief' that advance only 'positive proof' in support of their own positions. This has, on occasion, led to the presentation of methodologically flawed research that is supposed to enhance the supremacy of one discipline over the other (see Burch 1996).

As far as family and population research is concerned, contributions are made regularly from a wide variety of disciplinary backgrounds. The phenomena that needed to be explained, such as the fertility or mortality transitions or the change in demographic behavior and household structures, have drawn the attention of sociologists, historians, economists, anthropologists, geographers, biologists, epidemiologists, and statisticians alike. As a consequence, the field of population studies—or demography in a broad sense—has always had firm interdisciplinary roots. The leading journals, such as *Population Studies* (UK), *Demography* (US), *Population and Development Review* (US), and *Population* (France) have remained loyal to their interdisciplinary tradition, and they are still firmly in the lead over their more 'mono-disciplinary' competitors.

Yet, there have been skirmishes and frustrations. The former have often been similar to the nature–nurture debate in psychology. But in the field of family and population research, they emerge in the form of an economics vs. culture debate. The demand theory used by economists in explaining declining fertility or differential fertility between social groups is matched by a diffusion theory, advanced by economists and noneconomists alike. In the latter, the roles of social

learning and the worldwide spreading of new ideas and attitudes are being accentuated. Today, there is still a small cluster of neoclassic economists that hold the belief that nothing more is needed on the explanatory side than changing constraints and opportunity structures (see Becker 1991, Stigler and Becker 1977), but at the other extreme, there is also a hard core of diffusionists who claim that historical and contemporary fertility transitions are occurring under such highly contrasting economic conditions that only the cultural diffusion of new ideas—now also followed by family planning programs—can account for them (e.g., Cleland and Wilson 1987). Obviously such discussions have major policy implications, and the debate has had a lasting life (see the contrasting results on the effects of family planning in Schultz 1997 and in Bongaarts et al. 1990).

In contrast, the frustrations among the mainstream participants in family and population studies stem from the observation that the different explanations live side by side as distinct narratives, as separate 'good stories.' Van de Kaa (1996) describes an anchored narrative by the following two characteristics: (a) there must be an easily identifiable central action (such as a detailed statistical account of changing trends or of newly emerging differentials) and (b) there must be a setting or theory that allows for an easy interpretation of that action. A well-anchored narrative consists of subnarratives that can be ordered from the general to the highly detailed. The question is whether the various social science disciplines can achieve a sufficiently high degree of sophistication that lifts them beyond anchored, but still compartmentalized narratives.

A classic recipe for theory integration emerged from the great debate in philosophy of science on the evolution of scientific knowledge as it took place in the late 1960s and much of the 1970s with Karl Popper and Thomas Kuhn as major protagonists (see Lakatos and Musgrave 1970). The gist of Lakatos's (1970) strategy of progressive problem shift is that a scientific theory H_1 can be regarded as falsified only if another theory H_2 has been proposed such that:

(a) The counter-evidence to H_1 is corroborating evidence for H_2 ;

(b) H_2 satisfactorily explains all the empirical successes hitherto explained by H_1 (i.e., partial inclusion of H_1 into H_2); and

(c) H_2 is also capable of explaining or predicting new facts or facts that were unlikely or impossible according to H_1 (i.e. enlarged content of H_2).

If these three principles are applied to the theories and anchored narratives in the broader field of demography, one comes to the following observations (see Lesthaeghe 1998):

(a) Many theories, or substantial parts of them, are not mutually exclusive but merely complementary: H_1 and H_2 explain different aspects or account for different subnarratives;

(b) Mechanisms described in H_1 are often connected to mechanisms identified by H_2 . Theories escape from these connections through *ceteris paribus* clauses with respect to different but nonetheless highly relevant omitted variables. If put together, however, the respective mechanisms in H_1 and H_2 often produce mutually reinforcing (or neutralizing) consequences or lead to important interaction effects; and

(c) Nonoverlapping parts of H_1 and H_2 that produce opposite predictions have rarely been falsified. Rather, their respective predictions have often come true in different contexts. Hence, a substantial degree of historical path dependency or context specificity emerges. This points in the direction of conditionality on factors not specified in either H_1 or H_2 but to be identified in a broader theory H_3 .

These points will now be illustrated with respect to the changes in family formation and dissolution in industrialized countries since the 1950s.

2. A Comparison of Three Theories of the 'Second Demographic Transition'

The term 'second demographic transition' was introduced by Lesthaeghe and van de Kaa (1986) to describe the changes in family formation, union dissolution, and patterns of family reconstruction in Western nations during the latter part of the twentieth century. By now, these patterns are also emerging in the former Eastern bloc countries of Europe including Russia. The changes in family formation operate through the postponement of marriage, a rise in single living, prolonged residence in the parental household, increases in incidence and duration of cohabitation (both pre- and postmarital), a postponement of parenthood leading to overall subreplacement fertility, and an increase in procreation within consensual unions. Patterns of union dissolution are characterized by rising divorce rates occurring at shorter durations of marriage and by high separation rates among cohabitants. These phenomena lead to more single person households and to more female-headed single parent households. More children grow up in non-conventional families. Although countries differ with respect to incidence and timing (e.g. Lesthaeghe 1995), most changes have consistent international trends.

These features of the narrative have been explained in three ways:

(a) by the theory of increased female economic autonomy combined with other ingredients of the neoclassic demand theory in economics (Becker et al. 1977, Becker 1991, Hotz et al. 1997, Bergstrom 1997);

(b) by the theory of relative economic deprivation (Easterlin 1976, Easterlin et al. 1990);

(c) by the theory of ideational shift (Preston 1986, Lesthaeghe and Meekers 1986, Thornton and Camburn 1987, Lesthaeghe and Surkyn 1988, Bumpass 1990).

The neoclassic economic theory of the family posits that (a) increasing investment in the quality (e.g., education) of children rather than in their numbers is being made as individuals and societies become wealthier, and (b) that increased female education and the opening up of employment opportunities for women has led to increased female economic autonomy and to increased opportunity costs associated with household tasks and child rearing. Services provided by the market are substituted for such tasks. Economic autonomy also means that women are less dependent on husbands, so that entry costs into marriage have become higher and exit costs lower. From these core propositions one can easily derive the changing position of men in households (sharing of tasks and child rearing activities, the quest for gender symmetry, weakening of male authority, etc.). By the same token, the rules of the game are being altered, especially with respect to intrafamilial exchange: less trust from family members in the family head is exchanged for less altruism from the family head toward the members. All of this leads to an economic disinstitutionalization of the family, and the narrative of the 'second demographic transition' describes the demographic outcomes produced by these forces.

The theory of relative economic deprivation provides a number of other key ingredients. The outcomes of this narrative essentially are produced by the tension between the consumption aspirations and employment opportunities of successive male generations. Easterlin originally used the theory to account for the baby boom of the 1960s and the baby bust of the 1970s and thereafter. The generations that produced the marriage and baby boom were raised in periods of less economic growth and had not developed high consumption aspirations during their childhood. These cohorts, however, benefited from expanding economic opportunities during the 1960s, and as a result they could more easily satisfy both consumption aspirations and family formation desires. Subsequent cohorts experienced high levels of consumption in their families of origin and developed high consumption aspirations themselves, but could not as easily fulfill these as a result of less advantageous labor market conditions. All the crucial steps in family formation were, therefore, postponed and new arrangements with a more temporary character, such as prolonged residence in the parental household or cohabitation, emerged as a necessity while waiting for better times. Furthermore, high consumption aspirations in tandem with scarcer opportunities for males fostered the need for an additional partial or full income to be provided by the female partner. From there onwards, the story could easily be completed according to the lines provided by the 'female economic autonomy' version of the neoclassic school.

Yet, both the starting point and the mechanism of the narratives are different. Becker et al. (1977) start from the female side, whereas Easterlin et al. (1990)

start from the male point of view. Moreover, the Easterlin version has the built-in mechanism of cohort succession and a link to values (concerning consumption) fostered during childhood and adolescence. In the original version, a demographic factor was also given greater prominence: cohort size would exacerbate the consumption aspirations-income tension since larger cohorts would face more competition and hence also tougher employment conditions. Higher female education and a larger female labor force supply (see the neoclassic version) would of course result in similar effects.

From the above narratives it is clear that the two economic theories H_1 and H_2 are not mutually exclusive and that their respective mechanisms have, to put it metaphorically, several cog wheels that grip into each others.

The theory of ideational shifts links the features of the second demographic transition to long-term trends, often starting in the nineteenth century, toward greater individual autonomy in the ethical, religious, and political domains. Typical expressions of this trend are the steady rise of secularism in the West over the last two centuries, the growth of emancipation movements operating first in the domain of social stratification, and then in the area of gender relations, the steady rise of postmaterialist political aspirations (see Inglehart 1990) in each successive birth cohort since 1920, the historical alteration of priorities in child socialization values (see Alwin 1990); or more recently the growth of scepticism toward institutions and the weakening of civil morality. In this respect, the baby boom in the original six member countries of the EU corresponded with the arrival at childbearing age of a cohort of parents that had hardly progressed on the postmaterialism scale with respect to political aspirations, whereas the baby bust corresponds with the childbearing of the cohort that made the largest leap of all on the same scale (Lesthaeghe and Surkyn 1988). Very much the same result emerges with respect to the relative pace of secularization in Europe, and also among American youths, a significant shift away from community orientations and toward a self-centered preoccupation with material riches was observed (Easterlin and Crimmins 1991). On the whole, the rise of secular individualism and the concomitantly diminishing acceptability of institutional regulation in the sphere of the family are part of the same trend. This also implies that individuals are allowed more freedom to opt out of any arrangement if returns are judged to be inadequate and hence, that the cost-benefit calculus, which features so prominently in the neoclassic economic version, has gained legitimacy in most spheres of life.

A distinct feature of the ideational shifts theory is its preoccupation with intergenerational values transmission, the predictive power of value orientations on subsequent choices with respect to patterns of family formation and dissolution (selection effects), and the

feed-back effects in the form of value adjustments (affirmation or negation) in function of certain life cycle events. This line of research has made extensive use of panel data in assessing the strength of these effects net of other predictors commonly used in microeconomic empirical analyses (see Thornton and Camburn 1987, Axinn and Thornton 1993, Clarkberg et al. 1995, Barber 1999, Moors 1999). These meticulously designed studies all show that value orientations, even when measured at young ages (including those related to consumption aspirations and labor force commitment), have highly significant independent effects on future life course choices such as cohabitation vs. marriage, parenthood within and outside wedlock, timing and quantum of fertility, separation and divorce. While not refuting predictions derived from economic theory, these studies document the extra predictive power of ideational dimensions.

Finally, it is clear that several ingredients of the ideational theory can be linked directly to those featured by the economic views. For instance, the neoclassic central focus on female economic autonomy leads directly to the changes in gender relations, and these indeed matter in the unfolding of alternative forms of family formation. In addition, both the Easterlin version and the ideational theory pay ample attention to the role of socialization in successive generations and both are essentially cohort theories of social change. Ideational theory, however, traces the trends in causal factors much further back in time, and it does not of course restrict itself to consumption aspirations alone. But in doing so, the ideational shifts theory provides support for the neoclassic economic position by showing how increased legitimacy of individual economic rationality is taking the lead over the forces of institutional normative regulation.

3. Predictions and Contexts: The Example of Cohabitation

The three theories just discussed offer different mechanisms for explaining the rise of premarital cohabitation in the West since the 1960s. According to the neoclassic theory, the phenomenon results from increased economic independence of women, and expects, therefore, that better educated women with better employment opportunities and higher incomes take the lead. In the relative economic deprivation version, cohabitation is a preliminary stage typically emerging prior to the full establishment of economic self-sufficiency of a new household. Cohabitation is, therefore, more typical for lower income groups with less employment stability, or for subgroups that have not yet matched their incomes to their consumption aspirations. In the ideational version, cohabitation is not so much predicted on the basis of economic constraints or opportunities, but it stems from a set of value orientations already emerging during the 'for-

mative years.' As these values are often linked to more advanced education, a positive correlation between education and cohabitation is expected. Furthermore, antecedents in the family of origin (such as parental divorce) also contribute to earlier home leaving and cohabitation (see, for example, Cherlin et al. 1995, Kiernan and Hobcraft 1997), *inter alia*, via the accentuation of values of ethical autonomy.

The neoclassic economic theory and the ideational theory both draw support from the initial positive correlation between education and cohabitation in a number of countries. At the microlevel, attitudes associated with career commitment and gender symmetry, also tend to be positively related to premarital cohabitation. However, in several contexts (e.g., Scandinavia, Austria), historical patterns of premarital cohabitation had survived in specific subgroups, and the diffusion of premarital cohabitation did not originate from the college educated or in large urban areas, but rather in the working class and more remote areas (e.g., northern Sweden, Alpine valleys). In addition, in most other Western countries, younger generations from working-class backgrounds and with less education adopted the pattern of cohabitation very quickly, so that the initial positive correlation with education vanished or even reversed. The predictive power of specific value orientations is also weakening as cohabitation has ceased to be an expression of protest among the highly unconventional (for these trends see Villeneuve-Gokalp 1990, Oppenheimer and Lew 1995, Boulanger et al. 1997, Bracher and Santow 1998, Kravdal 1999).

To sum up, cohabitation has become widespread among very different socioeconomic strata, suggesting (a) the existence of particular diffusion patterns in various national contexts, and (b) a rapid differentiation of motivations behind the choice. This form of family formation can be initiated for reasons suggested by all three theories. As a result, a particular theory may have higher or lower predictive power depending on the phase of the process of diffusion and on national or historical contexts.

4. Conclusion

The scientific traditions in the different social science disciplines leave an important imprint on both the theoretical conceptualization and the research agenda. Neoclassical economists have a strongly deductive approach and typically start with theory construction and the specification of the corresponding equations. Often they stop short of adequate operationalizations of basic concepts or fail to incorporate noneconomic variables in tandem with economic proxies (Robinson 1997). Demographers, sociologists, and statisticians use both deductive and inductive strategies, and they are engaged continuously in concept operationalizations and statistical testing or exploration. The latter

group is, therefore, more likely to report 'mixed results' than the former. Empirical scrutiny of associations derived from theory invariably points in the direction of theoretical incompleteness. As shown by the example above, the separate narratives are seldom mutually exclusive, they complement each other in more than one way, and they produce predictions that are restricted to context, time, or social group. All three theories are, therefore, candidates for inclusion into a larger theoretical framework.

Finally, there are two particularly fruitful ways of advancing empirical research. First, the use of panel data, with measurements of both socioeconomic and attitudinal variables at each wave, is ideally suited to check predictions deduced from various theories. And second, much can be learned from comparisons across contexts if attempts are made to specify and to test the presumed contextual causes at work.

See also: Demographic Transition, Second; Families and Households, Behavioral Demography of; Family Size Preferences; Family Systems and the Preferred Sex of Children; Family Theory and the Realities of Childbearing Behavior; Family Theory: Competing Perspectives in Social Demography; Family Theory: Economics of Childbearing; Family Theory: Role of Changing Values

Bibliography

- Alwin D 1990 Historical changes in parental orientations to children. *Sociological Studies of Child Development* 3: 65–86
- Axinn W G, Thornton A 1993 Mothers, children, and cohabitation: The intergenerational effects of attitudes and behavior. *American Sociological Review* 58: 233–46
- Barber J S 1999 Ideational influences on the transition to parenthood: Attitudes toward childbearing and competing alternatives. Paper presented at the Annual Meetings of the Population Association of America, New York, March 25–27
- Becker G S 1991 *A Treatise on the Family* (enlarged edn.). Harvard University Press, Cambridge, MA
- Becker G S, Landes E M, Michael R T 1977 An economic analysis of marital instability. *Journal of Political Economy* 85(6): 1141–87
- Bergstrom T C 1997 A survey of theories of the family. In: Rosenzweig M R, Stark O (eds.) *Handbook of Population and Family Economics*. Elsevier, Amsterdam, pp. 21–76
- Berrington A, Diamond I 1999 Marital dissolution among the 1958 British birth cohort: The role of cohabitation. *Population Studies* 53(1): 19–38
- Bongaarts J, Mauldin W P, Phillips J F 1990 The demographic impact of family planning programs. *Studies in Family Planning* 21: 299–310
- Boulanger P-M, Deboosere P, Lambert A, Lesthaeghe R, Surkyn J 1997 *Ménages et Familles*, Census Monograph 4. National Institute of Statistics, Brussels
- Bracher M, Santow G 1998 Economic independence and union formation in Sweden. *Population Studies* 52(3): 275–94
- Bumpass L L 1990 What's happening to the family? Interactions between demographic and institutional changes. *Demography* 27(4): 483–98
- Burch T K 1996 Icons, strawmen, and precision: Reflections on demographic theories of fertility decline. *The Sociological Quarterly* 37(1): 59–81
- Cherlin A J, Kiernan K E, Chase-Lansdale P L 1995 Parental divorce in childhood and demographic outcomes in younger adulthood. *Demography* 32(3): 299–318
- Clarkberg M, Stolzenberg R M, Waite L J 1995 Attitudes, values, and entry into cohabitational versus marital unions. *Social Forces* 74: 609–32
- Cleland J, Wilson C 1987 Demand theories of the fertility transition: An iconoclastic view. *Population Studies* 41(1): 5–30
- Easterlin R A 1976 The conflict between aspirations and resources. *Population and Development Review* 2(3): 417–25
- Easterlin R A 1980 *Birth and Fortune*. Basic Books, New York
- Easterlin R, Crimmins E 1991 Private materialism, personal self-fulfillment, family life, and public interest: The nature, effects and causes of recent changes in the values of American youth. *Public Opinion Quarterly* 55: 499–533
- Easterlin R A, Macdonald C, Macunovich D J 1990 How have American baby boomers fared? Earnings and economic well-being of young adults 1964–1987. *Journal of Population Economics* 3(4): 277–90
- England P, Farkas G 1986 *Households, Employment, and Gender: A Social, Economic, and Demographic View*. Aldine De Gruyter, New York
- Folbre N 1996 *The Economics of the Family*. Edward Elgar, Cheltenham, UK
- Hotz V J, Klerman J A, Willis R J 1997 The economics of fertility in developed countries. In: Rosenzweig M R, Stark O (eds.) *Handbook of Population and Family Economics*. Elsevier, Amsterdam, pp. 276–347
- Inglehart R 1990 *Culture Shift in Advanced Industrial Society*. Princeton University Press, Princeton, NJ
- van de Kaa D J 1996 Anchored narratives: The story and findings of half a century of research in the determinants of fertility. *Population Studies* 50: 389–432
- Kiernan K E, Hobcraft J 1997 Parental divorce during childhood: Age at first intercourse, partnership, and parenthood. *Population Studies* 51(1): 41–56
- Kravdal O 1999 Does marriage require a stronger economic underpinning than informal cohabitation? *Population Studies* 53(1): 63–80
- Lakatos I 1970 Falsification and the methodology of scientific research programmes. In: Lakatos I, Musgrave A (eds.) *Criticism and the Growth of Knowledge*. Cambridge University Press, Cambridge, UK, pp. 91–196
- Lakatos I, Musgrave A (eds.) 1970 *Criticism and the Growth of Knowledge*. Cambridge University Press, Cambridge, UK
- Lesthaeghe R 1995 The second demographic transition—An interpretation. In: Mason K O, Jensen A-M (eds.) *Gender and Family change in Industrialized Countries*. Clarendon Press, Oxford, UK, pp. 17–62
- Lesthaeghe R 1998 On theory development: Applications to the study of family formation. *Population and Development Review* 24(1): 1–14
- Lesthaeghe R, van de Kaa D J 1986 Twee demografische transitities. In: Lesthaeghe R, van de Kaa D J (eds.) *Bevolking—Groeï en Krimp*. Van Loghum-Slaterus, Deventer, pp. 19–68
- Lesthaeghe R, Meekers D 1986 Value changes and the dimensions of familism in the European Community. *European Journal of Population* 2: 225–68

- Lesthaeghe R, Surkyn J 1988 Cultural dynamics and economic theories of fertility change. *Population and Development Review* **14**(1): 1–45
- Moors G 1999 Gender role attitudes, values, and family formation. Paper presented at the Annual Meetings of the Population Association of America, New York, March 25–27
- Oppenheimer V K 1988 A theory of marriage timing. *American Journal of Sociology* **94**: 563–91
- Oppenheimer V K, Lew V 1995 American marriage formation in the 1980s. How important was women's economic independence? In: Mason K O, Jensen A.-M (eds.) *Gender and Family Change in Industrialized Countries*. Clarendon Press, Oxford, UK, pp. 105–38
- Pollak R A, Watkins S C 1993 Cultural and economic approaches to fertility: Proper marriage or mésalliance? *Population and Development Review* **19**(3): 467–96
- Preston S H 1986 Changing values and falling birth rates. *Population and Development Review* **12**(suppl.): 176–95
- Robinson W C 1997 The economic theory of fertility over three decades. *Population Studies* **51**(1): 63–74
- Schultz T P 1997 Demand for children in low-income countries. In: Rosenzweig M R, Stark O (eds.) *Handbook of Population and Family Economics*. Elsevier, Amsterdam, pp. 349–430
- Stigler J, Becker G S 1977 De gustibus non est disputandum. *American Economic Review* **67**(2): 76–90
- Thornton A, Camburn D 1987 The influence of the family on premarital sexual attitudes and behavior. *Demography* **24**(3): 323–40
- Villeneuve-Gokalp C 1990 [From marriage to cohabitation—a recent history of changes in sensual unions.] Du mariage aux unions sans papiers: Histoire récente des transformations conjugales. *Population* **45**(2): 265–98

R. Lesthaeghe

Family Theory: Economics of Childbearing

Economic theories of childbearing relate decisions about family size and the timing of births over people's lives to elements of the constraints on choice that people face, particularly prices and incomes. This article presents the basic concepts and approaches used in these theories, discusses their main lines of argument and summarizes some of their empirical implications. It first examines 'static' theories of completed family size and then discusses dynamic theories of the timing of childbearing. While mainly concerned with theories of marital fertility, it concludes with a discussion of nonmarital fertility.

1. 'Child Quality' and Fertility

Economic theories of fertility have been around for over 200 years. The ones advanced by Thomas Malthus are the best known. In a few words, Malthus' view was that fertility was determined by the age at

marriage and frequency of coition during marriage. He argued that an increase in people's income would encourage them to marry earlier and have sexual intercourse more often. Modern economic theories of fertility date from a paper by Becker (1960), which generalized and developed the Malthusian theory. Until Sect. 4, it is assumed that fertility decisions are made in a stable marriage.

Becker's point of departure is that children are usually a source of psychic satisfaction for parents. An important aspect of his seminal contribution was to point out that this satisfaction is likely to depend on the amount that parents spend on children as well as the number of children that they have. For instance, parents are likely to receive higher satisfaction if their children obtain a university degree than if they leave school as teenagers. For short, he calls children who have more spent on them 'higher quality' children, hastening to add 'that "higher quality" does not mean morally better' (Becker 1960, p. 211). The basic idea is that if parents voluntarily spend more on a child, it is because they obtain additional satisfaction from the additional expenditure. It is this additional satisfaction that is called 'higher quality.'

The concept of 'child quality' continues to be an important element in the economic analysis of fertility today. The characterization of 'child quality' has moved in the direction of identifying child quality with the lifetime well-being of the child, which can be increased by investing more in the child's 'human capital' (i.e., his or her ability to earn income as an adult) or by the direct transfer of wealth to the child. That is, the parents receive more satisfaction from having children who are better off throughout their lives, and they make monetary transfers and human capital investments to influence their child's lifetime standard of living. Thus, child quality could be thought of as the child's future 'quality of life.'

Having incorporated the quality dimension of reproduction decisions, the well-developed model of consumer choice is applied to them. Parents are assumed to choose the best combination of the number of children, their quality and the parents' own standard of living subject to parents' lifetime income and the prices that they face in markets. An increase in parents' income is expected to increase both the quantity and quality of children. In light of the tendency for people to increase the quality of most consumer durable goods proportionately much more than the quantity of such goods when their income increases, Becker argued that this is also likely to be the case for children. In economics, the 'income elasticity' of a good is defined as the percentage increase in the demand for that good in response to a 1 percent increase in income. Thus, using this terminology, Becker's contention is that the income elasticity of the number of children ('quantity') is small compared to the income elasticity of child quality. Put somewhat differently, an increase in parents' income

may increase the amount spent on children substantially, but this would mainly take the form of higher quality rather than more children. The introduction of the child quality concept generalizes the Malthusian theory, which only allows responses to income along the quantity dimension.

1.1 Quantity–Quality Interaction

A new twist to the quantity–quality model was added in the influential papers by Willis (1973) and Becker and Lewis (1973). They assumed that parents view child quantity and quality as substitutes and that parents treat all their children equally, in the sense that child quality is the same for each child. The parents' lifetime budget constraint is that the sum of expenditures on their own consumption, expenditures on their children's human capital acquisition, and monetary transfers to their children equals the parents' lifetime income. The latter two elements of spending represent expenditure on child quality, which is the product of the number of children, quality per child and the full cost per unit of child quality (which is discussed later). This model can be generalized to incorporate costs of the number of children that do not depend on child quality and costs of quality that do not depend on the number of children. As an example of the former, a decline in the cost of averting births, say because of the introduction of the oral contraceptive pill, would increase the marginal cost of a birth without affecting the marginal cost of child quality.

This product in the budget constraint implies that the cost (or 'shadow price') of an additional child is proportional to the level of child quality, and the cost ('shadow price') of raising child quality is proportional to the number of children the parents have. As a consequence, there is an important interaction between family size and child quality. Suppose, for example, that there is a *small* exogenous decrease in fertility, say because of lower contraception costs. This lowers the shadow price of child quality, which in turn raises child quality, which raises the shadow price of children, which lowers family size further, and so on. Thus, a cumulative process favoring child quality and reducing fertility ensues. Similarly, a *small* exogenous increase in desired child quality could produce a *large* decrease in family size and a much larger increase in child quality.

This interaction between child quality and fertility does not derive from assuming that child quality and quantity are *close* substitutes for one another. Indeed, just the opposite must be the case, because if both fertility and child quality are to be nonzero, then they cannot be close substitutes. Thus, family size can be highly responsive to changes in prices and incomes, even though children have no close substitutes.

Another consequence of this interaction is that it may appear that the income elasticity of fertility is

negative, even though children are 'normal goods,' in the sense that parents want more of them when parental income increases. The reason is that the true income elasticity is defined with prices constant, but, because of the interaction, it is not possible to hold the shadow prices of an additional child and of child quality constant when measuring the elasticity. The measured income elasticity of family size is more likely to be negative (even though the true elasticity is positive), the more responsive is child quality to changes in income relative to the responsiveness of fertility (i.e., the higher is the quality elasticity relative to the quantity elasticity), and the better substitute is family size for child quality or parental consumption.

1.2 Child Mortality and Fertility

The ultimate manifestation of low child quality is a child not surviving to adulthood. In light of the 'demographic transition' (i.e., the change from a high fertility–high child mortality environment to a low fertility–low mortality one), an interesting question is how fertility responds to changes in the 'risk' of child mortality. Because parents may know the risk, but not exactly how many of their children will die for any given number of births, this is a problem of choice under uncertainty. Sah (1991) assumed that parents choose the number of births to maximize their 'expected utility.' This is the weighted sum of the utilities of having particular numbers of surviving children (inclusive of all benefits and costs), with the weights being the probabilities of obtaining that number of surviving children from given numbers of births.

Suppose first that only surviving children bring costs and benefits. Sah (1991) showed that the number of births cannot increase with an increase in the probability that the child survives to adulthood. While the parents may wish to have more births when survival chances are low, because some may die before adulthood and they will not have enough time to replace them in their fecund period, their incentive to practice such 'hoarding' is less when survival chances are better. Thus, their fertility will not be higher when child survival is more likely.

A cost may, however, be incurred by a birth regardless of whether or not the child survives. In this case, a higher probability of child survival reduces the effective price of a surviving birth, thereby encouraging higher fertility. Sah finds that this effect will be smaller than the 'hoarding effect' above if parents are sufficiently 'risk averse,' or there is a 'target fertility level,' in the sense that the marginal utility of the last child is nonpositive if all the children from an optimally chosen number of births were to survive. If either of these conditions hold, better survival chances for children tend to reduce fertility.

In the analysis by Cigno (1998), parents can influence the chances that their child survives (an element

of 'child quality') by spending more on him/her. In his model, an exogenous reduction in child mortality may either raise or lower fertility. When the level of child mortality is high, reductions in it are likely to raise both fertility and survival-enhancing expenditures on children, because it lowers the price of a surviving child. When, however, the level of child mortality is already low, further reductions in it are likely to reduce both fertility and survival-enhancing expenditures on children.

2. The Cost of Children

Consumer theory tells us that, like any other price, a higher full cost of child quality per unit ought to reduce both fertility and child quality. Elucidation of the factors affecting this cost was another important contribution to the literature on the economics of fertility. In particular, it acknowledged the key role of parental time, especially the mother's, in the rearing of and investment in children. The foundations for this analysis were laid in a paper by Mincer (1963) and in Becker's (1965) theory of time allocation, and Willis (1973) built on them in his analysis.

The concept of 'household production' is important in understanding what influences the cost of child quality. Following Becker (1965), the parents' standard of living, the number of children and quality per child yield satisfaction to the parents, but these cannot be purchased directly in the market. Rather, the parents' standard of living and child quality are both produced within the household by combining parents' own time and purchased goods and services. It is the costs of these inputs that determine the cost of child quality relative to the cost of the parents' living standard. Parental time in the production of child quality is primarily the mother's time, and the rearing of children is assumed to be *time-intensive* relative to other home production activities. That is, the proportion of the total cost of producing children represented by the value of the mother's time input to children is larger than the proportion of the total cost of producing the parents' standard of living represented by the value of her time input to that production. Thus, the unit cost of child quality relative to the cost of the parents' living standard is directly related to the mother's cost of time. If she has ever been in paid employment, her cost of time is the wage she could earn in employment (i.e., her forgone earnings). The higher her wage, the higher the cost of an additional child and of additional quality per child.

This analysis suggests, therefore, that men's and women's wages should have different effects on fertility. Higher wages mean higher income for the parents, encouraging them to have more children and to invest more in the human capital of each child or to make larger monetary transfers to them (i.e., higher quality). But higher wages also mean more income lost by those caring for children and investing in their

human capital. Because most child-rearing is done by the mother, higher women's wages raise the opportunity cost of a child as well as increasing family income, while higher men's wages mainly affect childbearing through their effect on the couple's income. If the opportunity cost effect of women's wages is larger than their effect on fertility through higher family income, then higher women's wages reduce family size and child quality. It then follows that an increase in the ratio of women's to men's wages would reduce fertility.

In this model, the fraction of the woman's time supplied to the labor market is also a choice variable, which depends on the wage that the woman can earn and her partner's wage. Women who can earn high wages, or whose partners have low earnings, tend to devote a high proportion of their time to the labor market and remain childless. At the other extreme, women who could only earn low wages, or who have partners whose earnings are high, do not work in the labor market at all and have large families. For this latter group, women's wages have no impact on their fertility and higher men's wages increase the opportunity cost of children, because home production time is scarce for these women (they cannot increase it by reducing time in paid employment) and the production of child quality is assumed to be time-intensive relative to other household production activities. Thus, higher husband's wages have a smaller effect on fertility among this group of couples than among couples in which the woman spends some time in paid employment.

In between these two extremes are the bulk of the population of women, who combine childbearing with paid employment. Within this group, there would tend to be a negative correlation between family size and the fraction of their life they spend in paid employment. This is a correlation between two choice variables. Variations in each partner's wages and in tastes drive both variables.

Purchased childcare can weaken the link between a woman's wage and the cost of an additional child. In Ermisch (1989) it is treated as an increasingly imperfect substitute for the mother's time in child-rearing as the amount of it increases. Mothers with high wages tend to purchase a much larger proportion of childcare time. For them, higher wages have little effect on the cost of children, making it more likely that they increase fertility by raising family income. Similarly, in countries with heavily subsidized childcare, mothers contribute much less of childcare themselves, making it more likely that women earning higher pay have larger families. At low to moderate levels of wages, a higher mother's wage reduces fertility, but its negative impact attenuates as her wage rises, or the price of childcare falls, because mothers purchase a larger proportion of childcare time under these conditions. The impact of the price of childcare on fertility displays a similar interaction, becoming more negative as the mother's wage rises.

Nevertheless, when examining changes in fertility over time, the cost of purchased childcare and women's wages tend to move together, because women's labor is such an important input to the provision of childcare services. Thus, over time women's pay relative to men's and fertility should be negatively related, because higher women's pay raises the cost of children.

3. *Dynamic Fertility Models*

The theories discussed so far are 'static' in nature, the relevant time unit being the parents' lifetime. In order to consider decisions about the timing of births, imperfect fertility control, or the consequences of unexpected outcomes like birth-control failure or child mortality, a dynamic model is needed. First, some features of dynamic life-cycle theories are discussed, and then a model of fertility dynamics across generations is introduced.

3.1 *Life-cycle Dynamics*

Dynamic models incorporating a sequence of decisions and imperfect, costly fertility control are more realistic than the static theories discussed above. Their predictions often depend on what is assumed about parents' ability to borrow and save. One of two polar assumptions has been made in the literature: either parents can borrow and save as they choose ('perfect capital markets') or they cannot do so at all (i.e., they must consume all of their current income).

In one of the earliest dynamic fertility models (Heckman and Willis 1976), a couple's fertility is controllable by the contraceptive strategies they adopt, but control is not perfect. Because it is assumed that couples start obtaining satisfaction from children as soon as they are born, they wish to have their children early. Couples may, however, find it best to adopt a 'precautionary' contraceptive strategy early in their partnership in order to reduce the risk of having more children than they would have chosen to have if fertility control were perfect and free. Such a strategy would result in postponing the age at first birth. If people cannot borrow or save, and their incomes rise with age, they would also wish to practice contraception in order to try to postpone childbearing to a time when the marginal benefit of income for parental consumption is lower (when they can 'better afford' children). Subsequent spacing of births (rather than having children as quickly as possible) would also be a response to this tension between their desire to have children early and the economic incentive to have them later, when parents' income is higher. The more rapid the increase in income with age, the greater the incentive to postpone the start of childbearing and to space births. In other words, births are timed in order to 'smooth' parental consumption over time.

Another common (but not universal) feature of dynamic models is allowance for the possibility that a woman's current participation in paid employment improves her future earning capacity, and that absence from employment may reduce her pay when she returns to employment. Thus, a mother's absence from paid employment because of a birth has two costs: a loss of current earnings and a loss of future earning potential. An early example of such a model is that of Moffitt (1984).

Even when there are perfect capital markets, couples may choose to postpone childbearing because these opportunity costs of childbearing are high when the woman is younger and on the steep part of her career-earnings profile. Women in jobs in which work experience has little effect on earnings have less incentive to postpone. In addition, if the care of young children is particularly intensive in the mother's time, then a birth produces a temporary increase in the value of her time in the home to a level above the value of her time in the labor market, causing her to leave paid employment. It also increases the cost of an additional child, encouraging her to postpone the next birth until the value of her time declines sufficiently through the aging of the child. Thus, even with perfect capital markets, there is an economic incentive to space births. An unexpected birth, because of contraceptive failure, would generate similar incentives for the mother to leave paid employment temporarily and postpone having another child.

The death of a child encourages its replacement, because the marginal benefit of an additional child is likely to increase substantially when a child dies. Replacement makes a woman's number of births increase with the experience of child mortality. Unless child mortality is expected when the mother is older, replacement is a better response to child mortality risk than 'hoarding,' because the latter involves a larger deviation from what the parents would have done in the absence of child mortality. Wolpin (1984) was the first to incorporate uncertainty about child survival into a sequential, dynamic fertility model, which permits analysis of replacement in response to child deaths and of the effect of infant mortality risk on fertility.

Recent detailed surveys by Hotz et al. (1997) and Schultz (1997) discuss the empirical strategies for applying the theories discussed above and the results of many empirical studies.

3.2 *Easterlin's Relative Income Hypothesis*

In Easterlin's (1980) view, an important factor affecting a young couple's willingness to marry and have children is their outlook for supporting their material aspirations. If they expect their earning potential to be high relative to their aspirations, they will be more optimistic and will feel freer to marry and have children. If they expect that they will find it difficult to

achieve the standard of living to which they aspire, they will hesitate to marry and have children. The ratio of the couple's potential earning power to their material aspirations is called their 'relative income.'

A person's material aspirations are formed through experience; for a young person, the primary relevant experience is the standard of living experienced in his/her parents' home. Young people's expectations concerning their earning potential are largely shaped by their experience of working.

A person's relative income may be related to the size of his/her generation relative to the size of previous generations. There are important differences between the characteristics of new labor force entrants and experienced workers. The former have not yet acquired a high level of skills and are engaged in a considerable amount of job search with consequent high job turnover. Older workers are more skilled and occupy career jobs with relatively low job turnover. As a consequence of these differences, the ability to substitute between young and older workers is limited. Thus, when young workers are scarce relative to older workers, their wages are higher relative to older workers and their chances of upward mobility improve. When they are in relative abundance, their relative wage declines and their promotion chances deteriorate. If there is a steady growth in labor demand, the earnings of young people (say, aged 15–29) relative to older people (say, aged 30–64) tends to be inversely related to the number of young people relative to older people ('relative generation size'), which suggests that their relative income also moves inversely with relative generation size.

The relative income theory of fertility would then lead us to expect that people from smaller birth cohorts would have higher fertility than people from larger cohorts. If couples feel constrained to choose family size within certain limits, say two to four children, then 'scarce' generations, who start childbearing earlier because of their favorable economic prospects relative to their material aspirations, would cut back later as they reach their upper limits, and people from 'surplus' generations would catch up later to attain their lower limits. Such behavior tends to make the fertility rates of women of different ages (different generations) move together. The period total fertility rate would then also move inversely with relative generation size, and it would fluctuate much more than completed family size.

4. Non-marital Fertility

Since 1990 there has been a shift away from the paradigm of the married couple making fertility decisions as a single entity. This reflects in part the fact that births outside marriage now represent a substantial proportion of births in many countries, which suggests that men and women should be treated as

individual agents in fertility decisions. Higher divorce rates would reinforce such a treatment.

This is clearest considering births outside a live-in partnership, drawing on the analysis of Willis (1999). Child quality is a 'collective good' to the parents (i.e., both the mother's and father's satisfaction depend on child quality). As a consequence, men fathering children outside a live-in partnership may choose to make transfers to the mother. Assume that the amount of transfer is determined in a noncooperative way: the mother has custody and chooses child quality and her own consumption taking transfer payments from the father as given. Fathers choose transfers and their own consumption, taking into account how the mother reacts to such a transfer. Then a father would voluntarily make a transfer if his income is high relative to the mother's income and he cares sufficiently for the well-being of his child. The higher is his income relative to hers, the larger is the transfer. But the mother, father, and child could be even better off if the couple cooperated in their decisions. Because of difficulties in monitoring expenditures on children, cooperation is facilitated by the parents forming a live-in partnership. Thus, these relatively affluent fathers would usually not want to father children outside a live-in partnership.

In a frictionless 'marriage market,' the collective-good nature of child quality encourages the highest income man to live with and have a child with the highest income woman, the next highest income couple would live together and have a child, and so on. If there are more men than women, then the men with the lowest incomes remain without a partner and all childbearing would be within live-in partnerships. If the number of women exceeds the number of men and if women at the lower end of the income distribution have incomes which are sufficient for them to want to raise children with their own resources, then fatherhood outside a live-in partnership would be free. If low-income men can father children by more than one woman, out-of-partnership fatherhood may be more advantageous for them than forming a union with a low-income woman, even though the mother and child would be better off in the union. This is more likely if the man's income is not much higher than the woman's, in which case he would not want to make transfers to her. The single low-income mothers gain by becoming mothers rather than remaining childless.

Thus, there may be what Willis (1999) calls an 'out-of-wedlock equilibrium' in which men with low incomes seek to father children outside live-in partnerships and women oblige them in preference to remaining childless. This outcome is more likely when there is an excess of unmarried women relative to unmarried men, men's incomes are not much higher than women's in the lower portion of the income distribution, and the combination of women's earnings and government transfer programs are adequate to raise children without the father's help. Thus, this

theory explains fertility outside live-in partnerships as a rational choice by low-income men and women, while at the same time higher income couples choose to have children within live-in partnerships.

In the world described by this model, people either form live-in partnerships or remain single forever. A simple model of time-consuming search for a partner, in which a woman may have a birth before entering a live-in partnership but still partner subsequently, yields a few similar predictions. Pre-partnership childbearing is more likely when the incomes of single mothers are higher relative to the incomes of single women (perhaps because of welfare benefits to single mothers), both because more partnership offers are rejected and because the value of becoming a single mother increases relative to the value of remaining single and childless. It is also more likely when single motherhood does not substantially reduce the chances of subsequent live-in partnering. The fathers of these children, whose partnership offers have been rejected, tend to have low incomes.

There has been little economic analysis of childbearing within cohabiting unions, which has grown in importance in many countries. In unions that are expected to have a short life, it may be akin to the analysis of out-of-partnership childbearing discussed above. More generally, it shares some features of the analysis of marital fertility, but there are also unique attributes of cohabiting unions that are probably important for fertility decisions. One is their high risk of dissolution. In light of it, it is essential to treat the fertility decision as problem of choice under uncertainty in which each partner is an individual agent whose preferences, expectations, and resources influence the decision. This presents new and difficult analytical problems, which are also relevant to marital fertility in high-divorce countries. These are new challenges in the economic theory of fertility.

See also: Family Size Preferences; Family Theory and the Realities of Childbearing Behavior; Family Theory: Competing Perspectives in Social Demography; Family Theory: Complementarity of Economic and Social Explanations; Family Theory: Feminist–Economist Critique; Fertility: Institutional and Political Approaches; Fertility of Single and Cohabiting Women; Fertility: Political and Political–Economic Perspectives; Fertility Theory: Caldwell’s Theory of Intergenerational Wealth Flows; Fertility Theory: Embodied-capital Theory of Life History Evolution; Mortality, Biodemography of; Motherhood: Economic Aspects

Bibliography

Becker G 1960 An economic analysis of fertility. In: *Demographic and Economic Change in Developed Countries*. National Bureau of Economic Research, Princeton University Press, Princeton, NJ, pp. 209–31

- Becker G 1965 A theory of the allocation of time. *Economic Journal* **75**: 493–517
- Becker G, Lewis H G 1973 On the interaction between the quantity and quality of children. *Journal of Political Economy* **81**: S279–S288
- Cigno A 1998 Fertility decisions when infant survival is endogenous. *Journal of Population Economics* **11**: 21–8
- Easterlin R 1980 *Birth and Fortune*. Grant-McIntyre, London
- Ermisch J 1989 Purchased child care, optimal family size and mother’s employment. *Journal of Population Economics* **2**: 79–102
- Heckman J, Willis R 1976 Estimation of a stochastic model of reproduction: An econometric approach. In: Terleckyj N (ed.) *Household Production and Consumption*. Columbia University Press, New York
- Hotz V J, Klerman J, Willis R 1997 The economics of fertility in developed countries. In: Rosenzweig M, Stark O (eds.) *Handbook of Population and Family Economics*. Elsevier, Amsterdam, pp. 241–347
- Mincer J 1963 Market prices, opportunity costs and income effects. In: *Measurement in Economics: Studies in Mathematical Economics in Honor of Yehuda Grunfeld*. Stanford University Press, Stanford, CA, pp. 67–82
- Moffitt R 1984 Optimal life-cycle profiles of fertility and labor supply. *Research in Population Economics* **5**: 29–50
- Sah R 1991 The effect of child mortality changes on fertility choice and parental welfare. *Journal of Political Economy* **99**: 582–606
- Schultz T P 1997 Demand for children in low income countries. In: Rosenzweig M, Stark O (eds.) *Handbook of Population and Family Economics*. Elsevier, Amsterdam, pp. 349–430
- Willis R 1973 A new approach to the economic theory of fertility behavior. *Journal of Political Economy* **81**: S14–S64
- Willis R 1999 A theory of out-of-wedlock childbearing. *Journal of Political Economy* **107**: S33–S64
- Wolpin K 1984 An estimable dynamic stochastic model of fertility and child mortality. *Journal of Political Economy* **92**: 852–74

J. Ermisch

Family Theory: Economics of Intergenerational Relations

In economics, the term ‘transfer’ is used to refer to the delivery of money or other assets, without counterpart. We do not talk of transfers, therefore, in the context of a commercial transaction, where money is the counterpart of goods or services. Public transfers, by contrast, are the direct subsidies that governments pay to households and firms. Although such subsidies are financed out of tax receipts, there is in fact no direct link between what any particular firm or household pays in taxes, and what it gets back in subsidies. A gray area is represented by certain types of old-age pensions, where the size of the benefit is related in some way to contributions made, but the rules on the basis of which the calculation is made are subject to change by political decision.

Transfers between generations may be voluntary, or forced by the tax system. Here, we are concerned with

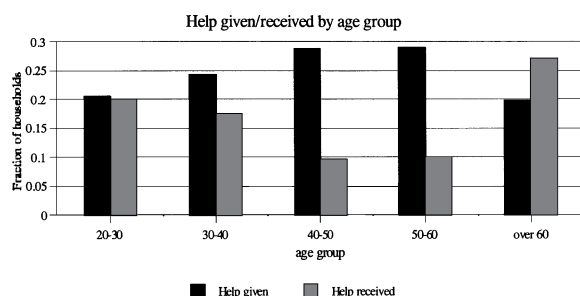


Figure 1
Help given/received by age group

voluntary transfers between different generations of the same family, but we are also interested in how these respond to public transfers. Typically, the latter redistribute income away from working-age cohorts, who pay more taxes and receive less subsidies than younger and older cohorts. A similar age pattern applies to voluntary *inter vivos* transfers: Fig. 1, drawn from Cigno and Rosati (2000), shows the distribution, by age of the household head, of monetary help to and from non-coresident 'friends or relatives' (overwhelmingly relatives) by Italian households during the late 1980s. It transpires that the probability of making a transfer is highest, and that of receiving one lowest, when the household head is in the 40–59 age range. Similar pictures can be constructed for other countries. What can be the reason for such a pattern?

Economic explanations fall into two broad categories. One follows what is called the altruistic approach, whereby individuals are assumed to derive direct utility (satisfaction) from the wellbeing of other family members. The other assumes that individuals are ultimately self-interested. Transfers are straight, no-strings-attached gifts according to the first of these explanations, part of a strategy that is expected to bring, directly or indirectly, benefits to the donor according to the second.

1. Altruistic Explanations

Simple altruistic explanations assume that sentiment is unidirectional: either from parents to children (*descending altruism*) or from children to parents (*ascending altruism*). More complicated versions allow for altruism in both directions.

Descending altruism underlies Richard Barro's and Gary Becker's theories of voluntary transfers. This assumption allows them to model transfers within the family as the outcome of a single optimization, whereby parents decide how much each family member will consume, subject to two constraints. One is that the sum of the consumption expenditures of all

family members cannot exceed the sum of their incomes. The other is that transfers received (the difference between consumption expenditure and income) by any family member other than the decision-maker(s) must be positive or zero, otherwise they would be rejected. Under standard conditions, the model predicts that transfers will not decrease (increase) as the income of the donor (beneficiary) increases. This is consistent with the observation (Fig. 1) that the probability of making a transfer is highest, and that of receiving one lowest, in middle age, when incomes are generally at their peak. It is difficult to reconcile, however, with the observation that people continue to receive *inter vivos* transfers after the age of 60 (or even 75), when few of them have living parents.

If it is assumed that parents make transfers in the form of bequests, rather than *inter vivos*, the descending-altruism model predicts that household saving is increased, dollar for dollar, by any current government deficit (Barro 1974). That is because, conscious of the government intertemporal budget constraint, altruistic taxpayers perceive any current deficit as a tax on their descendants (this is the so-called *Ricardian equivalence*, from the name of David Ricardo, who first detected it). According to the model, taxpayers thus respond to a government deficit by saving more, in order to be able to leave their descendants bequests large enough to pay the extra taxes. One of the mechanisms that might generate such a deficit is a pay-as-you-go public pension system, where current pensions are paid out of current contributions. Whichever the mechanism, however, any net public transfer will be matched by private transfers of opposite sign.

Ricardian equivalence does not hold if fertility is endogenous i.e., if parents can choose how many children to have (or, more realistically, condition the probability distribution of births), and parents derive satisfaction from the number of children, as well as from how much each of them consumes (Becker and Barro 1988). Suppose, then, that the government were to take a dollar off the social security contributions due by each of the present workers, without any corresponding reduction in their pension entitlements. Each member of the next generation would then have to pay (r/n) more dollars in taxes, where r is the intergenerational interest factor, and n the ratio of future to current taxpayers (roughly, the population growth factor). Knowing that, present workers would then want to compensate their own children by making larger bequests to each of them. But this would make it more expensive to have an extra child relative to increasing the wellbeing of existing children. Under standard conditions, the present generation would then choose to have fewer children, and to *more* than compensate each child for the heavier tax burden he or she will have to shoulder. Of course, the change in fertility behavior would result in a lower n , and thus raise the burden on each future taxpayer, but the effect

of each couple's actions on the aggregate fertility rate is infinitesimal. Couples, therefore, will not take this effect into account when they decide how many children to have. Hence, fertility would fall. Saving could rise or fall, depending on whether the rise in transfers per child is proportionally larger or smaller than the reduction in the number of children. By contrast, if the government were to promise higher benefits to current taxpayers already covered by old-age security, or extend coverage to new categories, and simultaneously raise taxes or contributions so as not to force a transfer in favor of present workers, there would be no effect on either saving or fertility behaviour. There is evidence (Cigno and Rosati 1996, among others) that old-age security *per se* (i.e., without intergenerational spillovers) encourages household saving and discourages fertility, while deficits discourage saving. This empirical finding is consistent with the Becker-Barro prediction that a deficit *could* result in lower savings, but contradicts that model's prediction that, in the absence of intergenerational transfers, saving and fertility would remain the same.

The implications of ascending altruism (Nishimura and Zhang 1992) are somewhat different. The direction of transfers is not from working-age parents to young children, as in the case of descending altruism, but from working-age children to retirement-age parents. This is consistent with the observation that old people receive *inter vivos* transfers, but difficult to reconcile with the one that they also make such transfers. As children are a source of income, parents treat them as an asset. If fertility is endogenous, working-age individuals (or couples) then choose the composition of their portfolios so as to equate the marginal return to having children to that of holding conventional assets (capital). According to this model, a gift to current tax payers at the expense of future ones would induce parents to save more, not in order to leave larger bequests, but in order to put some of the windfall aside for old age. By contrast, a fully funded expansion in pension coverage (a form of forced saving) would reduce voluntary saving. These predictions clash, however, with the empirical finding, already mentioned, that deficits discourage voluntary household saving, while social security coverage *per se* has the opposite effect.

Whichever the direction, unilateral altruism is thus an inadequate explanation of what we observe. Can bilateral altruism (parents love children, children love parents) do better? If family members are unanimous in their preferences, intrafamily transfers can again be described as the outcome of the maximization of a single preference function (this time agreed upon by all family members), subject only to the constraint that the sum of the expenditures cannot be greater than the sum of the incomes of all family members. There is now no restriction on the direction of money flows: depending on life-cycle patterns, transfers from middle-aged children to elderly parents may be positive in

some families and negative in others. This is consistent with evidence of two-way flows between generations, but the model also has a stronger behavioral implication: since the amount consumed by each family member depends on *total* family income, and not on his or her own individual income, changing the intrafamily distribution of income should not change the intrafamily distribution of consumption. Statistical tests based on this property of the model (Cox 1987, Altonji et al. 1992) reject the hypothesis of altruism. That property also has the strong implication that redistributive policies are ineffective: taking a dollar off the rich (the parents) and giving it to the poor (the children) would reduce the amount voluntarily transferred from the former to the latter by exactly one dollar.

In general, however, different family members are likely to have different ideas about the most desirable pattern of transfers. Even if all family members were altruistic towards one another, each of them might in fact appreciate his or her own consumption a little more (or, for that matter, a little less) than another family member. If that is the case, transfers cannot be viewed as the solution to a single optimization. They must rather be seen as the outcome of a game (in the sense of formal game theory) in which family members bargain with one another, or respond strategically to one another's moves, pretty much as if they were pursuing purely selfish aims (Manser and Brown 1980, Stark 1993). In such a situation, transfers depend on individual fall-back positions (what the player would get, if he or she dropped out of the game), as well as on preferences and characteristics. It is then difficult to distinguish empirically (McElroy 1990) bilateral altruism from self-interest, to which we now turn.

2. Non-altruistic Explanations

Non-altruistic explanations are based on the assumption that individuals are interested only in their own lifetime consumption stream. If a person voluntarily surrenders money to another, it will be for profit, or in order to improve the allocation of consumption over the life-cycle, or across states of the world. It must also be true that the person in question could not get a better deal from the market. There does not seem to be any difference of substance, therefore, between this and giving money to the grocer in exchange for a dozen eggs. Nonetheless, it is customary, maybe due to lack of information, to classify as transfers any moneys changing hands within the family whenever there is no explicit *quid pro quo* (not, however, in the case of a father paying his child to wash his car).

Mutual insurance ('I help you if things go well for me and badly for you, you help me if it is the other way round') is a typical form of intrafamily arrangement. But why should people want to enter into such informal, maybe even tacit, agreements, if it is possible to buy insurance from the market with all the safeguards

afforded by the law? The insurance market has the advantage over the family that individual risks are less likely to be correlated in a large heterogeneous population than in a small homogeneous group. Because of adverse selection and moral hazard problems, however, the market might not offer insurance, or offer it on very unfavorable terms, to certain individuals or for certain types of risk. Such problems are generally less acute within the family, the members of which have privileged information on one another's characteristics, and lower costs of monitoring one another's behavior. In some circumstances, it may thus turn out that intrafamily arrangements are more advantageous than market insurance, or that there is no alternative.

Mutual insurance may or may not result in intergenerational transfers (not if the agreement is among contemporaries). Intergenerational transfers, by contrast, are essentially another kind of insurance arrangement whereby a person (parent) transfers wealth or pre-commits to leave a bequest to a younger relative (child) in exchange for a promise of indefinite old-age support, effectively in exchange for an annuity (Kotlikoff and Spivak 1981). That way, the parent buys insurance from the child against the risk of running out of money after too long a period of retirement.

Another type of mutually advantageous intrafamily deal, also giving rise to intergenerational transfers, stems from life-cycle income variations and capital market imperfections. Standard microeconomic theory predicts that individuals smooth consumption over the life-cycle by borrowing from the market when they are young, and lending (buying assets) when they are middle-aged. In reality, however, minors cannot sign contracts, and young adults have difficulty in obtaining credit against future earnings. The very fact that people survive to adulthood is thus evidence that the young receive support in other ways, but from whom? We have just seen that nobody makes genuine presents. Self-interested relatives, on the other hand, will make loans, not presents, and only if the risk-adjusted return is at least as high as that offered by the formal capital market. And what is there to induce a selfish adult to honor an informal debt incurred 20 or so years earlier?

Imagine a set of family rules (family constitution) prescribing the amount that a parent must pay a young child, and the amount that a grown-up child must pay an elderly parent, subject to the clause that nothing is due to a parent who disobeyed the rules. Such a constitution is self-enforcing, in the sense that it is in everybody's interest to comply, if the rules are so designed that complying with them is at least as advantageous, for a middle-aged person, as the alternative strategy of having no children, giving nothing to parents, and providing for old age by buying capital (Cigno 1993). This line of reasoning explains transfers to the young as implicit loans, and transfers to the old as implicit loan repayments. In such a situation, paying

back one's own parents is a condition for getting the same treatment from one's own children. Now suppose that the level of old-age security is raised, or coverage extended to new categories. The present middle-aged affected by the policy will then want to shift less of their current spending capacity to old age. Since the amount they are supposed to pay to their elderly parents has not changed, however, some of these middle-aged persons will find it advantageous to disobey the constitution, and consequently give up the idea of being supported by their own children in old age. They will then default on their implicit debt towards their own parents, have no children, and possibly top up their expected pensions by buying assets. Conversely, if the government were to cut current taxes or social security contributions without reducing pension entitlements, some middle-aged individuals, who would otherwise have found it advantageous to buy assets, would have children instead. This is consistent with the already mentioned empirical evidence that aggregate household saving responds positively to old-age security, and negatively to deficits. It also explains why fertility responds negatively to old-age security.

See also: Age: Anthropological Aspects; Family and Gender; Family and Kinship, History of; Family, Anthropology of; Family as Institution; Family Systems in Europe; Family Theory: Competing Perspectives in Social Demography; Family Theory: Complementarity of Economic and Social Explanations; Family Theory: Economics of Childbearing; Family Theory: Feminist-Economist Critique; Family Theory: Role of Changing Values; Kinship in Anthropology; Parenting: Attitudes and Beliefs

Bibliography

- Altonji J G, Hayashi F, Kotlikoff L J 1992 Is the extended family altruistically linked? Direct evidence using micro data. *American Economic Review* **82**: 1177–98
- Barro R J 1974 Are government bonds net wealth? *Journal of Political Economy* **82**: 1095–1118
- Becker G S 1974 A theory of social interactions. *Journal of Political Economy* **82**: 1063–93
- Becker G S, Barro R J 1988 A reformulation of the economic theory of fertility. *Quarterly Journal of Economics* **103**: 1–26
- Cigno A 1993 Intergenerational transfers without altruism: Family, market and state. *European Journal of Political Economy* **7**: 505–18
- Cigno A, Rosati F C 1996 Jointly determined saving and fertility behaviour: Theory, and evidence for Germany, Italy, UK and USA. *European Economic Review* **40**: 1561–89
- Cigno A, Rosati F C 2000 Mutual interest, self-enforcing constitutions and apparent generosity. In: Gérard-Veret L A, Kolm S C, Mercier-Ythier J (eds.) *The Economics of Reciprocity, Giving and Altruism*. Macmillan, London
- Cox D 1987 Motives for private income transfers. *Journal of Political Economy* **95**: 508–46
- Kotlikoff L J, Spivak A 1981 The family as an incomplete annuities market. *Journal of Political Economy* **89**: 372–91

- Manser M, Brown M 1980 Marriage and household decision making. *International Economic Review* 21: 31–44
- McElroy M B 1990 The empirical content of Nash bargained household behavior. *Journal of Human Resources* 25: 559–83
- Nishimura K, Zhang J 1992 Pay-as-you-go public pensions with endogenous fertility. *Journal of Public Economics* 48: 239–58
- Stark O 1993 Nonmarket transfers and altruism. *European Economic Review* 37: 1413–24

A. Cigno

Family Theory: Economics of Marriage and Divorce

From an economic point of view, marriage is a voluntary partnership for the purpose of joint production and joint consumption. As such, it is comparable to other economic organizations that aim to maximize some private gains, but are subject to market discipline. This article describes the working of the marriage market, as a feedback system in which the aggregate divorce and marriage rates influence and are influenced by individual choices, and explains the features of this market that have led to the sharp increase in marital turnover.

1. The Gains from Marriage

The economic analysis of the family naturally begins with the gains from marriage. These gains are a major determinant of the decision to form a new union or to break an existing relationship. The main economic advantages of being married, rather than being single, are as follows.

(a) *Children*. Although children can be produced and raised outside the family, the family has a substantial advantage in carrying out these activities. There are two inter-related factors that cause this advantage: by nature, parents care about their own children and, because of this mutual interest, it is more efficient that the parents themselves determine the expenditure on their children. Because children are viewed as a public good by their parents, an efficient allocation of family resources between public and private uses requires some coordination between the parents. If the parents live separately, as either single or remarried, the noncustodian parent loses control of child expenditures and is less likely to contribute. In addition, lack of contact may reduce the parent's benefits from the children's 'quality' (i.e., their well being and success).

(b) *Division of labor*. Family members coordinate their work activities to exploit comparative advantage and increasing returns. For instance, one partner works in the market while the other works at home.

This pattern is economically efficient if the partners have different market wages or different productivity at home. Such differences may be inherent or, more likely, acquired through investment in skills. Investment in human capital is more profitable the higher its rate of utilization. This type of increasing returns leads to specialization that can, in principle, create mutual gains through redistribution.

(c) *Sharing*. There are collective (nonrival) goods that can be jointly consumed by both partners: for example, child quality, common leisure activities, housing expenditures, and shared information. The share of such public goods in family expenditures can be substantial. Lazear and Michael 1980 estimate that two single individuals can almost double their purchasing power by forming a union.

(d) *Credit and investment*. Married partners often engage in implicit loans that facilitate investment activities. For example, one partner works when the other is in school, expecting to share in the future gains from the investment. Similarly, parents often finance the schooling of their children, in part in the anticipation of old-age support. Such internal transfers can support investments that are mutually beneficial and would not be carried out because of market imperfections.

(e) *Risk pooling*. A family can spread the risks of idiosyncratic shocks to individual members. For example, one partner can work when the other is sick or unemployed. The family jointly accumulates precautionary savings that can serve *either* member in case of need.

There is plenty of evidence for division of labor within families. Typically, husbands work more in the market than do their wives. In addition, married men work longer hours in the market and have substantially higher wages than unmarried men. Similarly, married women have lower wages and work more at home than unmarried women. This division of labor is influenced by the higher market wages of males, which are reinforced by their higher participation in the labor market. Somewhat surprisingly, the evidence for mutual insurance is not easy to come by. In particular, an employed person is generally *more* likely to have an employed spouse than an unemployed person, suggesting that common factors affect the unemployment of both spouses.

The family is not the only organization that fulfills these functions. Markets and government institutions compete with and complement the family in this regard. If all goods and work activities are marketable, there is no need to form marriages to enjoy increasing returns or to pool risks. Similarly, with good medical or unemployment insurance one does not need to rely on one's spouse. Generally, the role of the family varies depending on market conditions and government policies. In addition, human partnerships need not be confined to couples of opposite sexes. One also observes 'extended families' of varying structures that

coordinate the activities of their members and provide self-insurance. The prevalence of male–female partnerships has to do with sexual attraction, which triggers some initial amount of blind trust and the production of children. These emotional and biological considerations are sufficient to bring into the family domain some activities that could be purchased in the market. Then, the accumulation of specific ‘marital capital,’ such as children and shared experience, increases the costs of separation and creates incentives for a lasting relationship. In this sense, there is an accumulative effect in which economic considerations and investments reinforce the natural attachment.

The potential gains from marriage generally depend on the personal traits of the two partners. Some individual traits are complements, and the gains from marriage rise if the partners have similar endowments. For instance, similarity in tastes facilitates decision on collective goods, such as the education of the children. In other cases, the traits are substitutes, and gains from marriage rise if the partners have different traits. For instance, the gains from division of labor and from credit are higher when the wages or incomes of the partners differ.

2. Family Decision Making

The existence of potential gains from marriage is not sufficient to motivate marriage and to sustain it. Prospective mates need to form some notion as to whether families realize the potential gains and how they are divided. Because families consist of several members with potentially conflicting interests, there is a question as to how families reach decisions. The old notion that consensus leads to joint maximization of a common objective appears too narrow. However, because of their proximity and repeated interaction, it is plausible to assume that married partners reach some sort of agreement that specifies an *efficient* allocation of resources and a *stable division rule*. These two principles replace the principle of rationality in individual choices.

In a special case, referred to as *transferable utility*, it is possible to separate the issues of efficiency and distribution. This situation arises if there exists a commodity (say money) that, upon changing hands, shifts utilities between the partners at a fixed rate of exchange. In this case, the family decision process can be broken into two steps: first choose an action that maximizes a weighted sum of the individual utilities, then use money to divide the jointly maximal outcome. The basic idea is that the partners can agree on most aspects of their choice and restrict their bargaining to transfers of money, which do not detract from total resources. In general, the problems of efficiency and distribution are intertwined. The family can still be described as maximizing a weighted sum of

the individual utilities, but the weights depend on the individual bargaining powers and any shift in the weights will affect the family choice. The bargaining power may depend on individual attributes, such as earning capacity, subjective factors such as impatience and risk aversion, and on market conditions, such as the sex ratio and availability of alternative mates.

Realization of the potential gains from marriage requires coordination between the two partners and durable relationship. Durable relationships are motivated by the long-term investment in children and the accumulation of marital capital, which is lost or diminished in value if separation occurs. However, to induce the actions that generate the gains from marriage, it is necessary for the partners to exchange *commitments*. For instance, a young wife is unlikely to support her husband by working while he is in medical school, if she expects him to leave her and marry a young nurse when he gets his degree. To some extent, emotional factors such as love and guilt replace formal commitments and facilitate efficient investments and effort levels. As an example, consider the role of an altruistic family head. Suppose that the head is given control over family resources and can make transfers as the head sees fit. The only requirement for becoming a head is that they should care about *all* family members, willing to raise their income, whenever the total resources at the head’s disposal rise. Under these conditions, selfish family members voluntarily act in the interest of the group. The reason is that any productive (destructive) action which increases (decreases) total family resources is rewarded by an increased (decreased) transfer from the family head (Becker 1991).

It is empirically testable whether families maximize a joint utility function, whether they act efficiently, and it is also possible to recover the sharing rule from observed family choices. Recent evidence shows that, holding total family income constant, an increase in the income of one member shifts the allocation of consumption goods in their favor. These findings reject the hypotheses that the family maximizes a joint objective, which would imply that within family, allocations are invariant under such redistribution of total resources, but are consistent with both cooperation and noncooperation. However, only cooperation implies efficiency, which yields testable cross-equation restrictions on family demand. Based on data on consumption and work patterns within families, efficiency is not rejected, suggesting that cooperation is the relevant mode (Browning et al. 1994). In the context of uncertainty, however, efficiency has some further implications that are rejected. Efficient risk sharing among linked individuals implies that, holding aggregate consumption constant, the consumption of each family member is independent of idiosyncratic shocks such as fall into unemployment or bad health. Stated differently, all individuals in the household are affected by a random shock to any individual income

and their consumption levels move together. Because of lack of data on individual consumption, the tests of this hypothesis involve larger units, such as villages or extended families, and, at this level of aggregation, efficiency is rejected.

3. Divorce and its Economic Consequences

Divorce is motivated by two general considerations. First, because it takes time to find a suitable match, people may enter a relationship which they intend to break in the future if a better match is met. Second, because of changing economic and emotional circumstances, the gains from marriage may dissipate. As time passes, new information on match quality and outside options is accumulated, and each partner decides whether to dissolve the partnership or to continue the marriage. In making this choice, partners must consider the expected value of each alternative, where the value of remaining married includes the option of later divorce, and the value of divorcing includes the option of later remarriage. Divorce occurs endogenously whenever one partner has an alternative option that the current spouse cannot, or is unwilling to, match by a redistribution of the gains from marriage.

The model outlined above yields several testable implications:

(a) It is the *unanticipated* changes in the characteristics of the partners or the quality of match which trigger divorce. It is clear that falling out of love can destroy the gains from the current marriage and cause divorce. It is less obvious how unanticipated changes in personal attributes, such as earning capacity, influence divorce, because they affect both the gains from the current marriage and the outside options. However the partners were matched, based on their traits as observed at the time of marriage, any unanticipated change reduces the gains from interaction. Thus, if the husband is unexpectedly wealthy, he will seek a better wife, and if he becomes unexpectedly poor, she will seek a better husband. Note that both positive and negative 'surprises' can disrupt the marriage.

(b) If the gains from marriage are substantial, small shocks will not lead to divorce. Therefore, the probability of divorce will be lower amongst couples who are well matched. Anticipating that, couples sort into marriage according to characteristics that are likely to enhance the stability of the marriage. In this respect, the decisions to marry and divorce are linked.

(c) By the same logic, costs of divorce mitigate the impact of unanticipated shocks on marital dissolution. The costs arise from loss of specific capital, such as information about the current spouse, and loss of coordination, especially with regard to child expenditures. There are also emotional and legal costs associated with the break up.

(d) Somewhat more controversial is the role of divorce laws, in particular whether the legal possibility unilaterally to walk away from a marriage influences the divorce rate. With transferable utility, such a change should only affect the shares in the gains from marriage but not the decision to separate, because the partner who wishes to continue the marriage can compensate the one who wishes to leave by giving up part of the gains from the marriage. Conversely, the person who wants to leave can pay damages to the one left behind. In general, because the unexpected shock differs between partners and because the options for compensations are limited, divorce is less likely if mutual consent is required.

These implications have been tested empirically. There is evidence that surprises matter. Specifically, an unexpected increase in the husband's earnings capacity reduces the divorce hazard, while an unexpected increase in the wife's earning capacity raises the divorce hazard. The important roles of search and costs of divorce are indicated by the findings that higher age at marriage and the presence of children stabilize the marriage. There is only weak evidence that divorce rates are higher in states where 'fault' is not a prerequisite for divorce. There is clear evidence that individuals sort into marriage based on their anticipation of divorce. Thus, couples with similar schooling attainments at the time of marriage are less likely to divorce, and individuals are more likely to marry if they have a similar amount of schooling. The same patterns hold for religion and ethnicity (Weiss 1997).

The costs of divorce are influenced by the post-divorce transfer of resources in the form of alimony and child support. It has been observed that divorced husbands, even if relatively well to do, fail to support their ex-wives and their children at the standard to which they were accustomed during marriage. Consequently, divorced women and children in their custody suffer a large decline in income. The possible explanations for this phenomenon are: lack of binding marriage contracts, the inability of noncustodial parents to monitor expenditures by the custodian, and loss of interest in the child by the noncustodian parent. In most cases, the wife obtains custody and controls the expenditure on children. This situation creates an *agency problem*, because out of every dollar transferred to the custodial wife with the intention of raising the welfare of the child, part is used for her own consumption. Because of this 'tax,' noncustodial fathers reduce their transfer, and consequently child quality is reduced. The assignment of custody to the wife is usually motivated by her comparative advantage in child care, but the agency problem implies that such a practice has shortcomings that are not easy to overcome, because the courts cannot verify within-household allocations.

The adequacy of transfers must be judged not only on the basis of the efficiency of allocation of family

resources in the aftermath of divorce, but also on the basis of partners' options and expectations at the time of marriage. From the latter point of view, efficiency requires that the consumption levels during marriage and divorce are tied together. Courts do take into account these two broad considerations in awarding divorce settlements. These are usually based on post-divorce incomes and the accustomed consumption of the wife and children during marriage. However, the evidence suggests that actual transfers fall short of the efficient level from both the *ex post* and *ex ante* points of view (Weiss 1997).

With deficient transfer mechanisms, the partners must better prepare for the event of divorce. One important instrument is the allocation of time within marriage. By investing in human capital, each partner can be less dependent on transfers in the event of divorce. However, such investments may detract from marital output. For instance, a wife who works is better defended against divorce but has less time to spend with children. Indeed, there is evidence that women tend to increase their investment in market work in anticipation of divorce. Thus, if the risk of divorce rises, children may suffer even before the marriage breaks.

4. The Marriage Market

Individuals in society have many potential partners. This situation creates competition over the potential gains from marriage. In modern societies, explicit price mechanisms are not observed. Nevertheless, the assignment of partners and the sharing of the gains from marriage can be analyzed within a market framework, because an undesired marriage can be avoided or replaced by a better one.

Any market solution is a particular assignment of males to females. The major questions of interest are: what types of matches are likely to form in equilibrium; for instance, would a class structure emerge where the rich marry the rich? What determines the rates of marital turnover and the proportions of married and single individuals in society? Does the marriage market operate efficiently and, if not, what type of intervention is required?

Matching models provide a starting point for such analysis. These models investigate the mapping from preferences over prospective matches into a *stable assignment* (Roth and Sotomayor 1990). Reflecting the assumption that marriage is voluntary, an assignment is said to be stable if no married person would rather be single and no two (married or unmarried) persons prefer to form a new union. To illustrate, assume that each male is endowed with a single trait, m , and each female is endowed with a single trait, f , which positively affects the gains from marriage, denoted by $z = g(m, f)$. Suppose, first, that z is a public good that *both* partners can consume

jointly. Then, the only stable assignment is that in which males with high m marry females with high f , and, if there are more (less) eligible men than women, the men (women) with the lowest endowments remain unmarried, a positive *assortative mating*. However, if one assumes, instead, that z can be divided between the two partners, a positive (negative) assortative mating occurs only if the two traits are complements (substitutes). Matching models refer to these two situations as nontransferable and transferable utility, respectively. With transferable utility, a man with low m may obtain women with high f by giving up part of his private share in the gains from marriage. Complementarity (substitution) means that the two traits interact in such a way that the benefits from a woman with high f are higher (lower) for a male with high m than for a male with low m . The type of interaction in the gains from marriage determines the willingness to pay for the different attributes. An important lesson is that in a marriage market, with sufficient scope for compensation within marriage, the best man is not necessarily the one married to the best women, because, with negative interaction, either one of them can be bid away by the second best of the opposite sex (Becker 1991).

The process of matching in real life is characterized by scarcity of information about potential matches. The participants in the process must spend time and money to locate their best options. The realized distribution of matches and the division of the gains from each marriage are therefore determined in an equilibrium which is influenced by the costs of *search* and the search policies of other participants. The main ingredients of the search model are as follows. There is a random process that creates meetings between members of society of the opposite sex. When a meeting occurs, the partners adopt a reservation policy where the match is accepted only if it provides expected gains above some prespecified value. Otherwise, they depart and wait for the next meeting to occur (Mortensen 1988). Because meetings are random and sparse in time, those who actually meet and choose to marry enjoy a positive rent. An important issue is the division of these rents between the partners.

The literature mentions two considerations that determine the division of the gains from marriage: outside options, reflected in the value of continued search, and the self-enforcing allocation that would emerge if the marriage continued without agreement. Combining these two considerations, the sharing rule will be influenced by both the value of search as singles (outside the current match) and the value of continued search during the bargaining process, including the option of leaving when an outside offer (whose value exceeds the value of potential agreement) arrives. In this way, a link is created between the division of gains and the market conditions. For instance, if there is excess supply of women in a particular marriage market, the value of continued search by females is

reduced and consequently their share in the gains from marriage declines.

With friction, there is still a tendency to positive (negative) assortative mating, based on the type of the interaction in traits. If the traits are complements, individuals of either sex with higher endowment will adopt a more selective reservation policy and will be matched, on the average, with a highly endowed person of the opposite sex. However, with sufficient friction, it is possible also to have negative assortative mating under complementarity. The reason for this result is that, because of the low frequency of meetings, males with low m expect women with high f to accept them, and if the gains from such a match are large enough, they will reject all women with low f and wait until a high f arrives (Burdett and Coles 1999).

Generally, one would expect negative sorting on wages and positive sorting on nonwage income, because the gains from division of labor are larger when wages differ, while the gains from sharing goods are larger when incomes are similar (Becker 1991). Empirical findings suggest positive sorting on both wage and nonwage income. In particular, there is a substantial correlation in the schooling achievements of partners to marriage. Similarity in schooling has opposing effects on the gains in marriage; it reduces the gaps in wages and thus the gains from the division of labor, but it also induces similarity in tastes, which facilitates the allocation of public goods. It seems that the latter effect is empirically more important. We should further note that meetings are not really random, and that unattached individuals select jobs, schools, and leisure activities so as to affect the chances of meeting a qualified person of the opposite sex.

Frictions and search explain why, in any given moment in time, part of the population is single, even though there are positive gains from marriage. The aggregate stock of unattached individuals is determined by the rates of entry and exit into marriage implied by the arrival rates of offers and the optimal reservation policies chosen by individuals. These decisions, in turn, depend on the aggregate proportion of singles, because if one meets an unattached person it is more likely that a new union will be formed than upon meeting an attached person, who already has some positive rents. Therefore, the higher the proportion of singles, including divorcees, the higher the private expected gains from divorce. This type of reinforcement can lead to multiple equilibria, whereby both high and low aggregate divorce rates can be sustained through the induced change in the individual incentives to marry and divorce.

The time pattern of aggregate divorces is consistent with multiple equilibria. For instance, the divorce rate in the USA doubled during the decade between 1965 and 1975, standing at roughly 10 percent from 1940 to 1965 and at roughly 20 percent from 1975 to 1995, suggesting a switch from a low to a high equilibrium. Explanations for the timing of the change include the

appearance of the contraceptive pill, break up of norms, and legal changes (Michael 1988). The main conceptual point, however, is that any change in exogenous factors, even a small one, can have marked impact on the aggregate divorce rate, because of the inherent reinforcements in marriage markets.

The marriage market determines not only the assignment of partners but also the division of resources and activities within the family. As market conditions change, a shortage of suitable partners of a particular kind leads to an increase in their gains from marriage. For instance, there is evidence that an increase in the demand for men, created by faster population growth combined with the tendency of men to marry younger women (a marriage squeeze), has led to an increase of dowries in rural India. In modern societies, up-front payments are rare and market forces are mostly revealed by the division of labor within families. The trend of rising female participation in the labor force among married people is probably associated with a larger share that women can extract from marriage, following the rise in female wages. Additional information on the (expected) gains from marriage is contained in the decisions to enter marriage and to stay married. For instance, it has been observed that black women in the USA delay their marriage and have children out of wedlock, because of a shortage of eligible black men (Willis 1999).

5. Some Policy Issues

Despite its firm roots in nature and its antiquity in human society, the future of the family institution has been recently put into question. This was motivated by recent trends, common to all Western societies, of declining marriage rates, declining fertility, higher divorce rates, and a rise in alternative arrangements such as cohabitation, single-parent households, and single-mother families. Social observers view these trends with considerable alarm, especially because of the potential harm to children (more than one million children in the USA are involved in a divorce every year).

Although there is ample evidence that divorce reduces the welfare of single wives and of children with single or step-parents, this is only one part of the picture. Continuation of marriage under adverse conditions can have equally harmful results, although these are harder to identify. Broadly viewed, divorce is a corrective mechanism that enables the replacement of bad matches by better ones. There is a risk that a better match will not be found, in which case the person who has divorced and the children are worse off. But rational agents take this consideration into account and can make financial arrangements to ensure that separations occur only if they improve the welfare of all parties. The issue, though, is whether the courts can enforce such binding contracts and the extent to which they should intervene.

Because meetings are random and uncoordinated, the search behavior and the decision whether to marry or divorce of each individual influence the marriage (and remarriage) prospects of all other members of society who are potential matches. Therefore, legal intervention must take into consideration the impact of the law not only on particular families but also on the marriage market as a whole. These two considerations may be in conflict. As an illustration, consider an increase in the size, or the enforcement, of child-support payments. Holding the aggregate divorce and remarriage rates fixed, each family in isolation may be better off as a result of the additional insurance. However, the reduced propensity to divorce can have a negative effect on the remarriage prospects of those who have suffered a large negative shock to their quality of match and who wish to separate. Another externality that must be considered, and may operate in the opposite direction, is that the availability of alternative mates can reduce investments that are specific to the current match and detract from its quality. This is indeed the logic behind the laws and religious norms aimed at restricting divorce.

The observed changes in family arrangements are closely related to the dramatic changes in labor-market participation of women, and the associated technological advances in the workplace and in the home. Whatever the causal relationships, it is clear that these processes reinforce each other and generate complex dynamics. While the reduction in marriage has been fairly smooth, the changes in divorce appear discontinuous, suggesting search externalities and perhaps the breaking of norms in the processes governing divorce and remarriage. These abrupt changes cause a special adjustment problem that requires social and legal restructuring. However, the expected outcome for the early twenty-first century is that the marriage market will settle at a new equilibrium with higher marital turnover, reflecting the lower gains from marriage as the disparity in the market productivity of men and women gradually declines. Although reduced, the gains from marriage are unlikely to disappear and the family, in one form or another, will continue to have a major role.

See also: Divorce and Gender; Divorce, Sociology of; Family Theory: Competing Perspectives in Social Demography; Family Theory: Economics of Inter-generational Relations; Marriage

Bibliography

- Becker G 1991 *A Treatise on the Family*, enl. edn. Harvard University Press, Cambridge, MA
- Browning M, Bourginion F, Chiappori P A, Lechene V 1994 Incomes and outcomes: A structural model of intra-household allocation. *Journal of Political Economy* **102**: 1067–96
- Burdett K, Coles M G 1999 Long-term partnership formation: Marriage and employment. *Economic Journal, Series F* **109**: 307–34

Lazear E P, Michael R T 1980 Family size and the distribution of real per capita income. *American Economic Review* **70**: 91–107

Michael R 1988 Why did the US divorce rate double within a decade? *Research in Popular Economics* **6**: 367–99

Mortensen D T 1988 Matching: Finding a partner for life or otherwise. *American Journal of Sociology* **94** (suppl.): 215–40

Roth A, Sotomayor M 1990 *Two Sided Matching: A Study in Game-theoretic Modeling and Analysis*. Cambridge University Press, Cambridge, UK

Weiss Y 1997 The formation and dissolution of families: Why marry? Who marries whom? and What happens upon divorce? In: Stark O, Rosenzweig M (eds.) *Handbook of Population and Family Economics*. Elsevier, Amsterdam

Willis R 1999 A theory of out-of-wedlock childbearing. *Journal of Political Economy* **107** (Part 2 suppl.): 33–64

Y. Weiss

Family Theory: Feminist–Economist Critique

1. Feminist Economics

The critique from feminist economics of economic family theory largely has been directed at the version developed in neoclassical economics, often referred to as the ‘new home economics,’ though Marxist economic approaches to the family have also come in for their share of the criticism. The latter approaches largely were explicated in the ‘domestic labor debates’ of the 1980s, with articles appearing on both sides of the Atlantic showing how Marxist concepts could be used to study household production. Both these bodies of family theory were developed mostly in the 1970s, though their origins go back to the nineteenth century and even earlier.

Feminist economics can also be traced back to earlier centuries, but it emerged as a discipline in its own right in the 1970s and has grown most strongly in the 1990s. It is the branch of economics that seeks to understand the economic aspects of the relations between women and men and the gendered aspects of economic activity. With this dual goal of inquiry, feminist economics substantially broadens the scope of human activity to which economics can be applied. While economics is most commonly defined as the study of the allocation of scarce resources to meet human needs, in practice that study has been limited to primarily market transactions. Scarcity implies choice and neoclassical economics generally posits that ‘economic man’ acts rationally to maximize his own well-being, choosing options subject to constraints. Feminist economics suggests broadening the scope to a comprehensive study of provisioning and distribution, or to how society meets all its needs for everyone (Nelson 1993).

This would include housework and home production as well as the labor of caring for others, much

of which is done by women within the family. Caring necessarily intertwines affection and work, and maximization of an individual's self-interest may not be the guiding principle of this type of production, requiring that, at a minimum, the simple neoclassical model be made more complex. Since a monetary measure is absent, it is difficult to determine precisely how much production occurs in the household sphere, but worldwide it has been estimated that one-third of all human production results from the unpaid work of women (UNDP 1995, p. 97). Feminist economists argue that if these aspects of economic activity are not considered, one has a very incomplete view of the economy.

As Peterson and Lewis point out in the introduction to their 800-page compendium on feminist economics (Peterson and Lewis 1999), this is a branch of economics that welcomes diversity and heterodoxy. Feminist economists borrow from many traditions of economic thought as well as insights from other social science disciplines (Kuiper et al. 1995). While some feminist economists work largely within the neoclassical tradition and seek to use its methods to answer feminist questions, others work more within the Marxist framework or institutional economic tradition, while yet others use philosophical, literary, or postmodern methods. Most agree, however, that traditional economics of all schools has failed to explicate the economic aspects of gender relations, specifically the differences in economic power between women and men or the subordination of women to men. Why do women earn less than men, on average, just about everywhere, why do women do most of the family care in virtually all human societies, why do men hold most of the positions of economic leadership as owners, top managers, financiers, and so on?

2. The Neoclassical Theory of the Family

Neoclassical economists did not turn their attention to the question of how economic factors affect decision-making within the family and the allocation of the time of family members among various pursuits until the market started to impinge on the family and to change family dynamics in an obvious way when labor force participation by married women began to grow in the United States in the 1960s. Before this period, the family was thought of mainly as a consumption unit spending income earned by the man of the family, who went outside the family to produce goods and services for the marketplace. As Albelda (1997) notes, this was the typical family arrangement of the white, middle-class men who founded the separate discipline of economics and increasingly professionalized it in the late nineteenth and early twentieth centuries in the United States and Europe.

Becker (1965, 1973) and Mincer (1962) developed a model of the family unit in which men who specialize in market work (at which they are assumed to be

relatively more efficient) and women who specialize in domestic production (at which they are assumed to be relatively more efficient) marry each other and maximize their joint satisfaction in a sort of trade exchange based on their comparative advantages. Each human needs both market goods and home-produced goods to survive (since the goods are substitutable but not perfectly so). If relative prices were to change, women and men would decide to trade different amounts with each other. The division of labor within the family (and the supply of family members' labor time to the market) can change with the male–female wage differential, changes in male–female productivity in housework, and the availability and quality of substitutes, all of which can be affected by technological change. Rising real wages, especially for women, would tend to lead women to work more outside the home, substituting market-based goods and services for their housework time. Fertility will decline as women will want to spend more time at market work and less at home production. Moreover, as technological change occurs and society becomes more complex, the quality of children (their education and socialization) becomes more important than their quantity. More market-based goods (more education and better food and clothing) can be combined with fewer children to yield higher quality output (children with more human capital). In this way the productivity of home production keeps pace with the growing productivity in market-based goods. In the several decades since its origins, the neoclassical theory of the family has been developed by the work of many economists carefully examining such issues as the allocation of time, the supply of labor, the development of human capital, population growth, and economic development.

The neoclassical approach views the household's role in the larger economy and society as the provision of labor to the market and the consumption of final products provided by the market. The household center (or factory) which performs these two functions combines market work, nonmarket work, and consumer products in an economically rational way: the mainspring of the system is the choice of individuals—who, faced with market options and given their own preferences and relative productivities, allocate their time and money accordingly. Thus, household production is the creation and maximization of utility to consumers in the household. It is assumed that the members of the household have harmonious interests which can be maximized by joint effort and decision-making (Hartmann 1974).

3. The Feminist Critique

Feminist economics challenges nearly every aspect of neoclassical economics. First, feminist economics challenges the limited role accorded the family or household, noting that caring labor is a large part of social reproduction and involves not only creating new

laborers by raising children and maintaining adult workers through daily care but also the care of those unable to work, not only the young, but also the sick, disabled, and elderly. Women's unpaid family work also includes maintaining kin networks and relationships, which provide a type of social insurance for the family. Women also contribute volunteer time to school, neighborhood, religious and community groups, which provide necessary services to family members, especially in countries without extensive welfare state services. Elson (1994) refers to all this crucial social reproductive work as the 'caring economy,' a term which brings its importance to the fore. The quantity and quality of the actual labor that takes place in the family is seldom examined carefully by neoclassical economists and thus the family portrait they paint is simplistic and incomplete. Too often women's caring labor is not seen as skilled work that must be learned but rather as an innate ability that women have, or as the manifestation of their natural affection.

Moreover, since in feminist economics caring labor combines affection and work, economic rationality and love, or its withering away, may dictate different 'choices.' A disabled child almost always requires more care and attention than an able-bodied one, yet investing more in such a child who most likely will be less productive hardly seems economically rational; such a 'choice' may be more a matter of affection and love than one of economic rationality. Economic rationality implies 'free' choice, choice that is not fundamentally coerced even though it is faced with constraints. If love or affection has waned, even otherwise enjoyable activities like sex could become oppressive. Women sometimes stay on even in violent relationships because their alternatives seem so limited to them—can such behavior be deemed a free or rational choice?

Second, the neoclassical economics model assumes choice by individuals of different relative productivities exercising their individual preferences, but it fails to explain why domestic labor is most often performed by women. It offers no convincing explanation of why this would be so (other than to assert that women's and men's abilities and preferences differ markedly on average, an assertion that has not been proved). Moreover, one cannot just assume that men's and women's preferences are determined individually. Instead, social roles may rather be collectively determined or politically enforced to such an extent that it is difficult for an individual to express a different preference. Even in the present-day United States, where it is common knowledge that there is increased flexibility and diversity in family structures, a typical married mother may seem to make an individual choice to stay home and care for the children because her wage rate is lower than her husband's, yet the wages women and men face in the labor market are affected by societal norms, including discrimination

against women. As much as she might like to work, given her wage and the cost of childcare, she may have no individual choice to make. Also, the neoclassical model assumes that on the margin men and women can make individual decisions about time allocation, and that an individual could choose to work 33 hours per week, or 17 hours, or 51. In practice, the 'choice' of the number of hours of work is also largely socially determined by the convention of the 40-hour work-week as full-time work. And in some countries, for the most part only full-time jobs come with all the fringe benefits (health insurance, unemployment insurance, and pensions) that families need. Such benefits are difficult to get elsewhere.

Third, it is hard to see whose utility is being maximized, since nothing in economic theory suggests that two adults often have identical utility functions. Sometimes neoclassical economics seems to assume a joint utility function; at other times the possibility of altruism enters, in which, for example, the male head makes altruistic decisions within the family but acts in a completely self-interested way in the market place. In reality it is most logical and quite common that men and women in families often have different interests, preferences, and utility functions, as do parents and children (Hartmann 1981, Folbre and Hartmann 1988, Woolley 1999). Neoclassical economics does not deal well with the possibility of serious conflict in the family.

Fundamentally, feminist economists challenge this model because it justifies the status quo of the gendered division of work, labor and care within the family. It assumes that women are disadvantaged socially, have less access to the economic resources of good wages and property, and are often oppressed by individual men. It, therefore, hides what feminists are most interested in explaining. As Galbraith (1973, p. 25) has put it, neoclassical economists have made the concept of the household 'a disguise for the exercise of male authority.' Albelda (1997, p. 119) writes: 'At the most fundamental level, neoclassical economics argues that women receive lower wages than men and perform more unpaid labor than men largely out of women's and men's own rational choices and desires.'

4. Marxist Domestic Labor Debates and the Feminist Critique

A debate about the unique aspects of production in the home took place within the Marxist economic school of thought in the 1970s, but it was never resolved. Does housework, like wage work, produce surplus value that is appropriated both by capitalists (who can pay workers a lower wage than they otherwise could) and/or by husbands (who appropriate personal services and who have a higher standard of living than they would otherwise)? Does home production constitute a separate mode of production, a feudal mode,

directly appropriated by the lord of the house, rather than a capitalist one based on wage labor?

As Himmelweit (1999) notes in a useful review, this debate was part of a feminist attempt to use Marxist economics to identify the material base of women's oppression. Since the Marxist framework gives prime attention to production, housework (i.e., the unique work of women) was scrutinized. Thus, earlier than in neoclassical economics, the actual content of housework was studied and theorized, but just as in the neoclassical domain, the role of housework had to be discovered by women. Just as the first neoclassical 'new home economics' treated the family as an unexamined black box with a unitary utility function, early Marxist analysis of the family generally suggested that the reproduction of the working class could safely be left to itself (i.e., to the invisible labor of women in the home). Since then the concept of housework as production has been firmly established and the importance of the changing nature of the boundaries between the spheres of domestic production and market production is understood everywhere. But the feminist critique also established that what is unique in women's work in the home to reproduce the next generation is 'its caring and relational aspects' (Himmelweit 1999, p. 133).

5. The Development of a Feminist Model

Using aspects from three major streams of economic thought, namely the neoclassical, institutional, and Marxist approaches, Table 1 describes the attributes of the 'traditional neoclassical; model of the family (or economic decision-making more generally) in column 1. The table also shows how they are modified when institutional perspectives and feminist critiques informed by Marxist and institutional economics are added. (The table has been adapted from Nancy Folbre's (1994) much cited *Who Cares for the Kids?*) In column 2 it shows that neoclassical institutionalists modify the simple model and emphasize that rules, norms, and preferences are not all exogenously given but are at least partially endogenous. As Folbre notes, it is often easier to act based on norms and on what everyone else is doing than to calculate each decision

anew. Individuals also sometimes act collectively. Folbre argues that these modifications come largely from institutional economics, which rejected both individual choice and Marxist class analysis and instead emphasized group behavior by voluntary interest groups, such as labor unions (Figart and Mutari 1999). This model notes that a transaction can involve bargaining and cooperation (based on love and affection) as well as simply buying and selling at fixed prices, particularly within the family, where there are no fixed prices at which a wife can sell her home-produced services to a new husband. Maximization of self-interest exists side by side with altruism in real life. One fruitful area of feminist research within this modified neoclassical framework is the use of game theory and models of cooperative and noncooperative bargaining to elucidate gendered outcomes in the family. (See Seiz 1999, for an overview, critique, and suggested extensions of this work.)

As the third column in Table 1 shows (the 'stylized feminist' description), not only are rules, norms, and preferences structural, but so are the assets with which the different parties begin the transaction. Using the concepts of material base and unequal ownership of assets from Marxist economics, feminist economists note that the genders do not come to the playing field as equals. Feminist economists also note that individuals do not choose all their group memberships, as the institutionalists often assume; many such memberships are assigned at birth. And, as in Marxist economics, important economic transactions are likely to involve coercion and production as well as cooperation and exchange. The relational aspect of caring is captured in the process of cooperation. Thus, in this example of a feminist model, the traditions of neoclassical, institutional, and Marxist economics are all challenged and incorporated in changed forms.

6. Impact of the Critique from Feminist Economics

The emergence of common elements of feminist approaches, from whichever economic framework they start, has contributed to a considerable policy impact of the feminist economic critique. The rec-

Table 1
Models of choice/action/collective identity

	Traditional neoclassical	Neoclassical institutionalist	Stylized feminist
Factors	Rules, norms, preferences (exogenously given)	Rules, norms, preferences (partially endogenous)	Assets, rules, norms, preferences (structural)
Agents	Individuals	Individuals, voluntary groups	Individuals, chosen groups, given groups
Processes	Exchange	Exchange, bargaining, coordination	Coercion, production, exchange, coordination
Sites	Markets	Markets, social institutions	Firms, states, markets, families

Source: Folbre N 1994, pp. 24, 49

ognition of the different work, interests, and utilities of wives and husbands, mothers and fathers, and the durability of unequal gender relations within the family, has informed the policies of economic development and family planning around the world. The success of these policies requires an understanding of women's roles in the family and nonmarket production. Empirical research has shown that mothers use economic resources more for the welfare of children than do fathers both in more and less advanced countries.

Hence empowering women through education and employment is thought to be the surest way to improve the health and education of children and to raise a country's productive capacity in the long run. More economically conservative structural adjustment policies, often fostered by international financial institutions to curb inflation and public deficits, seek to shrink the public sector and lessen wage demands. These policies have been shown to have disproportionately negative impacts on women, who try to make up for the lack of market wages and government aid through more intensive unpaid work. The limits of women's ability to absorb the costs of adjustment necessarily limit the success of these and other macro-economic policies. Feminist economic analysis has led to the understanding that integrating women into paid work or into production for market exchange is not sufficient to overcome gender roles that remain unequal in the family, nor is it enough to call for more grassroots activism to allow affected women to participate in policy discussions (Bakker 1999). It calls instead for 'gender mainstreaming' or putting gender issues at the center of policy changes everywhere, including in international agencies such as the FAO, ILO, and UNDP.

National and international statistical efforts have also become informed by the feminist economic critique of the family model. For the first time, the UN Platform for Action from the Fourth World Conference on Women held in Beijing in 1995 endorses the measurement of all unpaid labor activities, including those for own consumption in the household (i.e., housework). It also endorses their inclusion as satellite accounts in the System of National Accounts for each country (Vanek 1996). In its 1995 *Human Development Report* the UNDP for the first time included measures of work-time, covering both paid and unpaid time, for women and men. Its data for 31 countries show that everywhere women work more hours than men. Beneria (1996) has called for the continuous collection and analysis of such data from all countries to better measure the effects of policy changes, such as structural adjustment policies. In the United States, the Bureau of Labor Statistics is considering a regular, periodic survey of time use (Committee on National Statistics 2000).

Feminist economists call for new measures of family poverty that challenge the assumption of income

pooling and equal sharing within the family (Shaw 1999). In the United States feminist analysis of poverty, single mothers, and welfare reform calls attention to the similarity of women's income packaging strategies across class—women package income from men or other family members, their own labor market participation, and government programs (such as welfare, unemployment compensation, or social security). Just as most middle-class married women now work both inside the home for no pay and outside the home for some pay, most mothers who receive welfare or welfare-state benefits do not rely on it alone but combine income from paid work and from other family members with their government benefits. But we still know too little about who controls income within the family and how resource allocation decisions are made (Spalter-Roth et al. 1995).

7. The Growth of Feminist Economics

The establishment of a list serve for feminist economists (Femecon-L) in 1991; the formation of the International Association for Feminist Economics (IAFFE) in 1992; its annual conferences held in diverse locations around the world; its regular presence as the sponsor and co-sponsor of sessions at the annual meetings of the Allied Social Science Association in the United States and in many other associations in many countries; and the founding of its journal, *Feminist Economics*, first published in 1995, have all contributed to the growing visibility of feminist economists and their ideas (Shackelford 1999). As is characteristic of so much other growth in social science in the past several decades, the development of feminist economics is distinctly diverse in method and approach, distinctly international in scholars, and distinctly global in its impact.

See also: Economic development and women; Family Planning Programs: Feminist Perspectives; Family Theory: Complementarity of Economic and Social Explanations; Feminist Economics; Feminist Theory; Rationality and Feminist Thought; Rational Choice Theory: Cultural Concern

Bibliography

- Albelda R 1997 *Economics and Feminism: Disturbances in the Field*. Twayne Publishers, New York
- Bakker I 1999 Development policies. In: Peterson J, Lewis M (eds.) *The Elgar Companion to Feminist Economics*. Edward Elgar, Cheltenham, UK and Northampton, MA, pp. 83–95
- Becker G S 1965 A theory of the allocation of time. *Economic Journal* 75: 493–517
- Becker G S 1973 A theory of marriage, Part I. *Journal of Political Economy* 81: 813–46

- Beneria L 1995 Toward a greater integration of gender in economics. *World Development* 23(11): 1839–50
- Beneria L 1996 Thou shalt not live by statistics alone, but it might help. *Feminist Economics* 2(3): 139–42
- Committee on National Statistics 2000 *Time-Use Measurement and Research: Report of a Workshop*. National Academy Press, Washington, DC
- Elson D 1994 Micro, meso and macro: Gender and economic analysis in the context of policy reform. In: Bakker I (ed.) *The Strategic Silence: Gender and Economic Policy*. Zed Press and the North-South Institute, London
- Figart D M, Mutari E 1999 Feminist political economy: Paradigms. In: O'Hara P A (ed.) *Encyclopedia of Political Economy*. Routledge, London and New York, pp. 335–7
- Folbre N 1994 *Who Pays for the Kids? Gender and the Structures of Constraint*. Routledge, London and New York
- Folbre N, Hartmann H I 1988 The rhetoric of self-interest: Ideology and gender in economic theory. In: *The Consequences of Rhetoric*. Klammer A, McCloskey D, Solow R (eds.) Cambridge University Press, Cambridge, UK, pp. 182–206
- Galbraith J K 1973 *Economics and the Public Purpose*. Houghton Mifflin, Boston
- Hartmann H I 1974 *Capitalism and Women's Work in the Home, 1900–1930*. Ph.D. Thesis, Yale University
- Hartmann H I 1981 The family as the locus of gender, class, and political struggle: The example of housework. *Signs* 6(3): 366–94
- Himmelweit S 1999 Domestic labour. In: Peterson J, Lewis M (eds.) *The Elgar Companion to Feminist Economics*. Edward Elgar Publishing, Cheltenham, UK and Northampton, MA, pp. 126–35
- Kuiper E, Sap J, Feiner S, Ott N, Tzannatos Z (eds.) 1995 *Out of the Margin. Feminist Perspectives on Economics*. Routledge, London and New York
- Mincer J 1962 The labor force participation of married women. In: Gregg H (ed.) *Aspects of Labor Economics*. National Bureau of Economic Research, New York
- Nelson J A 1993 The study of choice or the study of provisioning? Gender and the definition of economics. In: Ferber M A, Nelson J A (eds.) *Beyond Economic Man: Feminist Theory and Economics*. University of Chicago Press, Chicago, pp. 23–36
- Peterson J, Lewis M (eds.) 1999 *The Elgar Companion to Feminist Economics*. Edward Elgar Publishing, Cheltenham, UK and Northampton, MA
- Shackelford J 1999 International Association for Feminist Economics (IAFFE). In: Peterson J, Lewis M (eds.) *The Elgar Companion to Feminist Economics*. Edward Elgar Publishing, Cheltenham, UK and Northampton, MA, pp. 486–9
- Seiz J A 1999 Game theory and bargaining models. In: Peterson J, Lewis M (eds.) *The Elgar Companion to Feminist Economics*. Edward Elgar Publishing, Cheltenham, UK and Northampton, MA, pp. 379–90
- Shaw L B 1999 Poverty, measurement and analysis of. In: Peterson J, Lewis M (eds.) *The Elgar Companion to Feminist Economics*. Edward Elgar Publishing, Cheltenham, UK and Northampton, MA, pp. 634–9
- Spalter-Roth R, Burr B, Hartmann H I, Shaw L 1995 *Welfare That Works: The Working Lives of AFDC Recipients*. Institute for Women's Policy Research, Washington, DC
- UNDP (United Nations Development Program) 1995 *Human Development Report*. Oxford University Press, Oxford, UK
- Vanek J 1996 Generate and disseminate! The UN platform for action. *Feminist Economics* 2(3): 123–4

- Woolley F 1999 Economics of family. In: Peterson J, Lewis M (eds.) *The Elgar Companion to Feminist Economics*. Edward Elgar Publishing, Cheltenham, UK and Northampton, MA, pp. 328–36

H. Hartmann

Family Theory: Role of Changing Values

1. Introducing the Issue

This article is about the significance of changing values in understanding demographic behavior. Questioning the significance of values may seem odd since in everyday life few people would doubt that values are important. Among social scientists, however, the issue of values still evokes feelings of ambivalence and controversy. Three issues that we think are pertinent to a sociological understanding of the significance of values in family research will be broached. First, the significance of values as such is dealt with, including a discussion of the concept of values. Second, we argue that if values theory is to play a role in family research, it should be able to complement other types of explanations as well as to have an intrinsic quality to explain demographic changes both at the individual and the societal level. Finally, two types of values theories, that is socialization and role theory are merged into one framework that highlights the dynamics of the values-behavior relationship.

A note on terminology: although logically one should write about 'value theory' (without the -s) the plural form 'values' is preferred because 'value' and 'values' are distinct concepts similar to 'medium' and 'media.' In writing about 'values' the totality of values is referred to as ideational factor, and not to 'value' in the sense of the worth of an object.

2. The Significance of Values

In general three types of attitudes toward the concept of values can be found among social scientists: denial, trivialization, and emphasis. Probably the most radical type of denial of the 'significance' of values is to refuse to give any role to 'values' altogether. In most economic theories for instance, 'value' (without the 's') enters the picture as a synonym of 'price.' 'Values' as a cultural factor are rarely considered. This is especially true for neo-classic economists (cf. G. Becker 1981). If ideational factors are allowed in economic models, it is often in terms of tastes, preferences, or needs, and not in terms of values in a sociological sense of the word (cf. infra). But even in sociology some approaches (e.g., structural theories) are notorious for denying the significance of culture (Thompson et al. 1990). Among approaches that explicitly give credit to

a cultural argument, several also raise doubt whether the concept of values is needed in the study of culture and social change (Adler 1956, Blake and Davis 1968; see also J. F. Scott's phenomenological approach (Cancian 1975). They question the usefulness of the concept of values in causal explanations because values can only be inferred from behavior itself or from specific norms and goals. Hence, in this view a direct explanation of human behavior by reference to values is a sociological tautology. Although these critics differ in opinion regarding the type of cultural factors that are significant, they all deny the significance of values.

Trivialization of values may manifest itself in several ways. Values may be trivialized by 'explaining it away,' that is by arguing that values are merely spuriously related to behavior. Such a view is congruent with the former critics regarding the tautological nature of any values explanation of behavior. Values may also be trivialized by making them a 'constant' or 'exogenous' factor in explaining behavior. Within rational choice theory the exogenous character of values is explicit: assuming that there is a rational actor is implying that 'rationality' is a natural characteristic rather than a cultural construct. Defining values as an exogenous factor or arguing that values are merely spuriously related to behavior is minimizing their role. However, maximizing their importance can also lead to trivialization. This is illustrated in the work of Parsons (1937) who—perhaps ironically enough—protested strongly against the aforementioned 'natural science' kind of attitude toward values. Parsons' own functionalist approach to values, however, can also be criticized for having an intrinsically trivial nature (Spates 1983). Part of the trivialization is due to the teleological conceptualization of values within the Parsonian tradition: 'a value is a conception...of the desirable which influences the selection from available modes, means and ends of action' (C. Kluckhohn, in: Parsons and Shils 1951, p. 395). Values determine behavior by definition. By consequence, the question whether they do so is irrelevant. Spates (1983) raises a second caveat, that is that Parsons' values theory faces the problem of deductive imposition by arguing that values can only be deduced by a careful theorist. Hence, observable reality is forced into accord with a preconceived model. Although theoretically values are emphasized, empirically they are trivialized. Dropping the element of action probably leads to a less trivial conceptualization.

The empirical tradition of studying values by definition emphasizes the significance of values. Conceptualization of values, however, has not proved to be its strongest point. Many researchers do not hesitate to use the concept of values implicitly in their work, assuming a mutually intuitive understanding of the concept. Others have claimed that the empirical tradition merely added to the conceptual confusion rather than anything else (Lautmann 1971). Although

there is a grain of truth in these claims, it is felt that they tend to exaggerate divergence. Within the ideational approaches of the concept of values, there is a tendency to fit in with two traditions. ('Fitting in' does not mean that scholars fully subscribe to the conceptualization within a tradition, but rather that they share its point of departure.) The Parsonian tradition has already been discussed in the preceding section. Researchers working within this framework seem to agree with Kluckhohn's idea that values are 'conceptions of the desirable.' The second tradition was initialized by the social psychologist Rokeach (1968, 1973), who explicitly opposed the Parsonian concept of values. Instead he argued that 'a value is an enduring belief that a specific mode of conduct or end-state of existence is personally or socially preferable to an opposite or converse mode of conduct or end-state of existence' (1973, p. 5). The major advantage of this definition is that it tries to avoid letting issues open to empirical questioning—such as reference to action—constitute elements of that definition. Presumably, this is one of the reasons why the two values theories that will be discussed in the last section adopt a similar concept of values. Nevertheless, one could interpret Rokeach's concept of values as being a conservative force since they are defined as enduring beliefs. Although socialization theories (cf. *infra*) tend to agree with the latter statement, Rokeach himself has pointed out that complete stability is not at issue. To avoid misinterpretation it is suggested the adjective 'enduring' is dropped from the nominal definition.

In spite of these arguments several researchers still reject Rokeach's definitions and prefer Kluckhohn's conceptualization (Mohler 1987 and Miceli and Castelfranchi 1989). This merely illustrates that both traditions are competing.

It is obviously important to distinguish 'values' from other ideational factors (Rokeach 1973, Cancian 1975, Fishbein and Ajzen 1975, and Miceli and Castelfranchi 1989). In family research, three such factors seem important, namely, norms, attitudes, and intentions. Especially the latter two concepts have received much more attention than values in studies of demographic behavior. Since there is little agreement on the meaning of these concepts a view is presented on the major differences. 'Values' and 'attitudes' differ because attitudes are more 'object' and 'situation specific' and because values are of a more general nature. Intentions are even more object-related. Consequently, an 'organization of attitudes' transcends the specificity of each attitude and may be interpreted as a 'values orientation.' Analyses of the European and World Values Studies often adopt such a strategy (Ester et al. 1993) by collapsing scores on different attitudes into one or several values dimensions. Hence, attitudes and values are strongly related. Intentions, on the other hand, are more closely related to behavior. For that reason measured intentions are often modeled as dependent variables as an alternative to indicators

of behavior in values research (Inglehart 1977, 1990). At least Fishbein and Ajzen's theory of reasoned action (1975) demurs at such a point of view. Instead they argue that the best predictor of behavior is the intention to do so. A second difference between values on the one hand and attitudes or intentions on the other is that the latter are individual predispositions, whereas values may also be characteristics of depersonalized entities, as in 'the values of the working class.' In such a connection values bridge the gap between attitudes and norms. Norms prevail at the society or macro level as products of cultural inheritance. They refer to prescriptions of how to behave. In fact they include sets of (normative) rules. Internalization of these rules may define values orientations.

3. The Role of Values in Family Theory and Research

The values argument belongs in a broader context concerned with the role of culture in social demography. As far as family research is concerned, one cannot 'neglect' noncultural types of explanation. One should not ignore the impact of G. Becker's New Home Economics (1981) and R. Easterlin's Social Deprivation Theory (1976, 1980) in explanations of demographic change in recent decades. Pollak and Watkins (1993) recently summarized the various perspectives on how cultural approaches relate to economic perspectives and argued that cultural and economic approaches are compatible when they are consistent or equivalent to one another or when cultural explanations attempt to specify the scope of economic rational choice explanations. Preston (1986) and Lesthaeghe and Surkyn (1998) are referred to in particular on this issue. Similarly, Lesthaeghe and Moors (1994) and Lesthaeghe (1998) have demonstrated the usefulness of a frame of reference that includes the complementarity of theoretical perspectives in studying family formation (see *Family Theory: Complementarity of Economic and Social Explanations*). They argue that our understanding of demographic transitions may improve if we avoid rigidly adopting a single grand theory or aiming at enhancing the supremacy of one particular paradigm. Treating theories as distinct but not mutually exclusive views gives rise to a multicausal theoretical framework that can be used to 'explain' historical and geographical variations (Preston 1986, Lesthaeghe 1995) as well as individual differences (Lesthaeghe and Meekers 1986, Lesthaeghe and Moors 1994). Irrespective of these notions, Oppenheimer (1994) remains skeptical of researchers who rely exclusively on the idea that values influence behavior. In her opinion, they face the problem that there is little empirical research that actually tests the presumed causal effect. However, the little evidence that does exist supports

the theory of ideational change. Using panel data, Moors (1997)—on values—and Thornton and co-workers (1992, see also Barber 1999)—on attitudes—clearly demonstrated that ideational factors influence demographic behavior autonomously. Hence, the ability of values theory to complement other types of explanations, its intrinsic quality in explaining demographic changes, and its empirical merits lead us to conclude that values have a role to play.

4. Values Theory?

In empirical research on values two theories developed in the 1970s drew much attention, namely, Inglehart's socialization theory (1977, 1990) and Kohn's role theory (1977, Kohn and Slomczynski 1993). They produced an independent body of research and stirred controversy, but have gained prominence in later decades. In demography Inglehart's theory has been used by the researchers focusing on values (cf. *infra*) because the theory is founded on the notion of ideational change. There have been some modifications of Inglehart's argument, but its core has remained intact. It is founded on two basic premises. First, Inglehart argues that an individual's values priorities reflect his or her socio-economic environment, people attach relatively more importance to relatively scarce objects. This scarcity argument is complemented by a socialization argument that stresses the importance of experiences in the so-called 'formative years.' Inglehart assumes that—in reaching adulthood—values tend to crystallize in the personality. By consequence he 'predicts' a generational pattern of change in values orientations toward increasing preference for autonomy and consequent behavior. The idea that a process of social metabolism (i.e., the succession of generations with distinctive profiles) may generate social change was nothing new to demographers familiar with the work of N. Ryder (1965). By consequence, it has been a small step to link a generational pattern of values change with changes in behavior, for example, rising levels of cohabitation, independent living and divorce, an increase in single parenthood, etc. Inglehart's values theory primarily emphasizes that values influence choices people make. The possibility of reverse causation is not explicitly discussed.

Whereas Inglehart focuses on historical conditions during the formative years, Kohn claims that values are directly influenced by 'here-and-now experiences' at the workplace. In Kohn's theory, values play a mediating role in relating social structure to behavior. As such he primarily focuses on 'contemporary conditions' rather than historical circumstances. The key argument of the Kohn thesis is that the normative requirements of a job generate values orientations which are appropriate to the actual circumstances of

the job. Hence, the relationship between social class and values can be explained by referring to occupational conditions, that is the level of occupational self-direction. Generalizing this point of view one could argue that people learn to value characteristics that are appropriate to their conditions of life. In contrast to Inglehart, Kohn highlights the adaptation of values to changing conditions. At first glance, the theories yield contradictory propositions. On consideration, however, these theories can be reconciled because they bring together different mechanisms that generate values orientations, that is, values may be socialized (Inglehart) as well as learned by taking up social roles (Kohn). Furthermore, even their difference in locus regarding the issue of values stability and values adaptation need not imply contradiction. To Inglehart values influence behavior because values are relatively enduring characteristics of individuals. Although Rokeach (1973) holds a highly similar position, he also demonstrated within experimental designs that values may change due to strong 'incentives,' for example, dramatic or intense experiences (Ball-Rokeach et al. 1984). Kohn's theory suggests that—given the centrality of work in the life of people—the everyday experience in the workplace may induce such change. Consequently, Inglehart's notion of relatively enduring values orientations after the formative years is compatible with Kohn's argument about the significance of temporal occupational conditions. After all, important changes in the life-course may re-affirm values rather than altering them. Young people who value traditional family life may decide to marry for that reason, and may find affirmation of their opinion by the very experience of marriage itself. Hence, even stability forms part of a dynamic process. As a consequence, the key question is no longer whether values change after the formative years, but when (e.g., under what conditions) they do so. By the same token the dynamics of the values-behavior relationship becomes manifest. Values influence the choices people make and the consequences of these choices—in turn—re-affirm or alter the values these people hold. It is important to reconcile both these theoretical perspectives in demography since there is little doubt that changes in family life can be intense experiences. As the research of Thornton et al. (1992, see also Axinn and Thornton 1993) has demonstrated, both historical and contemporary experiences affect demographic transitions. Hence, a life-course approach focusing on the sequential and recursive relationship of values and demographic transitions is most appropriate.

See also: Families and Households, Behavioral Demography of; Families and Households, Formal Demography of; Family and Gender; Family and Kinship, History of; Family, Anthropology of; Family as Institution; Family Processes; Family Theory: Competing Perspectives in Social Demography; Family Theory: Economics of Intergenerational Relations;

Family Theory: Economics of Childbearing; Family Theory: Economics of Marriage and Divorce; Fertility: Institutional and Political Approaches; Fertility: Political and Political-Economic Perspectives; Values, Sociology of

Bibliography

- Adler F 1956 The value concept in sociology. *American Journal of Sociology* 65: 272–9
- Axinn W G, Thornton A 1993 Mothers, children and cohabitation: The intergenerational effect of attitudes and behavior. *American Sociological Review* 58: 233–46
- Ball-Rokeach S J, Rokeach M, Grube J W 1984 *The Great American Values Test. Influencing Behavior and Belief Through Television*. Free Press, New York
- Barber J S 1999 Ideational influences on the transition to Parenthood: Attitudes toward childbearing and competing alternatives. Paper presented at the P.A.A. annual meeting, New York, March 1999
- Becker G 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA
- Blake J, Davis K 1968 Norms, values and sanctions. In: Faris R E L (ed.) *Handbook of Modern Sociology*. Rand McNally and Company, Chicago, pp. 456–84
- Cancian F M 1975 *What are Norms? A Study of Beliefs and Action in a Maya Community*. Cambridge University Press, Cambridge, UK
- Easterlin R 1976 The conflict between aspirations and resources. *Population and Development Review* 2: 417–26
- Easterlin R 1980 *Birth and Fortune. The Impact of Numbers on Personal Welfare*. Basic Books, New York (2nd edn. 1987)
- Ester P, Halman L, de Moor R 1993 *The Individualizing Society. Value Change in Europe and North America*. Tilburg University Press, Tilburg, The Netherlands
- Fishbein M, Ajzen I 1975 *Belief, Attitude, Intention and Behavior. An Introduction to Theory and Research*. Addison-Wesley, Reading, MA
- Inglehart R 1977 *The Silent Revolution*. Princeton University Press, Princeton, NJ
- Inglehart R 1990 *Culture Shift in Advanced Industrial Society*. Princeton University Press, Princeton, NJ
- Kohn M L 1977 (1st edn. 1969) *Class and Conformity. A Study in Values*. The University of Chicago Press, Chicago
- Kohn M L, Slomczynski K M (eds.) 1993 *Social Structure and Self-Direction: A Comparative Analysis of the United States and Poland*. Blackwell, Oxford, UK
- Lautmann R 1971 *Wert und Norm: Begriffsanalyse für die Soziologie*. Westdeutscher Verlag, Opladen, Germany
- Lesthaeghe R 1995 The second demographic transition: An interpretation. In: Mason K O, Jensen A M (eds.) *Gender and Family Change in Industrialized Countries*. Clarendon Press, Oxford, UK, pp. 17–62
- Lesthaeghe R 1998 On theory development: Applications to the study of family formation. *Population and Development Review* 24(1): 1–14
- Lesthaeghe R, Meekers D 1986 Value changes and the dimensions of familism in the European Community. *European Journal of Population* 2: 225–68
- Lesthaeghe R, Moors G 1994 Living arrangements, socioeconomic position and values among young adults: A pattern description for France, Germany, Belgium and the Netherlands 1990. In: van den Brekel J C, Deven F (eds.) *Population and Family in the Low Countries 1994*. Kluwer, Amsterdam

- Lesthaeghe R, Surkyn J 1988 Cultural dynamics and economic theories of fertility change. *Population and Development Review* **14**: 1–46
- Miceli M, Castelfranchi C 1989 A cognitive approach to values. *Journal of the Theory of Social Behaviour* **19**: 169–93
- Mohler P P 1987 Cycles of value change. *European Journal of Political Research* **15**: 155–65
- Moors G 1997 The dynamics of values-based selection and values adaptation. With an application to the process of family formation. Ph.D. thesis Vrije Universiteit, Brussels, Belgium
- Oppenheimer V K 1994 Women's rising employment and the future of the family in industrial societies. *Population and Development Review* **20**: 293–342
- Parsons T 1968 (1st edn. 1937) *The Structure of Social Action*. Free Press, New York
- Parsons T, Shils E A (eds.) 1951 *Toward a General Theory of Action*. Harper, New York
- Pollak R A, Watkins S C 1993 Cultural and economic approaches to fertility: Proper marriage or mésalliance? *Population and Development Review* **19**: 467–95
- Preston S H 1986 Changing values and falling birth rates. *Population and Development Review* **12** (suppl.): 176–95
- Rokeach M 1973 *The Nature of Human Values*. Free Press, New York
- Rokeach M 1976 (1st edn. 1968) *Beliefs, Attitudes and Values. A Theory of Organization and Change*. Jossey-Bass, San Francisco
- Ryder N B 1965 The cohort as a concept in the study of social change. *American Sociological Review* **30**: 843–61
- Spates J L 1983 The sociology of values. *Annual Review of Sociology* **9**: 27–49
- Thompson M, Ellis R, Wildavsky A 1990 *Cultural Theory*. Westview Press, Boulder, Co
- Thornton A 1992 The influence of the parental family on the attitudes and behavior of children. In: South S J, Tolnay S E (eds.) *The Changing American Family: Sociological and Demographic Perspectives*. Westview Press, Boulder, CO, pp. 247–66
- Thornton A, Axinn W G, Hill D H 1992 Reciprocal effects of religiosity, cohabitation and marriage. *American Journal of Sociology* **98**: 628–51

G. Moors

Family Therapy, Clinical Psychology of

The origins of family therapy can be traced to the 1950s, although its historical roots go back further. Several largely independent movements that deal with relationships among family members contributed to the development of family therapy. The oldest of these is the marriage counseling movement while, later, a separate movement toward conjoint marital therapy developed within the psychiatric community. Family therapy emerged from a union of several seminal thinkers from a variety of backgrounds. The main factors influencing its development were the expansion of psychoanalytic psychotherapy to include a broader range of emotional disorders and the application of

psychoanalytic principles to the entire family; the investigation of the role of the family in the development of schizophrenia and other mental and physical disorders; and the growth of the child guidance movement.

1. Historical Perspective of Family Therapy

1.1 Evolution of the Field

Social work, psychiatry, and psychoanalysis in both Great Britain and the United States all made enormous contributions to the development of family therapy. In her influential book *Social Diagnosis* (1917), Mary Richmond was one of the first proponents of studying the whole family and its needs. At the same time, the dominance of the psychoanalytic model in the mental health field focused on the mother–child dyad as the source of psychopathology. This model, with its linear–causal attributions, blamed maternal deficiencies for any childhood disturbance. Adler became the first of Freud's pupils to openly challenge such an explanation of human behavior. He offered an alternative and more socially rooted theory of psychodynamics, organized child guidance clinics in Vienna, and counseled children, parents, and teachers.

By the 1930s, a strong community of analysts had developed in the United States. Among the most influential to emerge was Harry Stack Sullivan. He focused on interpersonal relations in his research into family dynamics of schizophrenia. By the time family therapy pioneers began to experiment in the 1950s, there was a well-established bias toward social explanations among American analysts. This trend was underscored by a paradigm shift that occurred with the development of *General Systems Theory* (GST). GST implies that family therapy is grounded in a set of assumptions about the interchange between individual, family and social processes, which operate according to certain principles that apply to all human systems (Bateson 1971). Family members are inter-related such that every individual affects each other, while the group as a whole affects the first member in a circular chain of influence, making every action in a sequence also a reaction.

1.2 Pioneers of the Family Therapy Movement

John Bowlby's work at the Tavistock Child Guidance Clinic in London during the late 1940s exemplifies the transition from an individual to a family approach. These beginnings of family therapy were tentative in that Bowlby felt he had to 'reassure' readers that family interviews were 'rarely employed.' However, family interviews prior to the initiation of treatment

became routine, and served as a major stimulus to the development of this new approach in the United States. By the end of the 1950s family therapy emerged as a connected movement whose members exchanged correspondence and visits, and began to cite one another.

Nathan Ackerman came to family therapy from the field of child psychiatry. His involvement in a study of mental health problems among the unemployed, along with other experiences, convinced him that both the environment and the psyche could influence emotional problems. Although family interviews began as an experiment for Bowlby, Ackerman saw this as the major form of treatment in child guidance clinics. He saw the family as the unit of diagnosis and treatment, and began sending his staff on home visits to study families.

In 1958 Ackerman published the first book-length treatment of family relationships, *The Psychodynamics of Family Life*. By 1961 family therapists met to prepare the way for the first joint handbook and to found a common journal, *Family Process*. In addition, John Bell's respected monograph, *Family Group Therapy* (1961), constituted one of the founding documents of the profession. Carl Whitaker, who is noted as the most brazen of the founders, is generally credited with having called the first formal meeting in the family therapy movement in 1953. As family therapy developed, he played an integral role, as evident from his membership on the first board of editors for *Family Process*.

Among the pioneers, Lyman Wynne was particularly well prepared by formal training to become a family researcher and therapist. After his medical training in 1948, he embarked on doctoral studies in the Department of Social Relations. As a graduate student, he came into contact with many of the leaders in the field of sociology, social psychology, and social anthropology. Initially, he saw whole families only when individual treatment or joint interviews with the mother and patient were not effective. Gradually, however, he worked out his own theory of the family structure of schizophrenic patients. Murray Bowen, a psychiatrist who specialized in schizophrenia, felt that the parents—in particular, mothers, should be required to live in the hospital with their disturbed child. In 1954 he joined Lyman Wynne at the National Institute of Mental Health, where he established a program for treating families together.

The Palo Alto Group has probably one of the strongest claims to have been at the forefront of the development of family therapy. This group consisted of five richly individual persons, viz. Gregory Bateson, Jay Haley, John Weakland, Don Jackson and Virginia Satir. Bateson, an anthropologist and philosopher by training, and considered by many as the founder of family therapy, wrote important works such as *Steps to an Ecology of Mind* (1971) and *Mind and Nature* (1979). Bateson and his group hypothesized that

family stability is achieved by a feedback system, i.e., whenever the family system is disturbed, it will quickly move toward reestablishing balance/homeostasis. In 1956 this group generated one of the most discussed papers in the history of psychiatry. In *'Toward a Theory of Schizophrenia,'* they introduced the concept of double-bind as the crucial familial determinant of schizophrenia in children. Jackson published a seminal paper on conjoint family therapy, arguing that it was more effective than seeing family members individually (Jackson 1959). At the same time he founded the Mental Research Institute (MRI), and invited Satir to join him. The MRI, compared to the original Bateson group, was more focused on the treatment of families as opposed to research.

The influence of behavioral family therapy can also not be overlooked as it was launched as a specialist area in the late 1960s with the seminal work of Stuart (1969) and Liberman (1970). Behavioral approaches to family therapy can be characterized as goal oriented, a general reliance on principles of learning, directedness in interventions, and preference for short-term, problem-solving involvement.

While the pioneers dominated the 1960s and 1970s, they also saw new innovations. One innovator who has taken his place among the founders as one of the most influential of all family therapists is Salvador Minuchin. In the late 1960s he became the director of the Philadelphia Child Guidance Clinic and convinced Jay Haley, among others, to join him. The Minuchin-Haley team made substantial clinical and methodological contributions. 'Structural therapy' emerged from their interaction, and will be discussed below. In addition to carrying over the communications and systems elements developed at MRI, Minuchin and Haley gave added emphasis to the realignment of counterproductive family coalitions, and tied their theory to a family developmental framework.

2. The First Family Therapy Models

The field of family therapy encompasses many modalities, each with distinctly different ways of conceptualizing and treating families.

2.1 Psychodynamically Oriented Family Therapy

Nathan Ackerman was among the first to integrate psychoanalytic principles with systems theory. He viewed individual functioning as a reflection of constant interactions between the individual, the family, and society. For Ackerman, understanding a person requires an understanding of their intrapsychic processes and interpersonal interactions. Achievement of this goal involves helping family members adopt social roles that are flexible so as to permit role complementarity.

Murray Bowen's *Extended Family Systems Therapy* expands GST beyond the nuclear family. Behavioral disorders are the result of a multigenerational transmission process in which progressively lower levels of differentiation, i.e., independence of self from others, occur as the family's lack of differentiation is transmitted from one generation to the next. The primary goal of therapy is to increase the differentiation of family members. Bowen regards family relationships as involving 'triangles,' and therapy typically includes only two family members (usually the spouses) while the therapist becomes the third member in a 'therapeutic triangle.' As long as the therapist remains objective and neutral (does not become emotionally triangled), his/her presence helps family members resolve the fusion between them and achieve higher levels of differentiation.

Contemporary proponents of psychodynamically oriented family therapy include Christopher Dare, John Byng-Hall, and James Framo.

2.2 Problem-solving Approaches to Family Therapy

2.2.1 Communication interaction approach. This approach to family therapy grew primarily out of research conducted by Gregory Bateson and his colleagues. They recognized the importance of communication in family and individual dysfunction (e.g., the role of 'double-bind' communication in the development of schizophrenia), and developed a school of family therapy based on the principles of communication. This approach incorporates principles derived from information theory, cybernetics, and GST. Family therapists adopting this approach accept a circular model of causality that regards a symptom as both a cause and an effect of dysfunctional communication patterns. Although therapists do not deny that individual intrapsychic factors underlie family problems, the observable interactions between family members are viewed as appropriate targets of therapy. The primary goal of therapy is to alter interactional patterns that maintain the presenting symptom(s).

2.2.2 Structural family therapy. Salvador Minuchin (1974) originally developed structural family therapy through his work with poor, inner-city populations. In his seminal work, *Families and Family Therapy*, he argued that symptoms are the result of family structural imbalances. Structural family therapy focuses on concepts that describe space configurations: for example, malfunctioning hierarchy and boundaries, and maladaptive reactions to developmental and environmental changes. Structuralists use the key notion of 'complementarity' to denote a fit among matching parts of a whole. Symptomatic behavior is

seen as one such part that fits into a dysfunctional whole/organization. For instance, an adolescent's anorexia nervosa may be related to a mutual invasion of the patient's and their parents' territories. In a dysfunctional family, development has been replaced by inaction. When a family is stuck in a rigid arrangement, they cannot solve their problems and continue to grow.

Breaking away from such an organizational impasse requires the mobilization of resources that the family already latently possesses and which are often apparent in a different context. Systemic change, in the structural view, equals an increase in the complexity of the structure—an increment in the availability of alternative ways of transacting. The function of the therapist is to create a context for the family to experience these alternative patterns as accessible, possible, and necessary. What the structural therapist is trying to build through his or her restructuring efforts is more important than what he or she is trying to uncover. The therapist does not emphasize individual change or the prescription of specific solutions. Instead, he/she attempts to reorganize the family structure, or increase its flexibility. Through parental leadership and the creation of clear subsystems and boundaries, more adaptive coping is promoted.

2.2.3 Strategic or systemic family therapy. In addition to the strategic and systems models of the Palo Alto group, other innovative early approaches were Jay Haley (1976) and Cloe Madanas' (1981) problem-solving approach, and the Milan team approach (Selvini-Palazzoli et al. 1978). Haley and Madanas are now most associated with strategic family therapy, which combines the communication/interaction approach with Minuchin's structural family therapy. Haley coined the term 'strategic therapy' when describing the techniques used by the psychiatrist Milton Erickson. Like communication/interaction therapists, strategic therapists emphasize the role of communication in maladaptive behavior, but their focus is on how communication is used to increase one party's control in a relationship. Whereas struggles for control are inherent in any relationship, they become pathological when one or both parties denies his/her intent to control the other person, and when this results in symptomatic behavior. Symptoms are thus seen as interpersonal phenomena rather than intrapsychic events. The focus in strategic therapy is not on understanding the source of problems, but on alleviating current symptoms through altering family organization, hierarchy, and generational boundaries.

The Milan team consisted of Mara Selvini-Palazzoli, Luigi Boscolo, Gianfranco Cecchin, and Giuiana Prata. Their book, *Paradox and Counterparadox* (1978), presented the conceptual framework

and methodology of their systemic family therapy. Stimulated by Bateson's writings (Bateson 1971), the Milan group radically revised their analysis of systems. Until this point, they had seen the therapeutic process as one wherein the therapy team observed the family system in order to understand it. Working as a team, utilizing a one-way mirror, circular interviewing, and employing systemic analysis to describe the family's interactional map, interventions were designed to ensure that the therapist remains separate from the family. This analysis and approach was labeled 'first-order cybernetics.'

Through revisiting Bateson's work, the Milan clinicians began to take into account the context of observation itself, and hence the observational system as well. In first-order cybernetics a distinction is arbitrarily drawn between the family and the therapist, as though they were two distinct systems. The Milan clinicians began to see the family in the context of the therapist's interaction with them. This new awareness included a second level of observation that parallels the first. This second level of observation is that of the system created by the interaction between the therapist and family within the therapeutic encounter. This second level observation was termed 'second-order cybernetics.' Instead of only studying the interactional patterns of family members, the feedback loops that develop between interviewer and interviewees were also considered.

3. Recent Family Therapies

The 1980s ushered in a new stage in the development of family therapy, in part inspired by the Milan group. First, an era of skepticism and a reexamination of the old, under the banner of postmodernism, emerged. Postmodern psychologies concern themselves with how people make meaning in their lives; how they construct reality. Constructivism (how an individual creates his/her own reality) and social constructionism (how social interaction creates reality) took hold of family therapy in the early 1980s and have since exerted a powerful impact on the field (see *Constructivist Psychotherapies*). Second, a movement to deconstruct established knowledge expanded in the form of feminist critique. This exposed the patriarchal bias in the assumptions and practices of many family therapies (Hare-Mustin 1986). Third, postmodern family therapy became more collaborative, which led to a cross-fertilization among therapists, as well as between therapists and families. Fourth, the mechanistic aspects of systems thinking have become less popular. Instead the focus has shifted to the stories that govern a family's life rather than its structure or interactional sequences of behavior. Fifth, postmodernists have become less wedded to universal patterns, and emphasize diversity and pluralism.

The result of this review has been the realization that a favored therapeutic approach may only be one of many useful approaches. The field of family therapy has also shifted its view of the family as the source of symptoms to a biopsychosocial view. From this perspective, assessment and intervention attend to the impact of biologically based conditions as well as larger cultural influences, such as race, class, gender, ethnicity, religion, and sexual orientation on the family. Family therapy today is characterized as having a resource perspective as opposed to concentrating on deficits within the family. It has also shifted from a model of the therapist as expert to one of respectful collaboration with family members. Assessment and intervention have shifted their emphasis on problems and how they are maintained to the attainment of solutions (Walsh 1993).

3.1 Postmodern Approaches

Growing out of the strategic/systemic models, the post-Milan or 'second-order' family therapists invented techniques to engage families in 'conversation' about their problems; these are known as narrative (White and Epstein 1990) or conversational approaches (Anderson and Goolishian 1988). These therapies are based in constructivist and social constructionist views of reality. Normality is viewed as socially constructed. In narrative therapy, for instance, the therapeutic conversation and process of 'restorying' a problematic experience is emphasized. The goal is to envision optimistic life stories that are more empowering and satisfying. Another constructivist approach that has gained popularity is Steve de Shazer's solution-focused therapy (Berg and de Shazer 1993). This is similar to the narrative approach in that it uses leading questions to separate clients from their problem-focused stories. The therapist is affirmative in that he/she deconstructs the problem story by having clients focus on exceptions, i.e., times when their problem did not occur. No judgments are made about the solutions a family may discover. Goals in these therapies are achieved by externalizing problems, focusing on future-oriented potential, and the search for exceptions and unique outcomes.

3.2 Psychoeducational Approaches

The psychoeducational model was developed for family intervention with schizophrenia and other persistent mental and physical illnesses (Falloon et al. 1984). In schizophrenia, for instance, family management of the illness has been demonstrated as effective in reducing levels of 'Expressed Emotion' and lowering rates of relapse. Family therapy, combined with drug maintenance and social skills training, has produced the best results by reducing relapse rates and improving functioning. The process of change is

facilitated through information and management guidelines, social support, and a respectful collaboration between therapist and family.

4. Research in Family Therapy

The roots of family therapy are in research, as many founders came to this field through studying interactional patterns associated with problem families. Initially, no distinctions were drawn between researchers and therapists, as research had direct clinical relevance, and hypotheses were developed and tested in clinical settings. The growth in research depended disproportionately on intuitive appeal as opposed to evidence. This research was mainly qualitative and impressionistic in nature.

Since the early investigations, research in family therapy has grown dramatically and undergone at least two major transformations (Sprenkle and Moon 1996). First, the field has evolved from an impressionistic to a quantitative emphasis. This shift forced family therapists to operationalize less well-defined concepts, and to develop more reliable and valid measures. The early decades were characterized by a strong push for outcome research and, although the scientific rigor of some of the earlier studies may be questionable, it did give credence to an emerging discipline. Second and more recent, a transformation that is characterized by a shift from quantitative methods to the incorporation of qualitative methods can be discerned. This transformation was facilitated by the advent of the new epistemology constructivism, or social constructionism. Linear causality, assumed by traditional quantitative research, has been replaced by more systemic thinking. Proponents of the second transformation argued that family therapy concepts were operationalized before being truly understood and, consequently, seemed removed from clinical reality. Therefore, the second transformation emphasizes the 'context of discovery,' i.e., the intricate subtleties of the therapeutic process, as opposed to the 'context of verification.' The integration of outcome and process research that has characterized the first and second transformations, respectively, has gained the support of family therapy researchers. The Maudsley group in London exemplifies this integration with their family treatment studies for eating disorders (Eisler et al. 1997).

This 'third transformation' in family therapy research is called pluralism. A pluralistic research community values and appreciates different inquiry paradigms, methodologies and designs, and encourages an atmosphere of understanding, openness, and tolerance. Methodological pluralism encourages research from a wide variety of perspectives, and increases the tools at our disposal for understanding family therapy processes and outcomes (Gurman et al. 1986).

5. Future Directions

Family therapy is generally effective, and family treatments that focus on specific problem areas have been shown to have positive and enduring effects. If treated systemically, problems with clear family relations implications show a greater likelihood of a good outcome. The growing *diversity* of families poses new challenges for the family therapy field. Culture, class, race, ethnicity, sexual orientation, and religion contribute greatly to family beliefs and practices (Falicov 1995). A broader definition of normalcy has to be considered if family therapy is to remain responsive to this pluralism. Recent shifts from action to narrative, and from challenging to collaborating, have opened new possibilities for doing family therapy. Family therapy has developed greater humility about the influences of existing theoretical models, diminished the need to prescribe to and change people, and increased trust in a family's own resources.

See also: Child Abuse; Child and Adolescent Psychiatry, Principles of; Counseling Psychology; Divorce and Children's Social Development; Family and Psychiatric Illness/Family Therapy; Family as Institution; Family Health; Family Processes; Family Theory: Role of Changing Values; Marital Interaction: Effects on Child Development; Parenting: Attitudes and Beliefs; Parents and Teachers as Partners in Education

Bibliography

- Anderson H, Goolishian H A 1988 Human systems as linguistic systems: Preliminary and evolving ideas about the implications for clinical theory. *Family Process* **27**: 371-93
- Bateson G 1971 *Steps to an Ecology of Mind*. Ballantine, New York
- Berg I K, de Shazer S 1993 Making numbers talk: Language in therapy. In: Friedman S (ed.) *The New Language of Change*. Guilford Press, New York
- Eisler I, Dare C, Russell G F M, Szmukler G I, Le Grange D, Dodge E 1997 Family and individual therapy in anorexia nervosa: A five-year follow-up. *Archives of General Psychiatry* **54**: 1025-30
- Falicov C J 1995 Training to think culturally: A multi-dimensional comparative framework. *Family Process* **34**: 373-88
- Fallon I R H, Boyd J L, McGill C W 1984 *Family Care of Schizophrenia: A Problem-Solving Approach to the Treatment of Mental Illness*. Guilford Press, New York
- Gurman A S, Kniskern D P, Pinsof W M 1986 Research on the process and outcome of marital and family therapy. In: Garfield S L, Bergin A E (eds.) *Handbook of Psychotherapy and Behavior Change*, 3rd edn. Wiley, New York
- Haley J 1976 *Problem-Solving Therapy*, 1st edn. Jossey-Bass, San Francisco, USA
- Hare-Mustin R T 1986 The problem of gender in family therapy theory. *Family Process* **26**: 15-27
- Jackson D D 1959 Family interaction, family homeostasis, and some implications for conjoint family therapy. In: Maserman

- J (ed.) *Individual and Family Dynamics*. Grune and Stratton, New York
- Liberman R 1970 Behavioral approaches in family and couple therapy. *Journal of Orthopsychiatry* **40**: 106
- Madanas C 1981 *Strategic Family Therapy*. Jossey-Bass, San Francisco
- Minuchin S 1974 *Families and Family Therapy*. Harvard University Press, Cambridge, MA
- Selvini-Palazzoli M, Boscolo L, Cecchin G, Prata G 1978 *Paradox and Counterparadox*. Aronson, New York
- Sprenkle D H, Moon S M 1996 Toward pluralism in family therapy research. In: Sprenkle D H, Moon S M (eds.) *Research Methods in Family Therapy*. Guilford Press, New York
- Walsh F 1993 Conceptualization of normal family processes. In: Walsh F (ed.) *Normal Family Processes*, 2nd edn. Guilford Press, New York
- White M, Epstein D 1990 *Narrative Means to Therapeutic Ends*, 1st edn. Norton, New York

D. le Grange

Farmland Preservation

One might think that the escalation of the world's population and widespread famine would supply an ample reason for the preservation of agricultural land in the better-off countries where agricultural production is also the most know-how-intensive and efficient. But in fact, wherever conversion to urban development is more lucrative than the income from farming, agricultural land is in danger of losing out, unless special public policies are installed. This almost axiomatic conflict means that farmland preservation is a concern of every industrialized country. The relentless expansive energy of cities often affects precisely those areas that are most attractive to farming.

But despite the importance of preserving agricultural land from a worldwide perspective, the stimulus for farmland preservation in advanced economy countries is increasingly tilting away from concerns about agricultural production *per se*, towards environmental, landscape heritage, and quality of life concerns for the urban residents. Bruce (1998) shows how the interests and needs of the farmers themselves have receded almost to disappearance in the discourse about farmland preservation, while environmental and quality of life goals have risen. Moreover: the very goal of preserving farmland is not beyond dispute. It is sometimes challenged by competing goals, some stemming from the changing economics of particular types of agriculture, others from social equity policies or overriding national priorities, such as the supply to affordable housing. In the countries across the Atlantic, such goal conflicts are more intense than in North America.

Beyond the shared characteristics among advanced economy countries, there are significant differences in the specific goals and means for farmland preservation and in the degree of success achieved. While focusing mostly on the USA, this article takes a cross-national view of farmland preservation goals, means, and outcomes in selected advanced economy countries (see Table 1). The author bases this article on research about farmland preservation laws, policies, contexts, and degrees of success in the USA, Canada, the UK, France, The Netherlands, and Israel, and in more general terms, also Germany and Sweden. For a detailed exposition see Alterman (1997).

1. Comparing the Contexts

The conflict over land use has much to do with degree of urbanization, economic well-being, population

Table 1
Comparative background statistics of selected advanced-economy countries, 1999

Country	Population (in millions)	Average annual growth 1990–2000 (%)	Percent of population Urban	GNP per capita (US \$ current prices)	Fertility rate per woman	Surface area (sq. km)*	Density (residents per sq. km)
USA	272.0	0.91	76	29,080	2.0	9,363,520	29
Canada	30.3	1.14	77	19,640	1.6	9,970,610	3
France	58.5	0.41	75	26,300	2.2	551,500	105
UK	58.5	0.22	89	20,870	1.7	244,100	240
Germany	82.1	0.41	87	28,280	1.3	356,733	229
Netherlands	15.6	0.55	89	25,830	1.5	40,844	378
Sweden	8.9	0.41	83	26,210	1.6	449,964	20
Israel	6.0	2.88	91	16,180	2.7	21,056	286

Sources: The Economist World in Figures 2000, The Europa World Yearbook 1999, UN Statistical Yearbook 1999, Statistical Abstracts of the United States 1999, Israel Statistical Yearbook 1999.

growth, and population density. In the advanced economy countries, where the Gross Domestic Product per person is relatively high, growing food for subsistence is not a major problem because there is enough wealth to supplement what is locally grown with imported food. Local food self-sufficiency is not yet an item on the agenda of most local land-use planning agencies, though some have argued that it should be (see Pothukuchi and Kaufman 2000). As a group, these countries contrast sharply with less developed countries or East European countries, where the figures for annual GDP per person would typically run in the hundreds or very few thousands. The advanced economy countries also differ significantly among themselves (see Table 1).

The USA and Canada are large countries and less densely inhabited than most European countries. In view of their sophisticated farming technology and know-how, the USA and Canada not only supply most of their internal needs for food and fiber, they are major food exporters on which many other countries depend. While these two countries share some aspects of farmland policies, they differ on others. Canada has by far the lowest population density among the countries surveyed. The US population density of 27 per square km, and Canada's 3 (only Australia and Iceland are Western countries with similar densities), are low compared to most countries across the Atlantic. Although most of Canada's land area is quite inhospitable climatically and its 8 percent agricultural land is concentrated along a thin strip along the US border, the area of farmland per Canadian is considerably higher than in the US.

Crossing the Atlantic brings us to a different world of generally smaller farms than in the USA or Canada, much higher population densities, and distinctly more compact cities. France, with a lower population density than most continental European countries, has 105 persons per square km, while at the other extreme The Netherlands has 378 persons. Israel, though currently less densely inhabited than the Netherlands, is in the long term the densest because it has the highest population growth and fertility rates among advanced economy countries.

Even though it has less farmland per person than North America, Western Europe faces a reverse problem: agricultural surpluses endanger the maintenance of price levels and have led the European Union to adopt policies of compulsory 'set asides' of farmland. These pose a clear challenge to the very notions of farmland preservation.

There is no agreement cross-nationally about the size of farming units that should be the target for preservation. In the USA, hobby farming by wealthy ex-urbanites is regarded as a threat to commercial farming. Europeans, who are concerned more with preserving the visual and environmental character of the traditional countryside, tend to focus more on encouraging 'family farms.' Although there is also a

concern in the USA with supporting the family farm, the status of such farms is probably more secure there than across the Atlantic in view of the much-larger average size of American family farms.

2. *How Public Awareness and Planning Policies Emerged*

In both the USA and Canada, farmland preservation gained public support only in the 1980s; later than in most Western European countries. The landmark US National Agricultural Lands Study published in 1981 arose out of this growing concern and propelled it further. That study found that farmland conversion increased almost three-fold between the 1960 and the 1970s, rising from 1.1 million acres a year in the 1960s to 3 million acres a year in the 1970s. In Canada, policymakers' awareness about the depletion of farmland came approximately at the same time as the USA, and the major provincial programs, Ontario, British Columbia, and later Quebec, ensued (Reid and Yeates 1991).

In the USA and Canada, policies and methods for farmland preservation were a relative novice to the land-use planning and implementation kit of tools. No prescriptive formulae with proven results existed in the professional literature or in practice. Farmland preservation was not part of the routine work of local or state governments; someone had to 'invent' them. In the USA, the process whereby these tools emerged and tested (which, incidentally, is quite typical of other areas of land-use planning and implementation) may be called a 'survival of the fittest' process whereby more successful local policies survive and gain national recognition. While such policies were being developed and disseminated sporadically across local and state authorities in the 1980s, several 'how to' publications appeared that collated the emerging 'kit of tools' for planners and decision makers (for a listing of these see Alterman 1997). Scholarly evaluation of the degrees of success and impacts of alternative techniques has taken longer, and the jury is still out about some of them. In the Western European and Israeli context, where regional and national land-use planning is generally better entrenched in national legislation and tradition (for a comparative analysis of the broader land-use planning contexts in these countries see Alterman 2001), farmland preservation methods may therefore have been more 'natural.' That is not to say that they were more successful. Along with the general de-legitimization of direct government intervention, gradual changes occurred in the type of tools used for farmland preservation.

The other side of the coin of farmland preservation is containment of urban development and prevention of sprawl. This concern arrived in the USA and Canada much later than across the Atlantic where

compact cities are the tradition. American planners had to invent a special term, 'growth management' (and since the mid- 1990s, also 'smart growth') to discuss what is a routine part of land-use planning in much of Europe. 'Growth management' describes an open-ended set of policies and methods to phase and control the growth of cities and suburbs and to rationalize infrastructure investments. This approach often also prescribes a more coordinated institutional structure for land-use and fiscal planning (several of many titles on growth management include De Grove 1992, Stein 1993, Daniels 1999. A full review of growth management and 'smart growth' law and policy is presented by I Kushner 2000). However, only a minority of US states has seriously adopted growth management policies and tools.

3. Farmland Preservation Policies

With 50 states and innumerable localities, generalization about US policies is difficult. When compared with the other countries surveyed, American farmland preservation policies are characterized by a rich variety, innovativeness, and capacity for experimentation, but usually with only lukewarm effects.

A distinctive attribute of American farmland preservation policies is that planners and policymakers usually view them as a package of tools that cut across policy areas. By contrast, in some of the countries across the Atlantic, there is or was a dominant method of preservation usually based on top-down control that restricted the authority of lower-level governments to grant permits for conversion of farmland. The American-style package of tools offers not only land-use controls directly regulating the conversion of farmland to development, but also a variety of other tools such as taxation and developer exactions.

Direct control of land-use conversion from farming to non-farming use seems to be the most intuitively obvious tool. However, it does not hold center stage in the USA except in a few states, foremost Hawaii and Oregon. The most stringent version, called 'exclusive zoning,' restricts all uses of land except for farming or farming-related construction, but it is the least common of farmland preservation tool in the USA (Coughlin 1991, American Farmland Trust 1997). It is more prevalent in the countries cross the Atlantic, and ostensibly the most restrictive in Israel. In all the countries studied, exclusive zoning has been receding in importance and effectiveness. Yet in Oregon, where it is a major tool and is applied on a statewide level through a state institution, it has been found to be reasonably successful (Howe 1993, Nelson 1992). Much more common across the USA is a more flexible form of farmland conversion control, called 'non-exclusive zoning.' It permits many other land uses and puts few hurdles before conversion. Usually applied

on the local and county levels only, its effect is often quite lukewarm (Coughlin 1991).

Another direct control method that is so characteristics of British and some other Western European land-use policies is an effectively regulated 'greenbelt' around urban areas (an in-depth analysis of British greenbelt policies and their effectiveness is by Elson et al. This approach is not as common or effectively used in the USA. For example, the city of Madison, a mid-size Midwest city, discussed such a concept but interpreted it as the designation of several narrow one to three-mile buffers in a few spots along the circumference to be used as golf courses.

The dominant tools for farmland preservation in the USA are those that focus on the economics of conversion rather than seeking to limit the permission to build directly. These tools hope to defer the time when farming will no longer out-power real estate, but not to prevent it altogether. Among the most common are property or other tax relief programs such as existing use appraisal rules (where farming land is assessed for its value for agriculture, even if its real estate value has risen steeply). Also common are tax deferral programs, sometimes associated, as in Wisconsin, with agreements drawn with developers to continue farming for a preset number of years. Most widely used by all states are the so-called 'right to farm' laws that provide farmers immunity from lawsuits for causing a nuisance related to agricultural work. However, fiscal means are only mildly effective because once farmers feel that it is lucrative to sell, they do not deter conversions (see Nelson 1992).

Another economic tool, minimum lot size ordinances, aims to discourage suburban housing and thus defer conversion by making the land too costly for most population groups. However, research findings have shown that it is difficult to determine what is a viable cutoff line (Nelson 1992). Florida's infrastructure concurrency requirements are intended to render development in outlying areas less lucrative, but in fact at times work in the reverse.

Property-rights based tools (for which the USA is probably the world leader) are a promising set that attempt to prevent development by buying up the very right to develop, or in effect, by compensating for the farmer's willingness to forgo the opportunity for development. These include 'land trusts' by public agencies or, most commonly, voluntary sector organizations that purchase the development rights or easements from farmers while farmers retain the right to continue farming (Endicott 1993, American Farmland Trust 1997). Transfer of development rights is potentially a superb 'made in America' tool. In an area targeted for preservation, development rights are 'purchased' by landowners in another area, tagged a 'receiving area,' who wish to increase their development rights. However, research has shown that this tool can be used only in very particular market and land-use situations, otherwise, it is ineffective (see

Pruetz 1997). A simpler tool that can be effective in some situations is cluster zoning whereby development rights are concentrated in one part of a larger land area (see Daniels 1998). Significant in their absence from the typical USA package of tools (but used very selectively in the other countries studied) are those whereby governments directly intervene in land tenure. They do this either through expropriation ('condemnation') of land, consolidation of lots that are too small for viable agriculture (prevalent in France), redistribution of lots that are too large, or most ambitiously, 'land readjustment' that redistributes development rights more justly while preserving targeted land (used in Israel in spot cases). Also rare in the USA but prevalent in The Netherlands is land banking as an instrument for timing the conversion of farmland.

4. Levels of Government and Legal Constraints

An important variable in the effectiveness of farmland preservation efforts is the level of government involved. From an international point of view, the USA stands out in the extreme degree of freedom held by local governments in most US states. Despite the 'smart growth' movement that has called for greater state action on land management, in most states, especially those in the heartland, the fate of farmland is still largely in the hands of local government or local civic groups. An important factor is the extensive self-incorporation and annexation powers held by local residents and local governments in many states. A majority of landowners in a particular tract of land who wish to annex themselves to an urban jurisdiction can initiate such a procedure. Although an urban municipality is empowered to refrain from rezoning agricultural land, local interests often achieve both annexations and rezoning.

In a few states, often called the 'growth management' states, the state government does play a major role in farmland preservation. Oregon, recognized as the flagship state for preservation, has developed an especially rich package of policies. Compared with the national rate of farmland conversion, Oregon has been quite successful through a combination of exclusive agricultural zones implemented by a special state-level commission, alongside policies for containing urban growth. But even this comparative success is limited to only a few farm-size categories, not all (as shown by Nelson 1992, Howe 1993). Hawaii is a second flagship state. It is unique in that it has full statewide zoning, including agricultural, that covers approximately half the state. But Hawaii has been only mildly successful in preventing conversions to urban use and, indeed, the conversion rate has doubled as the farm economy faltered (Callies 1994 describes the Hawaiian system of land-use

controls. Empirical findings can be found in Ferguson and Khan 1992). Most states do not have state-level agricultural zoning. The Wisconsin tax relief and agreements program, while less illustrious than Oregon and Hawaii's, is a more typical degree of state involvement. Research into its implementation in Dane County shows little effectiveness and rampant circumvention.

What of the US federal level? National-level involvement in farmland conservation in the USA has been strikingly weak when compared with other environmental areas such as air, water, wildlife, coastlines, and other environmental areas. An attempt to enact a Federal Land Use Act was aborted in the 1970s and never reattempted (see Kayden 2001). The almost nonexistent federal role in farmland preservation (on the books is the US Farmland Protection Policy Act of 1981 oriented to minimizing the extent to which projects initiated by the federal government contribute to farmland conversion, but it has been called 'another case of benign neglect' because of its failure in implementation; see Ward 1991) contrasts with the UK, The Netherlands, and Israel (but interestingly, not France), where legislation has given central government a major role (however, the contents of the preservation goals themselves are changing across the Atlantic as well; see Alterman 1997).

A major legal impediment to farmland preservation would be if a designation of land for farming use were to be legal grounds to claim compensation from government. In none of the countries studied across the Atlantic would denial of an application to convert farmland currently designated as agriculture constitute a good ground for claiming compensation. The USA is a possible exception, but even there, the likelihood of a successful claim is very small. The inherent fear of all American planners is the 'taking issue': the constitutional limitation on excessive regulation that might be ruled by the courts to be tantamount to a taking of property without just compensation. Since the law on this issue in the USA is perpetually inconclusive and farmland preservation has not yet come up directly before the US Supreme Court (Myren 1993), there is still uncertainty about whether total denial of marketable development rights on farmland could be ruled a taking. In the past, several scholars have expressed optimism that restrictions on development would not likely be regarded as a 'taking' so long as some economically viable use is left. However, the 1992 US Supreme Court decision regarding 'Lucas' (1992) may have made the situation somewhat more uncertain; see Mandelker 1993, Lazarus 1992. The court applies a severe *per se* rule to any situation where no 'economically viable use' whatsoever is left, but leaves the law vague regarding situations of near total takings. So long as agriculture is economically lucrative, past optimism has no reason to recede. But it might prove to be less secure if farming economics were to render the 'beneficial use'

assumption tenuous, especially since farmland preservation often involves additional strata of regulations regarding runoff water, soil erosion, fertilizers, etc.

5. *The Relationships between Methods and Degrees of Success*

Achieving effective preservation is apparently much more complex and elusive than simply selecting from among the set of tools outlined above. Research shows that there is little direct relationship between the degree of success in preservation and the ostensible stringency of the tool applied (the research findings are presented in Alterman 1997). The two greatest success stories in farmland preservation among the countries studied—The Netherlands and Britain—do not primarily rely on direct control of farmland conversion. Furthermore, success in preservation does not depend on the characteristics of the broader planning system. The UK and The Netherlands have planning systems that, within the European vintage, are very different from each other (for a comparative analysis of national land-use planning in 10 advanced economy countries, including those analyzed in this article, see Alterman 2001), yet both are very successful. Nor is success intimately tied with the question of whether compensation for ‘taking’ is due. Most of the countries surveyed—indeed, all but possibly the USA—do not grant landowners compensation rights for denial of development permission on farmland, yet they exhibit widely varying degrees of success in preservation. In the USA there are significant variations between states in degrees of success in preservation, even though takings law in its constitutional aspects, is basically similar.

What, then, are the secrets of success in farmland preservation? First, the laudable achievements of the UK and The Netherlands could not have come about had the doctrine in these countries continued to focus on the preservation of agricultural land *per se*. In both The Netherlands and the UK, the concept of farmland preservation today draws much less attention than in the USA and Canada, despite the former countries’ higher population densities and smaller area of farmland per capita. Rather, in the UK and The Netherlands (and belatedly in Israel) there has been an overt redefinition of farmland preservation as countryside preservation. Second, both countries have achieved this outcome through national level policies that are planning driven, shared and applied effectively by local and regional planning authorities, and are less reactive to incremental developer initiatives than in the USA. Third, the UK and The Netherlands focus on the containment of growth of urban areas through infill and higher densities, many times higher than those prevalent in most US cities and suburbs, even where ‘smart growth’ is adopted. Fourth, and perhaps

most crucially, in both countries, preservation policies enjoy wide public support among electoral constituencies. Finally, planning policies in both countries are intolerant of sporadic exurban initiatives. They centrally determine the number and size of local governments, and notions such as ‘bottom up’ annexation initiatives are unthinkable. By contrast, since the 1980s, France, though relatively successful in the use of higher densities, is slack on local government management capacity and on national and regional planning. Now that the old tenure related preservation mechanisms are becoming economically obsolete, France is likely to achieve only lukewarm success in preservation.

Our most distinctive success story, The Netherlands, demonstrates that reliance on agricultural land conservation can no longer assure countryside conservation. The two are separate goals: economics may at times unite them, but will increasingly dissociate them, and may even place them at direct odds with each other by making industrialized farming in buildings lucrative. Countryside preservation must be considered in its own right.

See also: Agriculture, Economics of; Environmental and Resource Management; Environmentalism: Preservation and Conservation; Food Security; Green Revolution; Greening of Technology and Ecotechnology; Nature, Concepts of; Environmental and Ecological; Rural Planning; General

Bibliography

- Alterman R 1997 The challenge of farmland preservation: Lessons from a six-nation comparison. *Journal of the American Planning Association* 63(2): 220–43
- Alterman R 2001 National-level planning in democratic countries: A comparative perspective. In: Alterman R (ed.) *National-Level Planning in Democratic Countries: An International Comparison of City and Regional Policy-making*. Liverpool University Press, Liverpool, UK, pp. 1–42
- American Farmland Trust 1997 *Saving American Farmland: What Works*. American Farmland Trust Publication
- Bruce M 1998 Thirty years of farmland preservation in North America: Discourses and ideologies of a movement. *Journal of Rural Studies* 14(2): 233–47
- Burchell R W, Listokin D, Galley C C 2000 Smart growth: More than a ghost of urban policy past, less than a bold new horizon. *Housing Policy Debate* 11(4): 821–79
- Callies D L 1994 *Preserving Paradise: Why Regulation Won't Work*. University of Hawaii Press, Hawaii
- Coughlin R E 1991 Formulating and evaluating agricultural zoning programs. *Journal of the American Planning Association* 57(2 Spring): 183–92
- Daniels T L 1998 *When City and Country Collide: Managing Growth in the Metropolitan Fringe*. Island Press
- DeGrove J M, Miness D A 1992 *The New Frontier for Land Policy: Planning and Growth Management in the States*. Lincoln Institute of Land Policy, Cambridge, MA

- Endicott E (ed.) 1993 *Land Conservation through Public/Private Partnerships*. Lincoln Institute of Land Policy, Cambridge, MA
- Elson M J, Walker S, Macdonald R 1993 *The Effectiveness of Greenbelts*. The Stationery Office Books, HMSO, London
- Ferguson C A, Akram K 1992 Protecting farm land near cities: Trade-offs with affordable housing in Hawaii. *Land Use Policy* (Oct): 259–71
- Gale D E 1999 Land use planning, environmental protection and growth management: The Florida experience. *Journal of the American Planning Association* 65(3): 344–5
- Howe D A 1993 Growth management in Oregon. In: Stein J M (ed.) *Growth Management: The Planning Challenge of the 1990s*. Sage Publications, Newbury Park, CA, pp. 61–75
- Kayden 2001 National land use planning and regulation in America: Something whose time has never come. In: Alterman R (ed.) *National-Level Planning in Democratic Countries: An International Comparison of City and Regional Policymaking*. Liverpool University Press, Liverpool, UK, pp. 44–65
- Kushner J A 2000 Smart growth: Urban growth management and land-use regulation law in America. *Urban Lawyer* 32(2): 211–38
- Lowry R W 1998 *Preserving Public Lands for the Future: The Politics of Intergenerational Goods*. Georgetown University Press, Washington
- Mandelker D R, Payne J M 2001 *Planning and Control of Land Development: Cases and Materials*, 5th edn. Mathew Bender and Co
- Myren R S 1993 Growth control as a taking. *The Urban Lawyer* 25(2, Spring): 385–405
- Nelson A C 1992 Preserving prime farmland in the face of urbanization: Lessons from Oregon. *Journal of the American Planning Association* 58(4, Autumn): 467–88
- Pothukuchi K, Kaufman J L 2000 The food system: A stranger to the planning field. *Journal of the American Planning Association* 66(2), Spring: 34–45
- Pruetz R 1997 *Saved by Development*. Arje Press, Burbank, California
- Reid E P, Yeates M 1991 Bill 90—An act to protect agricultural land: An assessment of its success in Laprairie County, Quebec. *Urban Geography* July-August: 295–309
- Sax L J 1994 *Property Rights and The Economy of Nature: Understanding Lucas v. South Carolina Coastal Council*. Land Use and Environment
- Stein J M (ed.) 1993 *Growth Management: The Planning Challenge of the 1990s*. Sage Publications, Newbury Park, CA
- US Department of Agriculture and the Council on Environmental Quality 1981 National Agricultural Land Study
- Ward R M 1991 The US farmland protection policy act: Another case of benign neglect. *Land Use Policy* Jan: 63–8

R. Alterman

Fashion, Sociology of

While the word 'fashion' in English comes originally from the Latin *facere*, to make, in many European languages it derives from *modus*, to measure. Fashion can be said to be that which is understood as being right in a specific time and context. Its underlying

principal is revealed in the link between the *modus*-derivatives and the term 'modern,' with its original meaning of 'now' or 'for today.'

1. The Scope of Fashion

Knowing what is 'right for today' appears to be fundamental to all periods and societies. It was only at the turn of the twentieth century, however, that fashion became the subject of serious sociological study. Its universal nature taken as read, much has been made of its growing importance. People's lives have been increasingly affected in new ways, while more of the general population has the chance to follow fashion. Sapir (1931) contrasts fashion's functional irrelevance with its tremendous symbolic significance. This explains in part why fashion can affect so much. While usually expressed outwardly in clothing and adornment, there is nothing to prevent modes of thought, morals, art, and the like from being the equivalent of clothing fashions.

More recently, attention has been drawn to fashion's growing importance to the general population. Fashion was once a limited phenomenon, with very few in a society able to participate. In modern societies, however, increasing numbers have become conscious followers of fashion. Economic well-being has made it possible for the vast majority to participate in the 'fashion race.' Wealth is not the only factor, however; people also enjoy far greater psychological mobility, and the new form of communication it generates makes fashion more important today. In *Psychologie économique* (1902), Tarde coined the term 'inter-psychologie.' Society is the sum of people's mental condition in constant interaction. Free time and consumption become increasingly significant vehicles for such communication. New needs are born out of this mental interaction.

2. Fashion in Different Societies

Fashion's universal nature does not preclude strong variations in its social significance from society to society. Thus, Tarde (1903) differentiated between customary and fashionable societies. In the former, people see *custom* as governing their lives; they are more passionate about their country than about their period because custom prizes the past above all. In periods where *fashion* predominates, people are instead prouder of their era than their country.

Variations between societies are thrown up by different conditions. In economically stagnant periods, conventional stratification variables such as class and income are more decisive in people's lifestyles. In economically expansive periods, increasing importance is laid on fashion instead. Modern society is often contrasted with earlier epochs, and it is argued that

fashion's role has markedly increased, although many reasons are given for the change. An influential argument is that the differences in lifestyle of different groups have increased in number and scope, making it possible for people to exhibit their group's distinctive nature.

Simmel sees the 'need for distinction' and the 'need for union' as preconditions for fashion. Certain societies therefore lack the necessary motivation. Simmel (1904) here mentions that in fourteenth-century Florence, male dress was devoid of fashion. The 'need for union,' the need to express how certain groups were distinct from others with the help of fashion, was absent, and everyone could dress according to personal preference. Another of Simmel's examples was Venice at the same time. Again we find that fashion is absent from upper-class dress. Venetian aristocrats were ordered to dress in black so that the lower classes would not know how few they were. They did *not* want to differentiate themselves, and thus the 'need for distinction' that fashion requires was missing.

Variations in fashion's importance have also been tied to women's standing in society. According to Veblen (1899), it was woman's role to be a vicarious consumer; her consumption symbolized the man's wealth, and thus she became the vehicle of fashion. In Simmel (1904), however, we find an early formulation of compensation theory. Fashion acts as a safety valve. Deny women self-expression in other areas, and the only thing left for them is fashion. In fourteenth- and fifteenth-century Germany, the chance of personal development had increased, which, however, was denied to women. Rarely have more 'hypertrophic' modes of female dress been seen than in this period.

Simmel (1904) further explains fashion differences between societies. Fashion has two fundamental, symbolic functions: an ability to mark difference and an ability to express community. In many non-industrialized societies, the need for union is acute, while the need for distinction is weak. Simmel mentions certain African tribes who lacked both class differentiation and the concomitant need to signal differences. Where people in these societies want to note differences, it is often a matter of open hostility, which in turn is sterile ground for fashion. Given that fashion functions best where the distance between groups is otherwise only vaguely established, in instances where open hostility already exists, fashion as a sociological form becomes meaningless.

A similar argument can be found in König (1974), who identifies three different stages in society, each with a different type of fashion. The first significant historical change was the emergence of a fixed class hierarchy; at this stage, fashion is the preserve of an exclusive, feudal nobility. In the second stage, rigid differences between upper and lower classes are weakened; lower classes adopt the fashion of the elite, ultimately forcing it to alter fashion constantly to

retain its individuality. The third stage sees the development of egalitarian democracies and industrial technology. This modern form of society brings with it the spread of fashion through mass consumption.

3. *Fashion as a Process of Change*

Fashion prescribes what is 'the right thing' for 'now.' The question of what drives fashion's continuous changes has kept researchers busy, and a series of different theories has been given in answer. In Simmel's analysis of fashion we find a foretaste of many of the theories that were to follow.

The theory of *social differentiation* argues that fashion is based on class, the upper classes abandoning a fashion as soon as it is adopted by the lower classes (see Sect. 3.1). The theory of fashion's *internal dynamic* contends that changes are powered by an *Eigen-dynamik* in fashion (see Sect. 3.2). The theory of fashion as a manifestation of the *expression for other social change* is a variant of *spirit of the age* theory (see Sect. 3.3). Additionally, different *system theories* of fashion have been proposed, pointing to various factors, for example fashion producers, that accelerate the speed with which fashion changes (see Sect. 3.4). Theories of fashion as a *violation of convention* argue that fashion always breaks with the past, for example with prevailing conventions and fashions (see Sect. 3.5). Finally, there are theories of fashion as a channel of psychological and social *needs for expression* (see Sect. 3.6).

3.1 *Social Differentiation*

Many have stressed fashion's origins in class. The analysis by Veblen (1899) of patterns of fashion dispersal has become a classic. A fashion begins with the 'leisure class,' the group with most money. It is attractive to this class because fashion items cost a great deal more. By degrees the fashion spreads to the upwardly aspiring middle classes. Simmel (1904) also describes how fashion must constantly be supplied with new content if it is to give social expression both to similarities within groups and differences between them. To describe fashion's downward spread through society, the *trickle-down* model was adopted by Barber and Lobel (1953), who dismissed the argument that fashion is irrational as obscuring its connections with the economic system. Meanwhile, named for the imitation of the elite that brings in turn an alteration in elite habits, we have the *chase and flight* model and the *paper chase* model.

Bell (1976) follows Veblen in seeing economic competition as a force behind changes in fashion. Dress for the middle and upper classes is an expression of wealth. This wealth can take the form of *conspicuous consumption*, *conspicuous leisure*, and *conspicuous*

waste. To these, Bell adds his own category of *conspicuous outrage* to convey the fashion-setting class' conscious choice of clothing that does *not* conform with prevailing notions of good taste.

The theory that fashion is the means of social distinction for the upper classes in a hierarchical society has had great influence. Mimicry of social superiors and an urge to be different from 'inferiors' are often noted as the reasons for fashion's changes, while an individual's choice of style is seen as a significant means of securing his or her position in society. Such choices often establish their distance from those below them in the social pecking order, and express subtle distinctions of degree within one and the same class. Researchers who present other models often start from the theory of social differentiation, even if they distance themselves from its application to today's society. Some reason that the idea that fashion was the distinguishing mark of the elite was probably valid in the eighteenth and nineteenth centuries, but not today. It has been argued that the traditional upper class has lost its fashion primacy to be replaced, for example, by today's powerless elite. This new elite, termed *Divi* by Alberoni (1967), uses the mass media to present new lifestyles and new consumption.

3.2 Fashion's Inner Dynamic

Simmel's analysis included a further suggestion of what drives fashion. Fashion is made up of internal, conflicting forces that necessarily lead to change; it is distinguished by both differentiation *and* union, superficiality *and* profundity, freedom *and* dependence, individuality *and* imitation, destruction *and* construction, leadership *and* submission. Antagonistic forces are united in fashion in a way that ensures continual change. Nedelmann (1990) develops the idea of the inherently contradictory nature of fashion, describing its *Eigendynamik*. Fashion, by its nature, contains a stimulus for change derived from polar opposites. People like to imitate, thus spreading a certain fashion, but are equally drawn to differentiation. Imitation leads to the spread of fashion, while innovation and differentiation lead to its contraction. Nedelmann invokes Simmel, writing that the greater the degree of concord within the upper classes, the more frantic the search for imitation from beneath, and the more constant the search for new fashions by the upper classes. On the other hand, the more the members of the lower classes use style to differentiate between themselves, the less the upper classes are inclined to come up with new fashions. Fashion in this society thus has less significance.

Davis (1992) has also noted the ambivalence of fashion, and identifies its source in sexuality. Davis writes of a continuous dialectic in clothing between the erotic and the modest. Sellaerberg (1994) similarly stresses the significance of antagonism as a driving

force of change. Fashion reduces *and* creates complexity. It is a continuous break with authority, but its break is authoritative: 'this is what matters, for now.' In this contradiction one finds the motor of change.

3.3 Fashion as an Expression of the Spirit of the Age

A given fashion draws a line between the present and the past, expressing what is right 'for the day': fashion represents the present. This view is the basis for theories that fashion is the spirit of the age. Fashion not only grows out of the designers' world view, but also simply reflects a time. Laver (1937), who worked mainly on dress, drew up a fashion timetable for when an item of clothing is ahead of, of, or after its time.

Indecent	10 years before its time
Shameless	5 years before its time
Daring	1 year before its time
Smart	
Dowdy	1 year after its time
Hideous	10 years after its time
Ridiculous	20 years after its time
Amusing	30 years after its time
Quaint	50 years after its time
Charming	70 years after its time
Romantic	100 years after its time
Beautiful	150 years after its time.

Blumer (1969) approached fashion as an expression of the present. In interviews with designers intended to identify the stimuli that they felt influenced their creations, they mentioned trying to 'feel' what was current within art, literature, and public debate. World events left their trace in fashion. It has to express what happened at the time, it must be a 'Zeitgeist.'

3.4 System Theories

Davis (1992) differentiates between two types of sociological model used to explain the dissemination of clothing styles. The first, the *populist* model, centers on the consumer's role and the laymen's innovations in dress. Researchers who use the populist model study general consumption, for example, how groups such as teenagers, surfers, gays, skateboarders, or feminists set their own stamp on fashion. The other model is a *system* model. Some system models are predicated on the existence of established social centers for innovation, for example Paris or Milan, with their fashion shows, designers, editors, wholesalers, stylists, and so on. The fashion system is here seen as a closed circle of producers, distributors, and consumers.

Trickle-down theory has been criticized by system theorists because it does not focus on the complex

structures of organizations and marketing that channel and mediate the fashion process. Those who use a system model aim to analyze the influence of the different elements of the fashion system. One subject of controversy has been whether designers are creative interpreters or merely passive intermediaries in the commercially vital process of heeding consumers' wishes and reconciling them with the producers' requirements.

3.5 Fashion as a Break with Convention

The destruction of what exists and the construction of something new in the process are the elements that unite in fashion, according to Simmel. Researchers who today see fashion as breaking current rules usually identify groups other than the socioeconomic upper class as leading fashion. Campbell (1992) has criticized the class differentiation theorist's explanations and identifies different types of 'new': the new, the innovative, and the novel. The novel—new as different—is central to fashion as it breaks with what already exists. The necessary conditions for creating novelty exist in peripheral groups such as bohemians and artists. It is these groups who set fashion in motion, *not* an upper class. To argue that fashion is unconventional, however, is also to argue that fashion continuously re-establishes social conventions and social order. Blumer (1969) sees both similarities and differences between fashion and custom. In both cases it is a question of social ordering. Custom, however, is static or changes only slowly, while fashion is a constant creator of social order. In Blumer's argument, every fashion emerges from an intensive process of 'collective selection.' The alternatives that are ultimately selected become fashion in different areas. Fashion conjures order out of the vast jumble of possible styles where no objective norm can determine 'right' or 'wrong.'

3.6 Fashion as a Psychological and Social Expression

Psychologically oriented theories of fashion frequently center on general human needs such as the wish to be appreciated. It is often pointed out that people's need to display and strengthen their ego is particularly strong. Fashion has great shock value in the hunt for self-expression; it provides a tangible 'look at me.'

In a discussion of fashion as an expression of personality we can return to Simmel. Fashion makes it possible to express opposites. While conveying personality (paradoxically in the extreme imitation of the modern), it also offers the chance to *avoid* conveying individuality, and thus functions as the ego's 'iron mask.'

Fashion offers a means of psychological expression, but also a social expression of protest. People express

in fashion the impulses and wishes that are *not* sanctioned by normal social conventions. There is thus a constant tension between the more established 'domain of culture' and the marginal 'domain of fashion.' In fashion, young people express their stance through wearing what König (1973) calls a 'Gegen-uniform.'

The idea of fashion as a vehicle of social protest has been criticized, however. Protest styles are usually incorporated into fashion and exploited economically (Emberley 1987). Others argue that fashion's break with convention is actually a pseudodeviation. Although certainly defying convention, each break loses its original edge as soon as it is expressed in fashion (Wilson 1985).

Fashion has been interpreted as an expression of the search for a personal identity in an insecure world, and 'fads and fashions' as an expression of the worry that typifies our time. This has sometimes been interpreted as a negative phenomenon. Blumer (1969) thinks that the operation of fashion becomes natural in a changing world where people must continuously cut their ties with the past. Individuals are liberated from earlier models, and orient themselves towards the future.

4. Fashion's Cycles

Some researchers have been less interested in fashion's social dynamic than in an in-depth descriptive study of its cycles. Richardson and Kroeber (1940) use empirical data to identify six different formal aspects in women's dress taking, for example, the changes in length and width of skirts over 300 years. A number of researchers have followed Kroeber's empirical tradition, in which specific changes in fashion are seen to occur within a continuous cycle of longer periods. For example, Robinson studied beard fashions between 1842 and 1972 as time series analyses. To understand the elapse between the reappearance of the same fashion, Robinson argues that certain ground rules apply for fashion's cyclical changes, one being that as long as a number of people still follow a certain fashion, it prevents younger people from adopting it.

Fashion in this perspective is not born, but is rediscovered. Cross-cyclical changes receive particular attention as demonstrating sharp reactions to extreme styles. For example, the streamlined form's soft curves are contrasted with the ensuing period's particularly angular design.

5. Ambivalence Towards Fashion: The Vitality Unimportant

Fashion's symbolic significance has been contrasted with its lack of practicality. This has led many today to express ambivalence towards fashion, and there is a

tendency to legitimize it using functional arguments; the rational and practical is acceptable where the impractical is not. Function and symbolism can sometimes be reconciled in fashion, however. Aerodynamics has been described as a discovery of great significance for industrial design, and for a time all household appliances were streamlined. 'Streamlining' symbolized America's enlightened practice of throwing over old conventions to create a new and dynamic civilization. Those who chose, for example, a streamlined kettle, demonstrated that they also chose the rational, the future, the practical.

6. Fashion as Unpredictable

Increasingly, fashion's analysts have come to emphasize the inner dynamic in fashion. Individual theories are run together to reflect fashion's inherent contradictions: it is coercion *and* individual choice; it is social order *and* a break with social order; it is union *and* differentiation. The nature of fashion is seen as ever more complex.

There is a more simplistic perspective, however, common to those who base their interpretation of fashion on industrial products. Here one finds phrases such as 'the most thorough logic of planned aging,' and 'industry is based on extolling a perpetual neology.' Fashion, however, is to be found everywhere, not only in areas that are subject to commercial planning, while the aging of fashion is impossible to predict precisely, however much industry demands precise guidelines. The problems stem from fashion's most central characteristic: because it symbolizes 'now,' it can never be foreseen with any certainty. Fashion pinpoints what is right in the present, but the moment it can be predicted accurately, it is no longer a matter of fashion. The transient symbols of what is right for now can never be nailed down in advance.

See also: Action, Collective; Art and Culture, Economics of; Art, Sociology of; Collective Behavior, Sociology of; Consumer Psychology; Consumption, Sociology of; Cultural Expression and Action; Culture, Sociology of; Differentiation: Social; Dress and Fashion; Expressive Forms as Generators, Transmitters, and Transformers of Social Power; Hegemony: Cultural; Identity: Social; Sapir, Edward (1884–1939); Simmel, Georg (1858–1918); Taste and Taste Culture; Values, Sociology of

Bibliography

- Alberoni F 1967 *Consumi e società*. Società editrice il Mulino, Bologna, Italy
- Barber B, Lobel L S 1953 Fashion in women's clothes and the American social system. In: Bendix R, Lipset S M (eds.) *Class, Status and Power*. Free Press, Glencoe, IL, pp. 323–32
- Bell Q 1976 *On Human Finery*. Hogarth Press, London
- Blumer H 1969 Fashion: from class differentiation to collective selection. *Sociological Quarterly* **10**: 275
- Campbell C 1992 The desire for the new. Its nature and social location as presented in theories of fashion and modern consumerism. In: Silverstone R, Hirsch E (eds.) *Consuming Technologies: Media and Information in Domestic Spaces*. Routledge, London
- Davis F 1992 *Fashion, Culture and Identity*. University of Chicago Press, Chicago, IL
- Emberley J 1987 The fashion apparatus and the deconstruction of post-modern subjectivity. *Canadian Journal of Political Social and Theory* **11**: 38
- König R 1973 *The Restless Image: A Sociology of Fashion*. Allen & Unwin, London
- König R 1974 *A la Mode: On the Social Psychology of Fashion*. Seabury Press, New York
- Laver J 1937 *Tastes and Fashion: From the French Revolution Until Today*. G C Harrap, London
- Levine D N 1971 *On Individuality and Social Forms: Selected Writings*. University of Chicago Press, Chicago, IL
- Nedelmann B 1990 Georg Simmel as an analyst of autonomous processes: the merry-go-round of fashion. In: Kaern M, Phillips B S, Cohen R S (eds.) *Georg Simmel and Contemporary Sociology*. Kluwer, Dordrecht, The Netherlands
- Richardson J, Kroeber A L 1940 *Three Centuries of Women's Dress Fashions: A Quantitative Analysis*. University of California, Berkeley, CA
- Robinson D E 1976 Fashions in shaving and trimming of the beard: the men of the illustrated London news, 1842–1972. *American Journal of Sociology* **81**: 1133–41
- Sapir E 1931 *Fashion*. *Encyclopedia of the Social Sciences*. The Macmillan Company, New York, Vol. VI, pp. 139–44
- Sellerberg A-M 1994 *A Blend of Contradictions: Georg Simmel in Theory and Practice*. Transaction Publishers, New Brunswick, NJ
- Simmel G 1904 Fashion. *Fashion International Quarterly* **10**: 130–55
- Tarde G 1902 *Psychologie Economique*. F Alcan, Paris
- Tarde G 1903 *Laws of Imitation* [trans. from the 2nd French edition, Parsons E C] *Custom and Fashion*. Holt, New York, Chap. vii
- Veblen T B 1899 *The Theory of the Leisure Class: An Economic Study in the Evolution Of Institutions*. MacMillan, New York
- Wilson E 1985 *Adorned in Dreams*. Virago, London

A.-M. Sellerberg

Fast Food Enterprises

Fast food enterprises are companies that supply catering services on a fast turnover basis. The fast food sector comprises hamburger, chicken, pizza and other food outlets, coffee, pastry and icecream shops, ethnic food outlets as well as home delivery services for some of these. Many are franchised to international food chains. The rapid growth of this sector across cultural, national, and geographical frontiers has been taken, first, to signify fundamental changes in consumer

tastes and lifestyles, associated with changing family structures and employment patterns. Second, it is viewed as one of the most tangible features of cultural and economic globalization, marking the colonization by transnational organizations of one of the most important cultural terrains, namely what, how, when, and where people eat. Third, fast food enterprises seem to offer a prototype of the transition of the service industries from skill-based and people-based technologies to Taylorist and Fordist approaches which revolutionized manufacturing in the earlier part of the century. It is further argued that the fast food industry has become a paradigm for profound changes in the nature of contemporary societies. The term McDonaldization (Ritzer 1993–6) has been used to indicate the increasing dominance of the core principles embodied by the fast food restaurant in virtually every sector of society. Fourth, employment in fast food has become the first ‘real’ work exposure for increasing numbers of young people, affecting subsequent work attitudes and experiences. Lying at the meeting point of work and leisure, fast food enterprises represent a vital area on which to develop and test theories regarding the formation of contemporary identities as well as the interface between customer and employee.

1. Rationalization

Numerous factors account for the spectacular growth of this sector since the late 1950s. These include new franchising systems, pioneered by McDonald’s Ray Kroc, as well as broader social changes such as increasing participation of married women in the workforce, devaluation of housework, decreasing size of households, ever more hectic lifestyles, the greater importance of children as consumers, increasingly personalized and differentiated tastes in food, and ever greater consumer exposure to advertising images. All of these factors can all be seen as contributing to the success of ‘convenience eating.’ Above all, however, the success of fast food enterprises is often seen as the outcome of the belated rationalization of production, delivery, and marketing of a consistent, reliable, and inexpensive product. Writing in 1972, Levitt argued that fast food marked the industrialization of service, the application in catering of ‘technocratic thinking which ... replace[s] the high-cost and erratic elegance of the artisan with the low-cost, predictable munificence of the manufacturer’ (Levitt 1972, pp. 43–4). Taylorist and Fordist principles of management are clearly observable in the organization of the fast food service. These include a highly standardized and uniform product, routinized and fragmented food preparation procedures, highly rationalized technologies, and the employment of staff with virtually no catering skills and minimal training. Thus, Levitt views a McDonald’s retail outlet as ‘a machine that

produces, with the help of totally unskilled machine tenders, a highly polished product. Through painstaking attention to total design and facilities planning, everything is built integrally into the machines itself, into the technology of the system. The only choice available to the attendant is to operate it exactly as the designers intended’ (Levitt 1972, p. 46).

Rationalization in fast food enterprises extends into several other areas—the glamorization of their brands through relentless advertising, the systematic control of the customer through numerous discursive and nondiscursive practices (the location of outlets, menus, products, pricing, queuing, seating and waiting arrangements, background music, lighting, and outlet décor), and the organized attempt to control the emotional displays of the employees through the emblematic ‘service smile’ (Sturdy 1998). In all of these respects, fast food enterprises have moved rationalization far beyond Fordist and Taylorist regimes. Ritzer (1999) has argued along Weberian lines that excessive rationalization brings about the dissolution of magic and disenchantment. This leads to attempts on the part of producers to re-enchant consumption, by transforming their outlets into ‘cathedrals’ where fantasies are enacted—thus museums, cruise-ships, holiday resorts, theme parks, shopping malls, and even solid utilitarian institutions, like hospitals and universities, are gradually becoming arenas where consumers can fantasize, explore, experiment, and spend, after the model of fast food outlets. In a paradoxical way, then, attempts to re-enchant consumption are organized around the same rationalizing principles which account for disenchantment.

2. McJobs

Employment in fast food enterprises (‘McJobs’ in Ritzer’s expression) is characterized by extensive use of young workers, part-time employment, casualization, and high levels of staff turnover. Very few employees, including managers, view their work in fast food as anything other than short-term. Levels of remuneration are generally poor, as many employees live with parents or other relatives. Union representation is low, partly due to the casual nature of employment and partly because well-organized resistance of unions by employers. Job satisfaction is generally low, even by the low standards common in catering (Gabriel 1988). Discretion and individuality at work are very limited. While there is evidence that employees frequently break the rules even with the collusion of managers, this is most often in the interest of meeting demands from customers rather than in the form of resistance or self-expression. Gossip, joking, story-telling, day-dreaming, and some game-playing are all used as survival mechanisms.

Deskilling and control in fast food industries go beyond their counterparts in the manufacturing sector.

On the basis of fieldwork, Leidner (1993) and Reiter (1996) have argued that counter work entails extensive emotional labor (Hochschild 1983, Fineman 1993), a systematic attempt to control the employees' total emotional outlook towards the customers rather than simply their displays of emotion. While this argument has been contested, there is little doubt that much of the work performed in fast food enterprises combines low levels of skill with constant exposure to the demanding and critical gaze of the customer. Regular visits incognito by company inspectors add to the sense of constant surveillance. Interactions with the customers may not be scripted, but they are constantly monitored and assessed.

3. *The Consumers*

The high visibility of fast food enterprises, their restaurants occupying conspicuous positions in virtually every city and town, their logos, signs, and advertisements instantly recognizable, has made them an almost archetypal symbol of globalization. Some have argued that Americanization is a more appropriate term than globalization. Virtually all major fast food chains (including those associated with ethnic cuisine, like Mexican or Italian) originated in the United States, their products firmly developed on American food tastes. More importantly, fast food is symbolically linked to an American lifestyle, recognizable the world over, fast, informal, cool, fun, strong, reliable, open, modern, and democratic. The democratic quality of fast food is embodied in the orderly queues, the equal treatment of all customers, as well as well-publicized visits to outlets by celebrities, such as Prime Ministers, film stars, and tycoons, including Mr. Bill Gates, the world's richest man. The fun quality is highlighted by the ambience, the music, the smiling faces. The lifestyle symbolized by fast food outlets is especially attractive to children, who in many countries have been the consumer vanguard adopting fast food restaurants as their preferred venues for family entertainment. Fast food restaurants are not merely children-friendly places; they are places where the authority of the paying parent is diminished and replaced by an informal atmosphere of egalitarianism. The food has been prepaid, the tables have no heads to be occupied by the head of the family, the décor and music devised to mark the space as a space for children, where adults are welcome as long as they relinquish any claim to authority and discipline.

Based on a series of ethnographic studies of the growth of McDonald's in five East Asian countries, Watson claims that 'more than any other factor, ..., McDonald's success is attributable to the revolution in family values that has transformed East Asia' (Watson 1997, p. 19). The decline in family size, the weakening of extended kinship systems, and the erosion of

parental authority have all contributed to the considerable success of fast food in many developing countries. This has given rise to criticism of fast food as a type of American cultural imperialism, contributing, among other things, to the corruption of local traditions, the increasing uniformity of tastes, and the unstoppable growth of commercialization and consumerism. Watson and his colleagues observe unmistakable signs of cultural change brought about by the growth of transnational enterprises. An interesting example is the adoption of Western birthday parties in countries which neither celebrated birthdays nor even viewed them as significant dates. Moreover, they document evidence of direct resistance to the expansion of fast food restaurants ranging from rumor-mongering in Japan ('hamburgers contain cat-meat') to overt charges of antipatriotism leveled against fast food patrons in Korea. They conclude, however, that the impact of fast food on local cultures is neither wholly negative nor wholly one way. They observe that many local traditional restaurants have been forced to improve the standards of cleanliness, hygiene, and service in order to remain competitive against Western fast food chains. They also point at evidence of 'localization,' arguing that when an international institution is adopted by a new culture it is modified (occasionally as a token gesture) to suit local cultural and social conditions. While initially accepted as an exotic import, such an institution eventually becomes acclimatized to local culture. Some elements of fast food, like queuing, self-provisioning and self-seating, have been accepted, others have been modified. In Asian countries, McDonald's outlets are often turned into *de facto* leisure centers, coffee houses, and after-school clubs.

The charges leveled against the fast food business, however, go well beyond those of cultural imperialism. Some of these charges surfaced in 1994 during a well-publicized case in which McDonald's sued for libel two British activists who had published a pamphlet strongly critical of the company. This accused McDonald's of (a) causing widespread environmental damage, (b) exploiting young consumers through invasive advertising, (c) mistreating their employees, (d) causing unnecessary cruelty to animals, (e) promoting unhealthy eating habits, and (f) causing hunger in the Third World. The 'McLibel' trial became the longest trial in English history and was a public relations disaster for the company, spawning a much visited Website and becoming a focal point for antifast food activism. The verdict, handed down in 1997, amounted to a Pyrrhic victory for the company—the judge ruled that while some of the activists' allegations had been excessive, many were based on fact (Vidal 1998). In addition to giving the original pamphlet unprecedented publicity, the trial also had the effect of revealing the extent to which McDonald's resorted to litigation to silence its critics, resulting in further negative publicity for the company.

In spite of such incidents which indicate resistance to the fast food business on political, environmental, and cultural grounds, their growth appears unstoppable. The demographic, cultural, and employment factors which account for their early successes, if anything, appear to accelerate and spread to the newly industrialized countries. It appears that the major challenges to fast food enterprises in the future will come, first, from an increasing difficulty in recruiting casual, low-paid, young workers, and, second, from customer fatigue and the transience of fashions. It is quite possible that already some of the major fast food chains are losing customers to more up-market, fashionable, and at least temporarily alluring competitors.

See also: Consumer Culture; Consumer Psychology; Consumption, Sociology of; Customer Satisfaction; Globalization and World Culture; Leisure and Cultural Consumption; Leisure, Psychology of; Popular Culture; Retail Trade; Service Economy, Geography of; Services Marketing

Bibliography

- Fineman S 1993 *Emotion in Organizations*. Sage, London
- Gabriel Y 1988 *Working Lives in Catering*. Routledge, London
- Hochschild A R 1983 *The Managed Heart: Commercialization of Human Feeling*. University of California Press, Berkeley, CA
- Leidner R 1993 *Fast Food, Fast Talk*. University of California Press, Berkeley, CA
- Levitt T 1972 Production-line approach to service. *Harvard Business Review* 50: 41–52
- Reiter E 1996 *Making Fast Food: From the Frying Pan into the Fryer*. McGill-Queen's University Press, Montreal, PQ
- Ritzer G 1993–6 *The McDonaldization of Society*. Sage, London
- Ritzer G 1999 *Enchanting a Disenchanted World: Revolutionizing the Means of Consumption*. Pine Forge Press, Thousand Oaks, CA
- Sturdy A 1998 Customer care in a consumer society: Smiling and sometimes meaning it? *Organization* 5(1): 27–53
- Vidal J 1998 *McLibel*. London, Macmillan
- Watson J L 1997 Introduction: Transnationalism, localization, and fast food in East Asia. In: Watson J L (ed.) *Golden Arches East: McDonald's in East Asia*. Stanford University Press, Stanford, CA, pp. 1–38

Y. Gabriel

Fatherhood

Fatherhood defines a biological and social relationship between a male parent and his offspring. In common usage, 'to father' means to impregnate a woman and beget a child, thus describing a kinship connection and allowing for the intergenerational transfer of wealth

and authority in a patrilineal descent system. Fatherhood also refers to a patterned set of parenting behaviors and reflects a society's ideals about the rights and obligations of men in families. In general, fathers (including biological fathers, step-fathers, divorced fathers, and father-figures) are expected to love, support, protect, nurture, teach, discipline, and control their children, though many men do not fulfill all aspects of this idealized role. The concept of fatherhood also generalizes to other social and symbolic relationships, as when Christians refer to 'God The Father,' Catholics call priests 'Father,' Germans speak of their native country as 'The Fatherland,' and Americans label George Washington 'the father' of their country. Fatherhood thus reflects a normative set of social practices and expectations that become institutionalized within the family, religion, law, and culture.

1. Cross-cultural Variation in Fathering

Although folk beliefs suggest that fathering and mothering entail behaviors that are fixed by reproductive biology, humans must learn how to parent much like they learn other social behaviors. In every culture and historical period, men's family roles are shaped by social, economic, and cultural pressures, resulting in unique fatherhood ideals and practices. Although women have been the primary caretakers of young children in all cultures, the father role has varied from virtually no direct involvement to active participation in all aspects of children's routine care, feeding, and protection.

Anthropologists have identified two general patterns of male involvement in domestic life across the world's societies—one intimate and the other aloof. In the intimate pattern, men eat and sleep with their wives and children, talk with them during evening meals, attend births, and participate actively in infant care. In the contrasting aloof pattern, men often eat and sleep apart from women, spend their leisure time with men, stay away during births, and seldom help with child care (Whiting and Whiting 1975).

Cross-cultural research shows that about half of the world's known societies have exhibited close father-child relationships, with contact typically increasing as children grow older. Compared to societies with distant father-child relationships, those with involved fathers are more likely to be peaceful and include nurturing deities of both sexes in their creation stories (Sanday 1981). Distant-father societies are more likely to have stern male gods, with groups of men routinely excluding women from community decision-making, demanding deference from wives, and engaging in competitive displays of manliness (Coltrane 1996). Higher levels of father involvement with children are thus associated with more intimate husband-wife relationships and more gender equality

in the society, whereas lower levels of father involvement are associated with more distant husband–wife relationships and more overt male dominance.

2. Theories of Fatherhood and Families

To understand cross-cultural variation in fatherhood and identify the essence of family, Malinowski (1929) developed a theory based on the ‘principle of legitimacy.’ Malinowski observed that in Western societies, families were organized around the authority of fathers, with patrilineal descent and patrilocal residence assuring the position of the male household head. In matrilineal Trobriand Island society, in contrast, Malinowski observed that fathers were virtual strangers living among their wives’ relatives, making few contributions to the family household and exercising little power. The important link was not between husband and wife as in Western societies, but between sister and brother, with the latter being the symbolic and actual head of the household. These observations led Malinowski to question the universality of the father-centered household pattern and the importance of biological fatherhood, suggesting an alternative focus on the legitimacy of children. To be legitimate, a child must be socially approved through a formal marriage that has taken place between his or her parents. Malinowski thus argued that the important focus was social fatherhood, (not biological fatherhood). The purpose of social fatherhood was attaching a child to a particular man (whether father or uncle), thereby securing a place for that child in the social structure.

Subsequent theories of family and fatherhood suggest that Malinowski’s formulation may have been too narrow, as well as containing some gender-biased assumptions (Coltrane and Collins 2001). Societies have existed where the majority of children are not born to mothers married to husbands (e.g., in the Caribbean Islands), and similar patterns are becoming increasingly common in urban centers of modern industrial societies. Malinowski’s theory focuses on connecting children to social fathers because this brings them legal and social status, but it ignores the possibility that children could gain that status from mothers. Since males usually control more wealth and exercise more political power than females, it has been more important for a child’s economic fate to be connected to a man than to a woman, but it is important to see how this pattern can vary under different social and economic conditions.

According to Collins’ (1975) conflict theory, a strong emphasis on legitimacy is most common in societies with high levels of male dominance. Non-marital birth (‘illegitimacy’) is punished most severely in societies where men have the most extreme monopoly on power. In medieval Arab societies, for example, if a woman of the harem had an illegitimate

child, she was punished by death, with the sentence carried out by the woman’s husband, father, or brother. On the other hand, if the father of the illegitimate child was powerful enough, the mother could have a reasonably honorable status as a concubine or mistress, and the child could grow up with a recognized social position. This pattern has been repeated in male-dominated societies throughout the world. In medieval France, for instance, one duke’s son was known as William the Bastard; later, after he founded the English throne, he was called William the Conqueror. Considering variations in male dominance, Collins (1975) argues that illegitimacy rules do not so much say *no child without an official father* as they say there should be *no childbearing woman who does not officially belong to a man*. Collins extends Malinowski’s legitimacy theory by conceiving of the family as a property system regulating three kinds of property: rights of sexual possession, economic property rights, and intergenerational property rights. Such approaches help explain how fatherhood ideals and practices maintain men’s dominance over women in both premodern and modern societies.

3. Historical Development of Fatherhood

Fatherhood has been linked to the exercise of family authority at least since Ancient Greece and Rome, but it has been fathers’ public lives—their work, political exploits, and heroic battles—that have been chronicled in most histories. Fathers have also been care providers and teachers in different cultures and eras, and recent scholarship documents such practices. Before the modern era, childrearing was a collective enterprise and most family matters, including father–child relationships, were ruled primarily by duty and obligation (Shorter 1975). Because men’s work as farmers, artisans, and tradesmen occurred in the family household, most fathers were a visible presence in their children’s lives. Men introduced sons to farming or craft work within the household economy, oversaw the work of others, and were responsible for maintaining harmonious household relations. The family home was thus a system of control, as well as a center of production, and both functions tended to reinforce the father’s authority (Griswold 1993). Though mothers provided most direct care for infants and young children, men tended to be very active in the training and tutoring of children, and most parental advice was addressed to fathers. Because they were moral teachers and family heads, fathers were thought to have far greater responsibility for, and influence on, their children than mothers (Pleck 1987).

As market economies took over from home-based production, the father’s position as head and master of the household and moral instructor of his children was transformed. Men were increasingly called upon to seek employment outside the home and their direct

contact with family members declined. Although men's physical presence in the home was diminished, most did not stop participating in family activities. In fact, the first 'new' fatherhood movement emerged in the nineteenth century as middle-class fathers helped their wives through childbirth, romped with their offspring, forged close emotional bonds with daughters, and worried about sons' vocations and education (Griswold 1993). As the wage labor economy developed, however, men's occupational achievement outside the household took on stronger moral overtones and men came to be seen as fulfilling their family and civic duty not by teaching and interacting with their children as before, but by supporting the family financially (i.e., being a 'good provider'). The middle-class home, which was previously the normal site of production, consumption, and virtually everything else in life, was transformed slowly into a nurturant child-centered haven set apart from the impersonal world of work, politics, and other public pursuits. The ideal of separate spheres for men and women—work for him and home for her—became one of the defining features of the late nineteenth and early twentieth centuries.

4. Recent Trends

By the middle of the twentieth century, mothers had become almost solely responsible for home and children. Separate spheres ideology and the popularity of Freud's ideas about mother-infant bonding led to widespread acceptance of concepts like 'maternal deprivation.' Though few researchers bothered to ask who besides mothers took care of children, some researchers focused on 'father absence,' worrying that single-mother families could not set firm limits nor provide adequate masculine role models for sons. These studies have been criticized for confusing father absence with poverty, and for adopting the narrow view that aggressive and unemotional masculinity was both normative and desirable.

Starting in the 1970s, psychologists and sociologists began to report that a new fatherhood ideal was gaining in popularity. Films and television shows celebrated men's love for children, even if the men they pictured were initially shown as comically inept. Most social scientists had assumed that men were not capable of, nor interested in, actual hands-on parenting, but research showed that, at least in laboratory settings, fathers could care for infants and young children (Parke 1996). Following an individual differences research tradition in psychology, studies tended to focus on 'sex differences' in parenting style. Fathers typically were found to be more likely than mothers to engage in rough and tumble play, to be more directive, to give more attention to sons, and to treat boys and girls in sex-stereotyped fashion. Fathers' unique play styles were found to enhance children's emotional and

social competence, in part because they forced children to self-regulate (Parke 1996). Other studies by sociologists and psychologists showed that men and women performed child-care in different settings and in conjunction with different activities. Women most often did child-care in the home and frequently were engaged in multiple household tasks while simultaneously attending to children's needs. Fathers, in contrast, were more likely to watch children in public settings (e.g., parks and playgrounds) and to take care of them as a sole activity. The conventional pattern was thus for child care to be an ongoing and taken-for-granted task for mothers, but a novel and fun distraction for fathers (Coltrane 1998).

In response to the women's movement and maternal labor force participation, many observers predicted that fathers would begin assuming responsibility for a larger share of parenting and other family work. Several decades of research, however, have shown that while men are doing more, women continue to do most of the housework, remain the child care experts, and serve as emotional managers for their families. In general, family divisions of labor become more gender-segregated upon having children, with men increasing, and women decreasing, their paid work hours, even though both tend to increase their hourly contributions to family labor after a birth. Mothers are still more likely than fathers to take time off from their jobs to provide continuous child-care, spending significantly more time than fathers feeding, dressing, cleaning, and watching young children. Studies find that mothers spend double the time of fathers in these activities, even though many men did increase the time they spent with preschool and school-aged children during the 1980s and 1990s (Parke 1996, Pleck 1997). Two-parent, dual-earner families tend to share the most, with increasing numbers working different shifts and alternating routine child care between them. Research shows that couples share more family work when they are both employed full time, their incomes are similar, they believe in gender equality, and there are fewer, older, and male children (Coltrane 2000).

5. Measuring Father Involvement

Most researchers now agree that what fathers do with and for children is more important than simple co-residence or frequency of father-child contact (Parke 1996). The most influential approach distinguishing among various types of father involvement was offered by Lamb and his colleagues (Lamb et al. 1985) who suggested three principal components of fathering: interaction, availability, and responsibility. Interaction refers to the father's direct contact with his child through care giving and shared activities. Recent measurement strategies in this domain pay attention to the form and content of interaction, distinguishing between play, instruction, and care-taking, and at-

tempting to assess the quality of interaction, as well as the time spent interacting. Availability is a related concept focusing on the father's potential availability for interaction, by virtue of being present or accessible to the child. Responsibility refers to the role the father takes in ascertaining that the child is taken care of and arranging for resources to be available for the child both inside and outside the home. This type of parental management is one of the least studied aspects of fathering, but is one of the most important (Parke 1996). Mothers remain child-care managers in the vast majority of households, but evidence indicates that at least some fathers are taking a more active role in this domain (Coltrane 1996, Pleck 1997).

6. What Difference Do Fathers Make?

As scholars pay more attention to fathers, we are beginning to understand what influence their involvement might have on child development and gender relations. We know that fathers' resources improve children's life chances. When fathers also share child care and housework with their wives, employed women escape total responsibility for family work, evaluate the division of labor as more fair, are less depressed, and enjoy higher levels of marital satisfaction (Coltrane 2000). When men care for young children on a regular basis, they emphasize verbal interaction, notice and use more subtle cues, and treat sons and daughters similarly (rather than focusing on play, giving orders, and sex-typing children). These differences between everyday and occasional fathering deserve further study, for they portend differential social impacts. Most studies show that children with actively involved fathers demonstrate enhanced intellectual and socio-emotional adjustment, as well as developing more balanced ego structures and gender expectations (Coltrane 1996, Parke 1996).

At the same time that many fathers are spending an increasing amount of time with their children, high levels of divorce and increasing rates of nonmarital childbirth are keeping other men away from children. In many industrialized countries, growing numbers of men rarely see their children and do not support them financially. In the US, for example, single-parent households are increasing rapidly, but only about half of children eligible for child support from nonresident parents obtain court orders for it, and only about half of those receive full payments. Both trends in fatherhood—toward more direct involvement and toward less contact and financial support—are responses to the same underlying social developments, including women's rising labor force participation and the increasingly optional nature of marriage.

Longitudinal research shows that children from one-parent households (usually mother-headed) are at greater risk for negative adult outcomes (e.g., lower educational and occupational achievement, earlier

childbirth, relationship problems) than those from two-parent families (Cherlin et al. 1995, McLanahan and Sandefur 1994). Debates continue over whether these outcomes are specifically attributable to the actions or inactions of fathers, and whether state support and community programs might ameliorate negative effects. Some commentators call for a return to tougher fault-based divorce laws and gender-segregated models of fatherhood, but others label such responses a reactionary backlash against women's increasing autonomy. Reinstating fathers as household heads and expecting only minimal direct domestic contributions from them is likely to maintain hierarchical structures both inside and outside families. Whether absent or physically present and doing little, the ability of fathers to stay aloof from daily family life is linked to patriarchal family authority and the reproduction of systemic male dominance. Fathers' performance of everyday family work, in contrast, is associated with more intimate husband-wife relationships, more democratic family decision-making, and more gender equality in the society. The twenty-first century will undoubtedly see a continuing struggle between various intimate and aloof styles of fatherhood, with profound implications for men, women, and children.

Bibliography

- Cherlin A, Kiernan K E, Chase-Landsdale P L 1995 Parental divorce in childhood and demographic outcomes in young adulthood. *Demography* 32: 299–318
- Collins R 1975 *Conflict Sociology*. Academic Press, New York
- Coltrane S, Collins R 2001 *Sociology of Marriage and the Family*. Wadsworth, Belmont, CA
- Coltrane S 1996 *Family Man*. Oxford University Press, New York
- Coltrane S 1998 *Gender and Families*. Pine Forge Press, Thousand Oaks, CA
- Coltrane S 2000 Research on household labor. *Journal of Marriage and the Family* 62: 363–89
- Griswold R L 1993 *Fatherhood in America*. Basic Books, New York
- Lamb M E, Pleck J H, Charnov E L, Levine J 1985 Paternal behavior in humans. *American Zoologist* 25: 883–94
- McLanahan S, Sandefur G 1994 *Growing Up with a Single Parent*. Harvard University Press, Cambridge, MA
- Malinowski B 1929 *The Sexual Life of Savages in North-Western Melanesia*. H. Liveright, New York
- Parke R D 1996 *Fatherhood*. Harvard University Press, Cambridge, MA
- Pleck J H 1987 American fathering in historical perspective. In: Kimmel M S (ed.) *Changing Men*. 3rd edn. Sage, Newbury Park, CA, pp. 83–97
- Pleck J H 1997 Paternal involvement. In: Lamb M E (ed.) *The Role of the Father in Child Development*, Wiley, New York, pp. 66–103
- Sanday P R 1981 *Female Power and Male Dominance*. Cambridge University Press, Cambridge, UK
- Shorter E 1975 *The Making of the Modern Family*. Basic Books, New York

Whiting J W M, Whiting B 1975 Aloofness and intimacy of husbands and wives. *Ethos* 3: 183–207

S. Coltrane

Fear Conditioning

1. Basic Terms

The pairing of an initially neutral stimulus (the conditioned stimulus—CS) with a biologically relevant stimulus (the unconditioned stimulus—US) comes to elicit a response (conditioned response—CR) that is usually but not always similar to the response previously associated with the unconditioned stimulus (the unconditioned response—UR). In fear conditioning, the US is an aversive fear-eliciting stimulus such as painful electric shock or loud noise, the CS is a neutral tone or light stimulus. The unconditioned and the conditioned response consist of changes on the subjective, the behavioral and the physiological level and include (in humans) enhanced subjective fear and responses such as freezing, changes in heart rate and skin conductance, the release of stress hormones, reduced pain sensitivity and startle reflex potentiation.

The development of the CR is based on the formation of an association between a neutral stimulus and a stimulus with innate biological significance (Rescorla 1988). Most studies involving fear conditioning have used cue rather than context conditioning, i.e., discrete CSs were presented rather than using the environment of the animal (e.g., the cage) as CS. In addition, delay conditioning where the CS terminates with the US rather than trace conditioning where the CS and US are separated in time were used in most studies. Fear can be viewed as a specific reaction to threatening stimuli. It can turn into an anxiety disorder when the fear becomes disproportionate to the stimulus that elicits it or when fear is experienced in inappropriate situations.

2. Mechanisms of Fear Conditioning in Animals

The psychobiological basis of fear conditioning in animals, especially rodents, has been well established (LeDoux 2000, Davis 1992, Maren and Fanselow 1993). When a tone is used as CS and electric shock as US, the auditory CS reaches the lateral nucleus of the amygdala from the thalamus or the cortex, probably depending on the complexity of the CS. When contextual stimuli are employed as CS, the hippocampus is also involved and transmits the signal to the basolateral and the accessory basolateral nucleus of the amygdala. The nociceptive US also reaches the

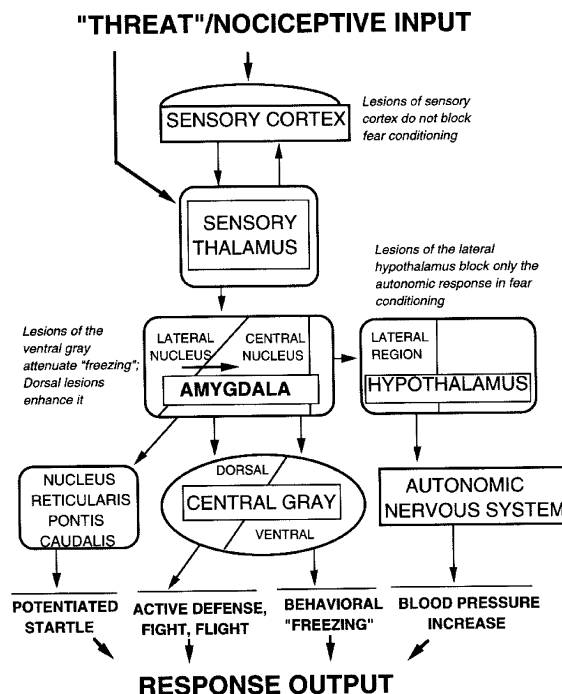


Figure 1

Fear stimuli proceed from sense receptor systems to the sensory cortex and/or sensory thalamus and then to the amygdala. There are three important connections efferent to the amygdala: (a) a projection from the central amygdala to the lateral hypothalamic area that mediates the autonomic components of the emotional response; (b) projections to the midbrain central gray region mediates defensive freezing and escape behavior and (d) a direct projection to the nucleus reticularis pontis caudalis modulates the startle reflex. (Based on Bradley Lang 2000 with kind permission)

lateral nucleus of the amygdala via thalamic and cortical pathways. Conditioning seems to require an intact lateral nucleus of the amygdala when cue conditioning and an intact basolateral and accessory basolateral nucleus when context conditioning is involved. The central nucleus of the amygdala receives input from the lateral and the basolateral nuclei and it projects to brainstem nuclei that initiate the expression of peripheral fear responses. Whereas the lateral or basolateral nucleus are thus important for the acquisition of the CR, the central nucleus is essential for the expression of the CR (see Figure 1).

The initial site of plasticity seems to be the lateral nucleus of the amygdala; however, thalamic and cortical reorganization have also been shown during fear conditioning (cf. Weinberger 1998). It is to date not clear if the memory trace is stored in the amygdala or if the amygdala only modulates memories that are stored elsewhere, e.g., in the cerebral cortex. If the

amygdala is inactivated during fear conditioning, no CR is formed. If the inactivation occurs after conditioning, then the memory trace is not abolished (LeDoux 2000). In addition, cognitive aspects of fear may involve other pathways.

3. Fear Conditioning in Humans

In humans, the role of the amygdala is less clear. In contrast to rodents, humans possess a much larger frontal cortex and it is likely that it takes over some of the functions that are related to the amygdala in rodents. Damage to the amygdala in humans, has, however, also been related to deficits in fear conditioning whereas lesions in the hippocampus have been shown to interfere with declarative learning and contextual fear conditioning (cf. Bechara et al. 1995). Most studies using positron emission tomography (PET) were unable to replicate the findings from animal experiments that assigned an important role to the amygdala in fear conditioning (for a review see Büchel and Dolan 2000). They instead reported activations in areas such as the hypothalamus, the thalamus, the cingulate gyrus, the premotor and the parietal cortex and, recently, also in the cerebellum (Fischer et al. 2000). Later studies, using differential conditioning paradigms, found some amygdala activation that was positively correlated with activations in the auditory cortex, the frontal cortex and the thalamus. An additional PET study used masked CSs and found significant amygdala activation with the right amygdala showing more activation with unaware (masked) and the left amygdala showing more activation with aware (unmasked) CS presentation. The notion that a direct pathway from the thalamus to the amygdala also exists in humans was recently supported by Morris et al. (1999) who showed that activation of the right amygdala in a masked conditioning paradigm was associated with activity in the superior colliculus and pulvinar but not with activity at cortical sites.

Several recent studies used functional magnetic resonance imaging (fMRI) to study fear conditioning in humans (for a review see Büchel and Dolan 2000). Probably due to the higher spatial and temporal resolution of this method compared to PET, amygdala activation was reported in all studies. In addition to the amygdala, the acquisition of a fear response created activation in the cingulate gyrus and the premotor cortex. Activation of the insulae has also been reported when painful stimuli were used as US. When trace conditioning was employed, activation in the hippocampus in addition to activation in the amygdala and cortical areas was noted.

Based on the observation that not all stimuli are equally easy to condition Seligman (1971) suggested that phylogenetically relevant stimuli such as snakes or spiders might serve as 'prepared stimuli' and be more conditionable than others. Subsequently,

Öhman (for a summary see Öhman et al. 2000) has shown that conditioned skin conductance responses to a prepared CS such as an angry face or a picture of a snake persist for a longer time (even when the presentation is masked) than when a neutral stimulus such as a picture of a mushroom is used as CS and combined with electric shock as US. Moreover, masked presentations of these stimuli also lead to the acquisition of conditioned responses suggesting that unconscious learning of these stimulus associations is possible.

More recent accounts of fear conditioning have emphasized the role of cognitive variables as well as genetic predisposition in the acquisition of fear (e.g., Davey 1995, Kendler et al. 1992). Part of this may be a cognitive bias termed covariation bias that suggests that aversive outcomes are overestimated in the face of fear-relevant stimuli (e.g., Tomarken et al. 1989). However, although these cognitive variables have recently been discussed extensively as important mediators of the acquisition of fear, the biases themselves can be acquired through conditioning processes and it is not clear to what extent altered cognitive processing precedes or follows the acquisition of fear.

4. Genetic Factors

Genetic factors seem to play a role in the development in anxiety disorders in humans. In familial studies, twin studies and adoption studies show genetic variances of 20 to 40 percent. The same was found for questionnaire variables such as neuroticism. Fear conditioning ability and predisposition to learn fear responses were extensively studied in animals. In rodents, fear processes or anxiety measured in the open field test or avoidance conditioning have revealed a strong genetic basis explaining most of the inter-individual variance. However, the genes responsible for that trait have not yet been identified.

The situation is different for antisocial personality disorder and criminal behavior. Assuming that a deficit in anticipatory fear conditioning is the central symptom from which all consequent social problems follow, all evidence ranging from twin to adoption studies and gene sequencing suggests a moderate to strong genetic influence. If sensitivity to alcoholism and novelty-seeking are included as comorbidity factors the genetic variance increases further. Whether this is related to genetic fear conditioning remains to be seen (Plomin et al. 2001).

5. Significance of Fear Conditioning for the Understanding of Anxiety Disorders

Although classical fear conditioning has been implicated as an important mechanism in the development of most anxiety disorders, studies that support

the etiological significance of fear conditioning have been rare. Retrospective data from either adult or child patients or the parents of children with specific phobias revealed that about 40–75 percent of the subjects or the parents could recall a traumatic event (e.g., Menzies and Clarke 1993), a substantial number of the remaining subjects experienced vicarious conditioning or had been verbally informed about potential dangers related to the feared object. In evaluating these data it must be kept in mind that all these reports are retrospective and that fear conditioning may be nondeclarative and therefore not be accessible to conscious recollection. Of greater relevance are therefore studies that examine fear conditioning in clinical groups or analogue samples.

In social phobias, several studies used differential conditioning with faces as CS and aversive odor or shock as US. In two studies, a conditioned response was already present prior to the acquisition phase with the social phobics showing enhanced amygdala and (in one study) orbitofrontal responses to the face CS suggesting that they displayed a fear response to neutral faces that never elicits amygdala activation in healthy humans (e.g., Birbaumer et al. 2000). In another study stronger amygdala activation was present to the CS+ than the CS– only in social phobics whereas the healthy controls showed deactivation. However, the scanning was performed late in the acquisition phase and it has been shown that the amygdala is only responsive early in the acquisition process. Additional studies that used conditioning paradigms or phobia-relevant cues during imaging also found activation increases in the anterior cingulate and the insulae in accordance with the findings of fear conditioning in healthy subjects.

In persons with psychopathy and antisocial personality disorder disturbed fear conditioning seems to be present with a concomitant complete absence of orbitofrontal, insular and cingulate activation during fear conditioning (Birbaumer et al. 2000). This absence of associative learning with respect to the aversive consequences of a stimulus might be the basis of the antisocial and often criminal behavior of these persons.

In posttraumatic stress disorder, higher conditionability as evident in enhanced and more difficult to extinguish fear responses has been observed. It is, however, not clear to what extent this conditionality predated the disorder or if it must be viewed as a consequence of the disorder and to what extent conditioning of trauma-related stimuli actually occurred because none of the studies involved trauma-relevant cues (e.g., Orr et al. 2000).

6. Future Perspectives

Future research must focus on the elucidation of the neuronal correlates of fear conditioning in healthy humans and patients. It is not clear to what extent the

amygdala plays the same role in fear conditioning in humans as it does in rodents. Furthermore, the role of awareness of the CS and the CS–US relationship has not been determined by neuroimaging methods. The importance of fear conditioning for the understanding of anxiety disorders needs to be explored using paradigms that involve fear-relevant and prepared stimuli as well as longitudinal measurements. Finally, the role of cognitive and genetic variables and their relationship to fear conditioning needs to be further explored.

See also: Anxiety and Anxiety Disorders; Anxiety and Fear, Neural Basis of; Classical Conditioning and Clinical Psychology; Classical Conditioning, Neural Basis of; Eyelid Classical Conditioning; Fear: Potentiation of Startle; Fear: Psychological and Neural Aspects

Bibliography

- Büchel C, Dolan R J 2000 Classical fear conditioning in functional neuroimaging. *Current Opinion in Neurology* **10**: 219–23
- Bechara A, Tranel D, Damasio H, Adolphs R, Rockland C, Damasio A R 1995 Double dissociation of conditioning and declarative knowledge relative to the amygdala and the hippocampus in humans. *Science* **269**: 1115–18
- Birbaumer N, Veit R, Flor H, Patrick C 2000 Excess and lack of fear: an fMRI study with social phobias and psychopaths. *Society for Neuroscience Abstracts* **26**: S1072
- Bradley M M, Lang P J 2000 Measuring emotion: behavior, feeling and physiology. In: Lane R D, Nadel L (eds.) *Cognitive Neuroscience of Emotion*. Oxford University Press, New York, pp. 242–76
- Davey G C 1995 UCS revaluation and conditioning models of acquired fears. *Behavior Research and Therapy* **27**: 521–8
- Davis M 1992 The role of the amygdala in fear and anxiety. *Annual Review of Neuroscience* **15**: 353–75
- Fischer H, Andersson J L, Furmark T, Fredrikson M 2000 Fear conditioning and brain activity: a positron emission tomography study in humans. *Behavioral Neuroscience* **114**: 671–80
- Kendler K S, Neale M C, Kessler R C, Heath A C, Eaves L J 1992 The genetic epidemiology of phobias in women: the interrelationship of agoraphobia, social phobia, situational phobia and simple phobia. *Archives of General Psychiatry* **49**: 273–81
- LeDoux J E 2000 Emotion circuits in the brain. *Annual Review of Neuroscience* **23**: 155–84
- Maren S, Fanselow M S 1993 The amygdala and fear conditioning: has the nut been cracked? *Neuron* **16**: 237–40
- Menzies R G, Clarke J C 1993 The etiology of childhood water phobia. *Behavior Research and Therapy* **31**: 499–501
- Morris J S, Öhman A, Dolan R J 1999 A subcortical pathway to the right amygdala mediating 'unseen' fear. *Proceedings of the National Academy of Science USA* **96**: 1680–5
- Öhman A, Flykt A, Lundqvist D 2000 Unconscious emotion: evolutionary perspectives, psychophysiological data and neuropsychological mechanisms. In: Lane R D, Nadel L (eds.) *Cognitive Neuroscience of Emotion*. Oxford University Press, Oxford, pp. 296–327

- Orr S P, Metzger L J, Lasko N B, Macklin M L, Peri T, Pitman R K 2000 De novo conditioning in trauma-exposed individuals with and without posttraumatic stress disorder. *Journal of Abnormal Psychology* **109**: 290–8
- Plomin R, DeFries J, McClearn G, Roth M 2001 *Behavioral Genetics* 4th Edn. Freeman, San Francisco, CA
- Rescorla R A 1988 Pavlovian conditioning: It's not what you think it is. *American Psychologist* **43**: 151–60
- Seligman M E P 1971 Phobias and preparedness. *Behavior Therapy* **2**: 307–20
- Tomarken A J, Mineka S, Cook M 1989 Fear-relevant selective associations and covariation bias. *Journal of Abnormal Psychology* **98**: 381–94
- Weinberger N M 1998 Physiological memory in primary auditory cortex: characteristics and mechanisms. *Neurobiology of Learning and Memory* **70**: 226–51

H. Flor and N. Birbaumer

Fear: Potentiation of Startle

When people are afraid they tend to show a bigger startle response to a loud sound. Rats show a similar effect. This phenomenon, called 'fear-potentiated startle,' has been used to study the neural mechanisms involved in fear and anxiety in both rats and humans.

1. Fear-potentiated Startle in Rats

Capitalizing on anecdotal evidence that people startle more when they are afraid, Brown et al. (1951) demonstrated that the amplitude of the acoustic startle reflex in the rat can be augmented by presenting the eliciting auditory startle stimulus shortly after onset (e.g., 3s) of a cue (e.g., a light) previously paired with a footshock. This fear-potentiated startle occurs with an auditory, visual, or tactile conditioned stimulus (CS), and when startle is elicited by either a loud sound or an airpuff in rats (Davis et al. 1993). In this test a central state of fear is considered to be the conditioned response. Conditioned fear is defined operationally by elevated startle amplitude in the presence of a cue previously paired with a shock. Thus, the CS does not elicit startle. Furthermore, the startle-eliciting stimulus is never paired with a shock. Instead, the CS is paired with a shock, and startle is elicited by another stimulus in the presence or absence of the CS. Fear-potentiated startle only occurs following paired rather than unpaired or random presentations of the CS and the shock, indicating that it is a measure of conditioned fear. Discriminations between visual and auditory conditioned stimuli, and between auditory cues and visual cues that differ in duration, have also been demonstrated with potentiated startle. Pairing one cue with shock does not lead to generalization of fear-potentiated startle in testing with other cues. Fear-potentiated startle still occurs reliably at least

one month after original training, making it appropriate for the study of long-term memory as well.

It has been suggested, however, that potentiated startle may not reflect increased fear in the presence of a CS, but instead results from the animal making a postural adjustment (e.g., crouching) which is especially conducive to startle, in anticipation of the impending footshock. However, spinally transected rats that were held rigidly in a modified stereotaxic instrument, which prevented obvious postural adjustments, still showed fear-potentiated startle, measured with the pinna component of startle. Potentiation of startle measured electromyographically in neck muscles also occurs in the absence of any obvious postural adjustment. In addition, the magnitude of potentiated startle correlates highly with the degree of freezing, a very common measure of fear.

1.1 Effects of Different Drugs

Drugs that reduce fear or anxiety in humans decrease potentiated startle in rats (Davis et al. 1993). Benzodiazepines, opiates, β -noradrenergic antagonists, drugs that decrease norepinephrine release, certain serotonin agonists, nicotine, and perhaps alcohol decrease or fully block fear-potentiated startle. In many cases these treatments do not depress startle levels in noise-alone trials, although this is not always the case. Conversely, drugs that increase norepinephrine release or decrease GABA (γ -aminobutyric acid) mediated neurotransmission and exaggerate anxiety in anxious people increase the magnitude of potentiated startle in rats.

1.2 Neural Systems Involved in Fear-potentiated Startle

A major advantage of the fear-potentiated startle paradigm is that fear is measured by a change in a simple reflex. Hence, fear is expressed through some neural circuit that is activated by the CS and ultimately impinges on the startle circuit. Figure 1 shows a schematic summary diagram of the neural pathways we believe are required for both the acquisition and expression of fear-potentiated startle in rats using a visual CS. These pathways involve convergence of the CS and the unconditioned shock stimulus at the lateral and basolateral amygdala nuclei, which project to the central nucleus of the amygdala, and this in turn projects directly to the startle pathway.

1.2.1 The primary acoustic startle pathway. Because the acoustic startle reflex has such a short latency (e.g., 8 ms measured electromyographically in the hindlegs), it must be mediated by a simple neural pathway. Using a variety of techniques, we have de-

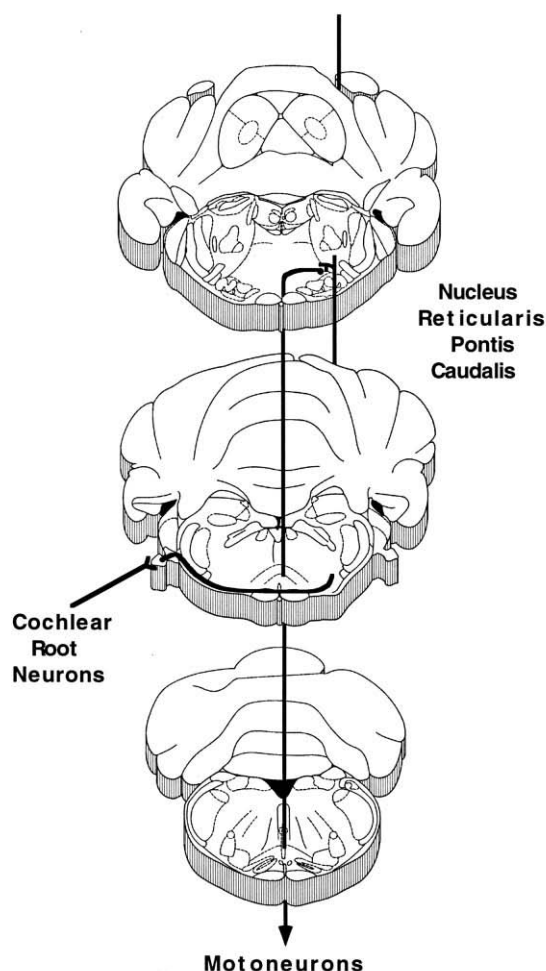


Figure 1
Schematic summary of the pathways believed to be involved in fear-potentiated startle in rats using schematic drawings of coronal brain slides in the rat (reproduced by permission of the Society for Neuroscience from *Journal of Neuroscience* 1996, **16**: 3775–89)

terminated that the primary acoustic startle reflex in rats is probably mediated by three central synapses (Lee et al. 1996). The first involves terminals of the auditory nerve that synapse with large neurons embedded in the auditory nerve, called cochlear root neurons. These send exceedingly thick axons (sometimes as large as 7 μm in diameter) to neurons in the nucleus reticularis pontis caudalis (PnC). These cells project directly to motoneurons in the facial motor nucleus (pinna reflex and perhaps the eyeblink component of startle) and spinal cord (whole-body startle). Chemical lesions of these neurons or local

infusion of glutamate antagonists in the vicinity of these synapses eliminated acoustic startle. Although other investigators have proposed that additional connections between the cochlear nucleus and the PnC may be important for startle (Koch and Schnitzler 1997), we believe that this simple, three-synapse pathway is the main acoustic startle pathway. For example, although the dorsal cochlear nucleus also projects directly to the PnC, lesions of the dorsal cochlear nucleus only decreased startle amplitude when high-intensity startle stimuli were used.

1.2.2 The role of the thalamus. Visual information uses parallel pathways that involve either (a) direct retinal inputs to the lateral posterior nucleus of the thalamus, which projects directly to the perirhinal cortex, or (b) retinal inputs to the dorsal lateral geniculate nucleus, which projects indirectly to the perirhinal cortex via the visual cortex. Lesions of both of these areas of the visual thalamus together, but not either one alone, blocked fear-potentiated startle using a visual CS (Shi and Davis 1996).

If an auditory CS is used, this involves parallel inputs from the auditory thalamus to the perirhinal cortex, either directly or indirectly via the auditory cortex (Campeau and Davis 1995a). Pre- or post-training lesions of the entire auditory thalamus completely blocked fear-potentiated startle to an auditory CS but not to a visual CS. Post-training lesions restricted to the main body of the medial geniculate, which projects to the perirhinal cortex via the auditory cortex, also specifically blocked fear-potentiated startle to the auditory CS.

Pain information reaches the amygdala via parallel pathways that include the caudal granular/dysgranular insular cortex and the posterior intralaminar nuclei of the thalamus. Pretraining lesions of both the insular cortex and the posterior intralaminar nuclei of the thalamus, but not lesions of either structure alone, blocked the acquisition of fear-potentiated startle (Campeau and Davis 1995a, Shi and Davis 1999). However, post-training lesions of both of these areas together did not prevent expression of conditioned fear. These results suggest that parallel cortical and subcortical pathways are involved in relaying shock information during fear conditioning.

1.2.3 The role of the anterior perirhinal and insular cortex. The perirhinal cortex, which receives either visual or auditory CS information, projects directly to the lateral and basolateral amygdala. Post-training lesions of the anterior perirhinal cortex completely blocked the expression of fear-potentiated startle using a visual CS (Rosen et al. 1992), provided the lesion destroyed both the dysgranular and agranular

portions of the perirhinal cortex. Post-training lesions of the perirhinal area (including secondary auditory cortices) blocked fear-potentiated startle to both auditory and visual CS (Campeau and Davis 1995a). However, reliable potentiated startle was observed after retraining in animals that had sustained main geniculate body lesions (which would destroy cortical connections between the thalamus and perirhinal cortex), or following pretraining lesions of the perirhinal area. These data suggest that cortical areas are normally used for the expression of fear conditioning, but that subcortical areas can take over if the cortex is damaged. Finally, as mentioned before, shock information seems to require the insular cortex, which in turn projects directly to the lateral nucleus of the amygdala.

1.2.4 The role of the amygdala. The perirhinal cortex projects directly to the lateral and basolateral amygdala. Selective destruction of cell bodies in the lateral and basolateral nuclei by local infusion of neurotoxins caused a complete blockade of fear-potentiated startle when the lesions were made, whether before or after training (Sananes and Davis 1992). This blockade did not seem to result from a disruption of vision, and other studies found that neurotoxic lesions of these amygdaloid nuclei also blocked fear-potentiated startle using an auditory CS. In contrast, lesions of a variety of other subcortical areas, such as the hippocampus, the septal area, the cerebellum, or the bed nucleus of the stria terminalis, had no consistent effect. The lateral, and especially the basolateral, nuclei of the amygdala project directly to the central nucleus of the amygdala, lesions of which blocked the expression of fear-potentiated startle using either a visual (Hitchcock and Davis 1986) or an auditory CS (Campeau and Davis 1995b).

Because CS and shock information converge neurally at the lateral and basolateral amygdala nuclei, this could be the site of plasticity for fear-potentiated startle. In fact, local infusion of NMDA (*N*-methyl-D-aspartate) antagonists into the amygdala blocked the acquisition, but not the expression, of fear-potentiated startle using either a visual or an auditory CS (Davis et al. 1993). This was probably not a result of shock information failing to reach the amygdala, because this treatment also blocked the acquisition of second-order fear-potentiated startle, which does not involve shock during second-order training.

By eliciting startle-like responses electrically from various points along the startle pathway before and after presentation of a light previously paired with a shock, we concluded that fear ultimately alters transmission at the PnC. Consistent with this, local infusion into the PnC of several compounds, such as glutamate antagonists and corticotropin-releasing hormone an-

tagonists, blocked fear-potentiated startle at doses that have no effect on baseline startle (Koch and Schnitzler 1997).

The central nucleus of the amygdala projects directly to the PnC, and lesions at several points along this pathway blocked the expression of fear-potentiated startle. Low-level electrical stimulation of the amygdala markedly increased the acoustic startle amplitude (Rosen and Davis 1988) at currents and durations that did not produce any other signs of behavioral activation. Transit times from the amygdala to the startle circuit appear to be about 4–5 ms or even less. Electrical stimulation of the amygdala in anesthetized rats facilitated tone-evoked activation of cells in the PnC (Koch and Ebert 1993).

However, it is now clear that there is at least one critical synapse between the amygdala and the startle pathway, located in an area in or very close to the mesencephalic reticular formation, dorsal periaqueductal gray or deep white layers of the superior colliculus (Koch and Schnitzler 1997). Neurotoxic lesions of this area or local infusion of muscimol blocked the expression, but not the acquisition, of fear-potentiated startle.

2. Fear-potentiated Startle in Humans

Fear-potentiated startle can also be measured in humans by using the eyeblink component of the startle reflex. In one test, people are told that when a certain colored light comes on they might get a shock on the wrist, whereas they will not get a shock when a different colored light comes on. Startle is elicited with bursts of noise through earphones, and the eyeblink component of startle is measured electromyographically from the orbicularis oculi muscles. Startle amplitude was consistently higher in the presence of the light that signaled shock (Grillon et al. 1991). The size of this increase depended on the time when the subject expected the shock on the basis of verbal instructions, and also occurred when conditioning procedures were used.

Thus far, there is a close correspondence between results gathered in rats and humans. Brain imaging studies show that the amygdala is activated during this verbally mediated fear-potentiated startle test, and that patients with lesions of the amygdala fail to display fear-potentiated startle. People also show an increase in startle when they see scary pictures—for example, of a gun in their face, a dog about to bite them, or mutilated bodies (Lang et al. 1990). The size of the increase in startle was related directly to how the subjects rated the pictures in terms of negative valence and arousal.

See also: Anxiety and Anxiety Disorders; Anxiety and Fear, Neural Basis of; Autonomic Classical and

Operant Conditioning; Cardiovascular Conditioning; Neural Substrates; Classical Conditioning, Neural Basis of; Conditioning and Habit Formation, Psychology of; Eyelid Classical Conditioning; Fear Conditioning; Fear: Psychological and Neural Aspects

Bibliography

- Brown JS, Kalish H I, Farber I E 1951 Conditional fear as revealed by magnitude of startle response to an auditory stimulus. *Journal of Experimental Psychology* **41**: 317–28
- Campeau S, Davis M 1995a Involvement of subcortical and cortical afferents to the lateral nucleus of the amygdala in fear conditioning measured with fear-potentiated startle in rats trained concurrently with auditory and visual conditioned stimuli. *Journal of Neuroscience* **15**: 2312–27
- Campeau S, Davis M 1995b Involvement of the central nucleus and basolateral complex of the amygdala in fear conditioning measured with fear-potentiated startle in rats trained concurrently with auditory and visual conditioned stimuli. *Journal of Neuroscience* **15**: 2301–11
- Davis M, Falls W A, Campeau S, Kim M 1993 Fear-potentiated startle: A neural and pharmacological analysis. *Behavioural Brain Research* **58**: 175–98
- Grillon C, Ameli R, Woods S W, Merikangas K, Davis M 1991 Fear-potentiated startle in humans: Effects of anticipatory anxiety on the acoustic blink reflex. *Psychophysiology* **28**: 588–95
- Hitchcock J M, Davis M 1986 Lesions of the amygdala, but not of the cerebellum or red nucleus, block conditioned fear as measured with the potentiated startle paradigm. *Behavioral Neuroscience* **100**: 11–22
- Koch M, Ebert U 1993 Enhancement of the acoustic startle response by stimulation of an excitatory pathway from the central amygdala/basal nucleus of Meynert to the pontine reticular formation. *Experimental Brain Research* **93**: 231–41
- Koch M, Schnitzler H-U 1997 The acoustic startle response in rats—circuits mediating evocation, inhibition and potentiation. *Behavioural Brain Research* **89**: 35–49
- Lang P J, Bradley M M, Cuthbert B N 1990 Emotion, attention, and the startle reflex. *Psychological Review* **97**: 377–95
- Lee Y L, Lopez D E, Meloni E G, Davis M 1996 A primary acoustic startle pathway: Obligatory role of cochlear root neurons and the nucleus reticularis pontis caudalis. *Journal of Neuroscience* **16**: 3775–89
- Rosen J B, Davis M 1988 Enhancement of acoustic startle by electrical stimulation of the amygdala. *Behavioral Neuroscience* **102**: 195–202
- Rosen J B, Hitchcock J M, Miserendino M J D, Falls W A, Campeau S, Davis M 1992 Lesions of the perirhinal cortex but not of the frontal, medial prefrontal, visual, or insular cortex block fear-potentiated startle using a visual conditioned stimulus. *Journal of Neuroscience* **12**: 4624–33
- Sananes C B, Davis M 1992 N-Methyl-D-aspartate lesions of the lateral and basolateral nuclei of the amygdala block fear-potentiated startle and shock sensitization of startle. *Behavioral Neuroscience* **106**: 72–80
- Shi C, Davis M 1996 Anatomical tracing and lesion studies of visual pathways involved in fear conditioning measured with fear potentiated startle. *Society for Neuroscience Abstracts* **22**: 1115

Shi C-J, Davis M 1999 Pain pathways involved in fear conditioning measured with fear-potentiated startle: Lesion studies. *Journal of Neuroscience* **19**: 420–30

M. Davis

Fear: Psychological and Neural Aspects

Fear is an innate behavioral–cognitive–emotional system that evolved because of its evolutionary success in protecting animals (including humans) from danger. While fear to certain kinds of stimuli is unlearned, fear can also be acquired rapidly and lastingly to other kinds of stimuli, allowing animals to respond adaptively to new environmental threats. This entry will present some of the behavioral and physiological aspects of fear that make it a useful model system for understanding the psychological and neurobiological mechanisms of emotion, learning, and memory, as well as the development of a variety of human fear-related disorders.

1. Fear—Definition and Background

All organisms are subject to threats and dangers from their environments, such as predation, dominance, and territoriality. Through the process of natural selection, evolution has endowed all organisms with defensive mechanisms (e.g., chemical toxins in plants; fight-flight responses in animals) for self-preservation. Fear serves this same purpose. In humans and other higher functioning animals, ‘fear’ as a scientific term describes a behavioral–cognitive–emotional state in which a set of adaptive responses is activated in the presence of danger. As the term is more commonly used, ‘fear’ refers to a sensation (or feeling) of strong unpleasant emotion aroused by stimuli or events that are perceived as threatening or dangerous.

While this more anthropomorphic (psychological) sense of the word ‘fear’ may be useful in describing a person who is experiencing an environmental threat (e.g., electric shock) and exhibiting certain stereotyped behaviors (such as a fearful facial expression, autonomic arousal, behavioral readiness, and other defensive responses), it is probably inaccurate to use the term to describe a goldfish (or any other lower-functioning animal) that is exhibiting similarly stereotyped behaviors in response to electric shock (such as avoidance or escape). In this entry, fear will be defined as a by-product (or an emergent property) of a coordinated pattern of brain activities and bodily changes that appear in response to threats and dangers. Thus, fear serves an adaptive function, but evolved relatively recently among the continuum (or phylogeny) of defensive mechanisms.

Fear as an emotion has been a topic of interest to many subdisciplines of psychology, including psychoanalysis, behaviorism, cognitive psychology, and

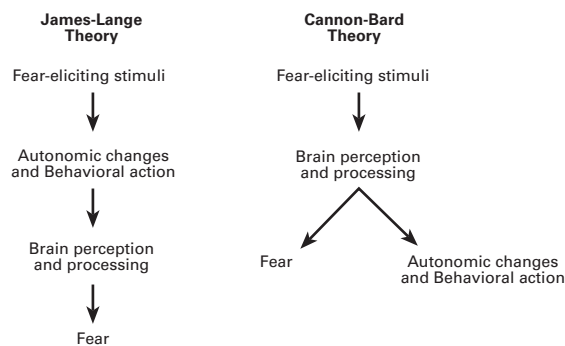


Figure 1

Two influential theories of emotion propose different sequences by which threats and danger trigger fear. The main tenet of the James–Lange theory is that the animal feels fear after its body reacts (emphasizing bodily feedback in the brain representation of fear). In contrast, the Cannon–Bard theory states that threats and danger simultaneously activate the brain representation of fear and the bodily responses (bodily feedback is not a precursor of the brain representation of fear)

biological psychology (for detailed treatment and further references, see LeDoux 1996). Historically, two opposing theories of emotion, the James–Lange theory and the Cannon–Bard theory, have been especially influential (Fig. 1). According to the James–Lange theory, threatening or dangerous stimuli trigger physiological arousal and behavioral responses (via the autonomic nervous system), which are perceived and then processed as fearful by the brain (emphasizing the importance of bodily feedback to the brain). In contrast, the Cannon–Bard theory posits that threats and dangers simultaneously trigger brain activities representing fear and initiate physiological arousal and behavioral readiness (emphasizing the importance of the brain activity). Thus, the critical distinction between the two theories is whether the appearance of immediate danger (e.g., a bear) is postulated to activate the central fear system in parallel with fleeing and autonomic responses (Cannon–Bard theory), or to first trigger fleeing and autonomic responses, which are then detected and interpreted as fear by the brain (James–Lange theory). There is evidence in support of both theories, suggesting that central processes and bodily feedback (as well as other, cognitive processes) both contribute to the emergence of fear as an emotion.

2. Neuronal Substrates of Fear

Because humans share many biological and behavioral features with other animals, experimental studies with nonhuman animals provide valuable information

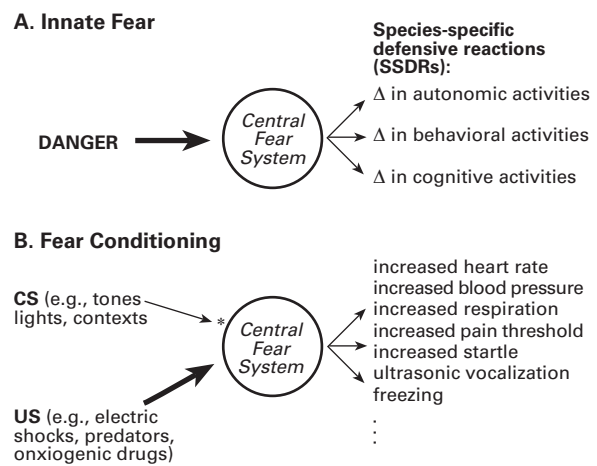


Figure 2

Simple models of fear system and fear conditioning. A. Innate fear: presence of danger triggers a coordinated set of programmed behavioral physiological and cognitive responses, which rapidly prepares the animal to contend with the immediate threat. B. Fear conditioning: fear of novel stimuli can be acquired through the process of classical conditioning. By virtue of association with the US, the CS comes to activate fear

* The putative locus of neural plasticity involved in fear conditioning

toward understanding the physiological as well as the psychological aspects of fear. Darwin postulated in *The Expression of Emotions in Man and Animals* (1872) that emotions such as fear are innate, highly specific to situations, and the product of coordinated brain activities, and that different animal species with diverse environmental demands (or histories) have evolved dissimilar specialized sets of fear (or defensive) responses to maximize survival. It follows from this line of reasoning that more related species will exhibit more similar fear responses than will less related species, whose nervous systems share fewer characteristics. Thus, it is highly likely that the neuronal substrates underlying fear (or other defensive mechanisms) in humans are more similar to those of other mammals (e.g., chimpanzees and rats) than those of nonmammalian animals (e.g., fish and birds).

A repertoire of innate fear responses that is characteristic of a particular animal species is often referred to as species-specific defensive reactions (SSDRs) (Bolles 1970) (Fig. 2, A). For instance, the universal expressions of fear in humans include particular facial expressions (seen in members of all cultures), changes in autonomic nervous system activities (e.g., increased heart rate, increased blood pressure, rapid respiration, perspiration, elevated pain threshold, pupil enlargement, and decreased digestive activity) and beha-

vioral readiness (e.g., freezing, increased propensity to startle, and sharpening of cognition) (see LeDoux 1996). These fear responses in humans are 'hardwired' (innate) as indicated by the observation that newborn infants, when exposed to aversive stimuli (e.g., a loud noise) for the very first time, exhibit responses, such as sudden eye closure, rapid breathing, clutching of hands, and crying, which are indicative of a general state of fear (e.g., Watson 1924). The innateness of the fear response is also demonstrated in animals; for example, laboratory-reared rats confronted with danger (e.g., a cat) for the very first time exhibit changes in autonomic nervous system activities and behavioral readiness (responses that are remarkably similar to those seen in humans), and ultrasonic (22kHz) distress vocalization (a response that is an example of an SSDR) (e.g., Blanchard and Blanchard 1971, Blanchard et al. 1991, Bolles 1970).

Several lines of evidence point to temporal lobe structures (specifically the amygdala, an almond-shaped group of nuclei buried deep within the temporal lobes) as the key neural system underlying the central state of fear (see LeDoux 1996). The classic studies of Klüver and Bucy (e.g., 1937), who examined the effect of temporal lobe ablations in rhesus monkeys, were pioneering in this respect. Klüver and Bucy noted dramatic postlesion behavioral changes, including hyperactivity, hypersexuality, and a so-called 'psychic blindness' that was manifested as an apparent indifference to the significance of previously frightening and/or arousing stimuli. While the monkeys were wild and aggressive prior to surgery, they were remarkably tame and agreeable to human contact afterwards, suggesting that a crucial emotive brain center had been removed or significantly affected by the lesions. Subsequent work, such as that of MacLean and Delgado (1953) using focal stimulation of temporal lobe structures in awake, behaving monkeys, pointed to the amygdala as the critical site. The importance of the amygdala in fear has been reconfirmed in numerous studies employing diverse species of animals (e.g., rats, mice, cats, rabbits, monkeys) and various experimental techniques (e.g., lesions, stimulation, drug infusions, recordings), and most current theories of amygdalar function reflect the importance of the structure in the processing of emotional stimuli (see LeDoux 1996). Consistent with this view, recent human neuropsychological and brain-imaging studies have found that patients with amygdalar damage are selectively impaired in the recognition of facial expressions of fear (Adolphs et al. 1994), and that there is a significantly increased blood flow to the amygdala (as measured by functional magnetic resonance imaging, fMRI) when normal subjects are presented with pictures of fearful faces (Morris et al. 1996).

The amygdala, as one of the principal structures of the limbic system, has access to sensory input from diverse areas of the brain (e.g., olfactory cortex,

hippocampus, thalamus, hypothalamus, neocortex), and projects to various autonomic and somatomotor structures that mediate specific fear responses (e.g., bed nucleus of stria terminalis for activating stress hormones, periaqueductal gray matter for defensive behavior, lateral hypothalamus for sympathetic activation) (see LeDoux 1996).

Given this pattern of connectivity, it is not surprising that damage to the amygdala eliminates stereotypical fear responses, such as defensive freezing (characterized as a motionless, crouching posture that functions to reduce the probability of detection by predators) exhibited by rodents in the presence of predators (Blanchard and Blanchard 1971). Nevertheless, while the amygdala is apparently an important component of the central fear-processing system, it does not necessarily follow that the amygdala is the seat of fear as an emotion (or a feeling). Given its intricate reciprocal connections with the neocortex, a structure implicated in higher cognitive functioning, it is very likely that the interactions between amygdala and neocortex (particularly sensory and limbic association cortices) are responsible for the emotion of fear. Thus, although the amygdala is essential in activating various fear responses (or SSDRs) in both humans and rats, the fact that there are great differences in the neocortices of humans and rats suggest that perhaps there might be important differences in fear as experienced by the two species.

3. Acquired Fear

While humans and other animals are genetically predisposed to fear certain stimuli/events, it is also adaptive for animals to have the capacity to learn about new dangers in their environments. For instance, although newborn infants innately exhibit fear to certain stimuli (e.g., loud noises), they do not show innate fear to flame or height (two stimuli that most children learn and adults remember to avoid). Thus, fear responses to many stimuli and events in the environment appear to be acquired.

Classical or Pavlovian fear conditioning has been widely employed for studying the mechanism(s) by which fear is acquired. Fear conditioning occurs when initially neutral conditioned stimuli (CSs) such as tones, lights, or distinctive environmental settings, are contingently paired with aversive unconditioned stimuli (USs) such as electric shocks, which reflexively activate unconditioned fear responses (URs). Through CS-US association formation, the CS comes to elicit various conditioned fear responses (CRs) that are similar to innate fear responses (Fig. 2, B). Perhaps the best known example of fear conditioning is the Little Albert experiment by Watson and Rayner (1920). Little Albert was an 11-month-old infant who initially exhibited curiosity (no fear) to a white rat by touching and playing with it. Watson and Rayner attempted to

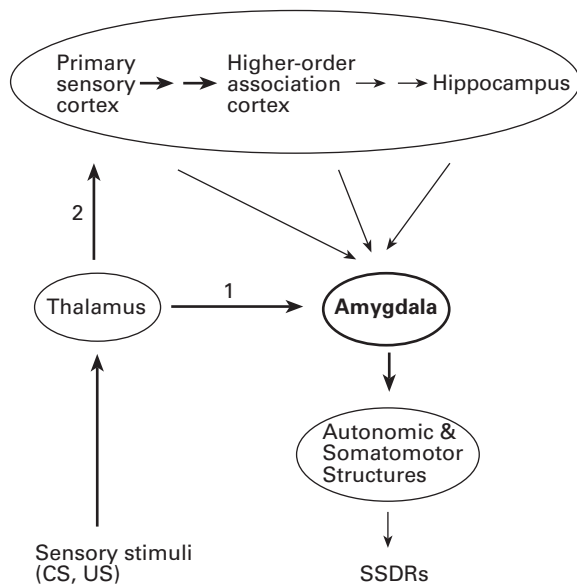


Figure 3

A simplified putative fear-conditioning circuit. Conditioned stimuli (e.g., tones) are processed via two separate pathways to the amygdala (where fear conditioning is thought to take place). One pathway is via the direct (1) thalamo-amygdalar projections. The second pathway is via the indirect (2) cortico-amygdalar projections

Source: Adapted from Armony and LeDoux 1997

condition Albert to fear the rat by striking a steel bar with a hammer behind his head (US) as his hand touched the rat, causing him to startle, fall forward, and cry (UR). A week later, when the rat (CS) was placed near Little Albert's hand, he withdrew his hand and began to cry (CR). This fear of the rat was also said to be observed in the presence of other white furry animals and objects (e.g., rabbits, dogs, fur muffs).

Modern investigations of fear conditioning typically employ rats as subjects and use a tone as a CS and a mild footshock as an US. Under these circumstances a single CS-US pairing results in robust (rapid and enduring) fear learning, as evidenced by a variety of conditioned fear responses exhibited upon subsequent presentations of the tone. Thus, the biological significance of fear conditioning is reflected by the quick and permanent nature of changes in the behavior toward the initially neutral stimulus that has been paired with an aversive experience.

Fear can also be rapidly acquired through instrumental or operant conditioning, in which the presentation of an aversive stimulus is contingent upon the behavior of the animal. A widely employed procedure with rodents is the passive (or inhibitory) avoidance task (see McGaugh et al. 1996). In this paradigm, the animal's response (e.g., entering a dark compartment

of a box when placed in an adjacent lighted compartment, or stepping down from a platform onto a grid floor) is paired with an aversive experience (e.g., a footshock). As a function of this response-stimulus pairing, the animal learns to avoid making the response that was followed by the aversive experience.

Similar to innate fear, evidence from lesion, pharmacological, recording, and stimulation studies indicates that the amygdala is essential not only to innate fear, but also to the acquisition of fear, as with fear conditioning (see Lavond et al. 1993, LeDoux 1996). Thus, amygdalar lesions (particularly lesions of the central nucleus) impair the acquisition (learning) and retention (performance) of various fear-related behaviors in rats, rabbits, and cats; neurons in the central nucleus of the amygdala are responsive to both conditioned and innate fear stimuli and undergo plastic changes during fear conditioning; and restricted stimulation of the amygdala elicits specific fear responses (e.g., freezing). Recent findings from human studies that patients with amygdalar damage are impaired in fear conditioning, and that blood flow to the amygdala is increased during fear conditioning in normal subjects, further add to the evidence that the amygdala is crucially involved in fear learning.

One hypothesis of amygdalar function in fear conditioning is presented schematically in Fig. 3. In this model, an emotional stimulus (such as a tone that has been paired with a footshock) is presented, and its sensory characteristics are processed by way of two parallel pathways that ultimately converge in the amygdala. One of these pathways is 'direct' and provides relatively 'raw,' unprocessed information about the stimulus, but does so very quickly. If the neocortex is the seat of consciousness, this thalamic pathway to the amygdala can be considered the subconscious (nondeclarative or implicit) fear route. The second pathway is 'indirect' and provides highly processed information about the stimulus, but does so at the cost of an increased processing time (LeDoux 1996).

This cortical (conscious, declarative, or explicit) pathway consists of projections through primary auditory cortex to higher-order auditory association areas and the hippocampal formation, which then project to the amygdala. Upon reaching the amygdala (specifically the lateral/basolateral nuclei), these two converging representations of the stimulus are subjected to additional processing within the amygdala itself. Most current theories assume that the essential plasticity underlying this form of learning occurs in the amygdala and corresponds to a neural instantiation of the association between the CS (the tone) and the US (the footshock) long assumed by psychologists to lie at the heart of classically conditioned behavior. The precise functions of the inputs to and outputs from the amygdala, as well as the intraamygdaloid processing of CS- and US-related information, remain a topic of intensive research.

4. Fear Conditioning and Clinical Implications

Fear conditioning serves as a model system for understanding not only the neural basis of acquired fear in nonhuman animals, but also the etiology of a variety of fear-related disorders in humans (e.g., anxiety, phobias, post-traumatic stress disorder, and panic disorder). Since the characteristic behavioral and physiological symptoms of human anxiety disorders closely resemble many of the fear responses observed in fear conditioning paradigms, it is possible that anxiety and fear are closely related (Table 1). (It is important to note, however, that the fact that the symptomatic manifestation of anxiety disorders is similar to the fear responses observed in fear conditioning does not necessarily mean that the underlying neuronal substrates of fear and anxiety are the same (*post hoc, ergo propter hoc*).) While anxiety disorders can arise unpredictably, it is known that most patients experience their first symptoms in early adult life, indicating that maladaptive anxiety may be at least partly acquired. Indeed, clinicians from the time of Freud have speculated that anxiety patients have acquired their disorder as a result of traumatic learning experiences. Since the paradigm *par excellence* of traumatic learning is fear conditioning, fear conditioning appears to be useful in understanding the genesis of anxiety disorders (Fig. 3).

The question of whether anxiety is analogous to fear conditioning might be one of semantics; the difference is between normal, adaptive fear and abnormal, maladaptive fear (anxiety, phobias). However, there is an important distinction between anxiety disorders and fear conditioning (e.g., Seligman 1971) that might be taken to suggest that the two have very different bases. That is, while conditioned fear responses may be extinguished by, for example, actively preventing an animal from making escape or avoidance responses to a danger-signaling stimulus, fear-related disorders in humans (e.g., phobias) are quite resistant to extinction. In fact, anxiety disorders are characterized by fears that the patients recognize to be irrational but are unable to control.

Based on Fig. 3, it is interesting to speculate that fear-related disorders (irrational fears) might be due to

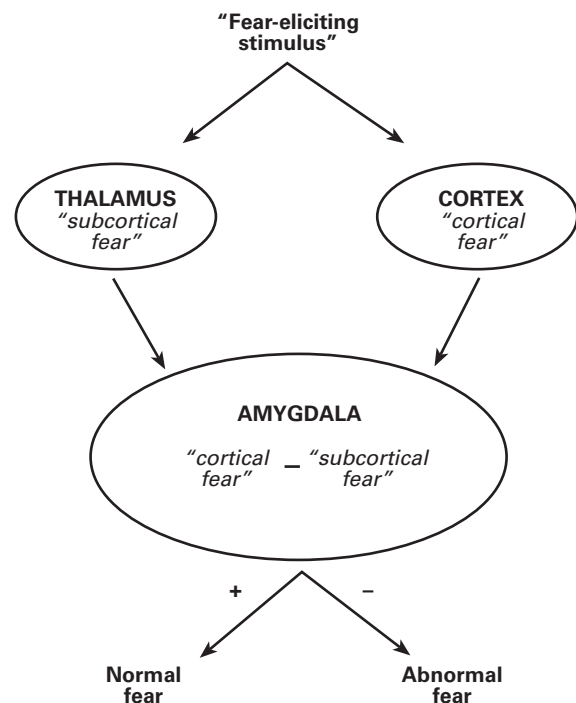


Figure 4

A simple model of normal and abnormal fear based on the putative circuitry of cortical and subcortical fear-conditioning pathways. The subcortical fear route provides a quick but imprecise representation of sensory information for activating fear (a nonconscious pathway), whereas the cortical fear route provides relatively slower but more detailed information for activating fear (a conscious pathway). In normal fear expression there is a balance between the cortical fear and subcortical fear pathways, such that the cortical component is stronger than or equal to the subcortical component. Abnormal fear might arise when this balance is disrupted such that the activity of the subcortical fear system overwhelms the activity of the cortical fear system

Table 1

Behavioral and physiological responses associated with anxiety disorders and fear conditioning

Anxiety disorders	Fear conditioning
Palpitations; pounding heart, or accelerated heart rate	Increased blood pressure, tachycardia
Sweating	Increase in Galvanic skin response measures
Sensations of shortness of breath or smothering	Increase in respiration, quick, shallow breathing
Nausea or abdominal distress	Ulcer
Paresthesia (numbness or tingling sensations)	Hypoalgesia or analgesia
Exaggerated startle responses	Potentiated startle

unbalanced activities in the cortical and thalamic pathways to the amygdala. For example, in most people a snake seen behind a glass window will elicit a different response than will a snake encountered in the woods. Both cortical and subcortical routes will process information about the snake, but the cortical (conscious) route will integrate the contextual information of the glass window and assess the safety of the situation (possibly acting as a gate). However, if the cortical pathway is somehow disrupted, the subcortical activity will be left unchecked to activate fear, even in the inappropriate context. It is possible that this type of unregulated subconscious activation of fear contributes to the etiology of fear-related disorders.

Figure 4 illustrates a hypothetical model of normal, adaptive fear and abnormal maladaptive fear. It is conceivable that genetic predisposition and environmental factors (e.g., stress) interact in such a way as to disrupt the normal balance between cortically activated fear and subcortically activated fear, leading to fear-related disorders. If it is then supposed, as some have suggested (LeDoux 1996), that the direct thalamic inputs to the amygdala are relatively resistant to extinction, whereas the cortical inputs are not, it follows that fear disorders based on overactivity in the subcortical pathway should be 'irrational,' in the sense that they do not subside even when the patient recognizes them as such. A more comprehensive understanding of the neuronal substrates and, subsequently, the molecular mechanisms of the fear system will be useful in understanding and developing treatments for fear-related disorders.

5. Conclusion

The study of fear, a form of emotion that evolved relatively recently among the phylogeny of successful defensive mechanisms, is but one means of understanding emotion, but it is the most fruitful one to have been employed to date. This is likely due to the primacy of fear: unlike other emotions, fear is apparently shared by a wide variety of species. This ubiquity speaks to its importance as a mechanism for self-preservation.

See also: Amygdala (Amygdaloid Complex); Anxiety and Anxiety Disorders; Anxiety and Fear, Neural Basis of; Emotion, Neural Basis of; Emotions, Evolution of; Fear Conditioning; Fear: Potentiation of Startle

Bibliography

Adolphs R, Tranel D, Damasio H, Damasio A 1994 Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature* **372**: 669–72

- Armony J L, LeDoux J E 1997 How the brain processes emotional information. *Annals of the New York Academy of Science* **821**: 259–70
- Blanchard R J, Blanchard D C 1971 Defensive reactions in the albino rat. *Learning and Motivation* **2**: 351–62
- Blanchard R J, Blanchard D C, Agullana R, Weiss S M 1991 Twenty-two kHz alarm cries to presentation of a predator, by laboratory rats living in visible burrow systems. *Physiology and Behavior* **50**: 967–72
- Bolles R C 1970 Species-specific defensive reactions and avoidance learning. *Psychological Review* **71**: 32–48
- Darwin C 1872 *The Expression of the Emotions in Man and Animals*. Murray, London
- Klüver H, Bucy P C 1937 'Psychic blindness' and other symptoms following bilateral temporal lobectomy in rhesus monkeys. *American Journal of Physiology—Regulatory Integrative and Comparative Physiology* **119**: 352–3
- Lavond D G, Kim J J, Thompson R F 1993 Mammalian brain substrates of aversive classical conditioning. *Annual Review of Psychology* **44**: 317–42
- LeDoux J 1996 *The Emotional Brain: the Mysterious Underpinning of Emotional Life*. Simon and Schuster, New York
- MacLean P D, Delgado J M R 1953 Electrical and chemical stimulation of frontotemporal portion of limbic system in the waking animal. *EEG Clinical Neurophysiology* **5**: 91–100
- McGaugh J L, Cahill L, Roozendaal B 1996 Involvement of the amygdala in memory storage: Interaction with other brain systems. *Proceedings of the National Academy of Sciences of the United States of America* **93**: 13508–14
- Morris J S, Frith C D, Perrett D I, Rowland D, Young A W, Calder A J, Dolan R J 1996 A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature* **383**: 812–15
- Seligman M E P 1971 Phobias and preparedness. *Behavioral Therapy* **2**: 307–20
- Watson J B 1924 *Behaviorism*. Norton, New York
- Watson J B, Rayner R 1920 Conditioned emotional reactions. *Journal of Experimental Psychology—Human Perception and Performance* **3**: 1–14

J. J. Kim and K. M. Myers

Feature Representations in Cognitive Psychology

One of the most basic assumptions of cognitive psychology is that entities in the world can be described in terms of simpler components called features. For example, when asked to list the features that tend to characterize a bird, the typical undergraduate would mention that a bird has wings, flies, lays eggs, builds nests, has a beak, has feathers, and so on. This article discusses psychological research on how people represent features, and examines a number of challenges to developing a theory of feature representation.

1. Introduction

Cognitive psychologists have often attempted to characterize the entities of a domain in terms of a

relatively small set of such features, with different entities having different combinations of these features. To take a simple example, consider capital letters. Each capital letter can be described uniquely by some combination of seven basic features: vertical, horizontal, or oblique lines; continuous or discontinuous curves; and right or acute angles between the junctions of lines. For instance, the letter R has a vertical line, two horizontal lines, an oblique line, a discontinuous curve, and three right angles. In contrast, the letter O has a single feature—a continuous curve.

Being able to characterize the entities of a domain in this manner would be an important step in specifying how entities are identified. For example, suppose that in a person's long-term memory, capital letters were represented as combinations of basic features like those described above. Then by using processes that identified such features and matched them to ones represented in long-term memory, a person could identify a letter. For example, upon detecting an oblique line and a continuous curve, a person would determine that these features matched those of the representation of Q and identify the letter as a Q. Evidence suggests that the human visual system has feature detectors (i.e., neurons). For example, some neurons respond to lines of different orientations and to junctions of lines (i.e., corners) (Hubel and Wiesel 1965).

Identifying entities is the starting point for many important, everyday cognitive tasks. In this example, identifying the letter would be an important step in identifying a word, which in turn would be an important step in reading. As another example, by identifying an object, we gain access to a large amount of stored information that is useful in interacting with that object. For instance, by recognizing something as a bear you can then access stored knowledge which lets you anticipate things about the bear that may not be immediately present (e.g., that the bear may approach and attack you).

In many domains, the representation (and subsequent processing) of features is considerably more complex than implied by the example of capital letters. In contrast to capital letters, it is not always clear what features characterize an entity. In addition, multiple factors often determine how a feature is identified. Further, relations between features can be important for identifying entities. Finally, people not only identify features but must also learn new ones. In the remainder of this article, I discuss these and other issues related to feature representation and processing.

2. Context-dependent Feature Representations

The features that we use to represent entities in the world are often influenced by the surrounding context (what other information is present at encoding, our expectations, etc.). A classic example is illustrated by a

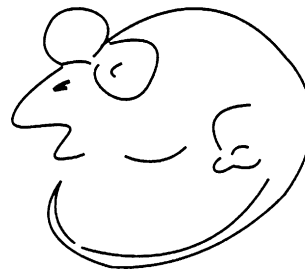


Figure 1

The rat-man figure used by Bugelski and Alampay (1961)

study involving the rat-man figure (see Fig. 1). Bugelski and Alampay (1961) found that when preceded by several pictures of human faces people tend to see the rat-man figure as a man, but when preceded by several pictures of animals they tend to see the rat-man figure as a rat. In these cases, people represent the same perceptual input in terms of different features. For example, the small circles in the figure are seen as ears or as eyeglasses depending on the preceding pictures. As another example, whether a white area on an X-ray indicates a lung tumor, a bone, or an artifact of the procedure can depend on a variety of contextual factors, including the radiologist's knowledge of the patient's case history, the location of bones, whether the X-ray is a chest X-ray, and the likelihood that the X-ray was underexposed. These and many other examples suggest that the representation of a feature is not just based on perceptual input but also on what is in the head of the perceiver. As result, the features associated with a perceptual input can vary depending on the perceiver's prior knowledge.

3. Relations between Features

Feature representations often capture relations between other features. These relations are necessary in order to identify an entity accurately. For example, a bucket and a coffee cup can both be described by the same two basic features (an open cylinder and a handle). Thus, it would not be possible to distinguish these objects if they were only represented by such features. Note, though, that the spatial relationship between the two features differs in the two objects. That is, the handle of the bucket is attached to the rim of the open cylinder, whereas the handle of the coffee cup is attached to the side of the open cylinder. Thus, people could distinguish these objects if their featural representations included such spatial relationships. Biederman (1987) describes a model of object recognition that represents objects in terms of part-like features (called *geons*), and the relations between them. As another example, the words 'team' and 'meat' are made up of the same basic features, called

phonemes (i.e., the smallest significant sound units in a language). However, the temporal relationship between these phonemes is different. For example, the 't' sound is pronounced first in 'team,' but last in 'meat.' Thus, people could distinguish these spoken words if their featural representations included such temporal relationships.

4. Construction of New Features

In a number of areas of cognitive psychology, the researcher and subject agree implicitly on the features which characterize the cognitive task being investigated. For example, in studying how people learn about new categories of entities, researchers often present subjects with stimuli whose features are well-specified, familiar, and unambiguous. The question of interest is how people determine which features are important for distinguishing one category from another. However, learning does not just involve selecting relevant features. In any domain, people must sometimes figure out the features themselves (i.e., acquire new features; see Biederman and Shiffrar 1987, Lesgold et al. 1988, Schyns et al. 1998, Wisniewski and Medin 1994, for examples).

To illustrate this point, consider a study by Schyns and Rodet (1997) in which subjects learned about categories of novel stimuli called 'Martian cells' (see Fig. 2). One group of subjects first learned a category whose members had x blobs followed by a category

with y blobs, and then a category with xy blobs (a fusion of the separate x and y blobs; see Fig. 2a). A second group learned the categories in the reverse order (see Fig. 2b). These different orders affected how subjects determined the features of the category with xy blobs. The first group perceived the xy blob as a conjunction of the x and y features. They had previously learned that x was diagnostic of the x blobs category, and that y was diagnostic of the y blobs category. Thus, previously learning the x and y features biased the first group to segment the xy blob into these two features. In contrast, the second group perceived the xy blob as a single, unitary feature xy . Because the initial category that they learned was the one with xy blobs, their perception of the xy blob was not influenced by having learned features x and y .

5. Local vs. Global Feature Representations

In many areas of cognitive psychology, researchers use symbols to represent features that correspond closely to language, and which tend to bottom out at the level at which the language bottoms out (Solomon and Barsalou in press). For example, in the English language, people typically use the word 'stripes' to refer to the stripes of zebras, barber poles, the American flag, and so on, rather than a more detailed term. In turn, researchers represent features using similar, language-like terms. Often, these representations take the form of attribute-value pairs. For

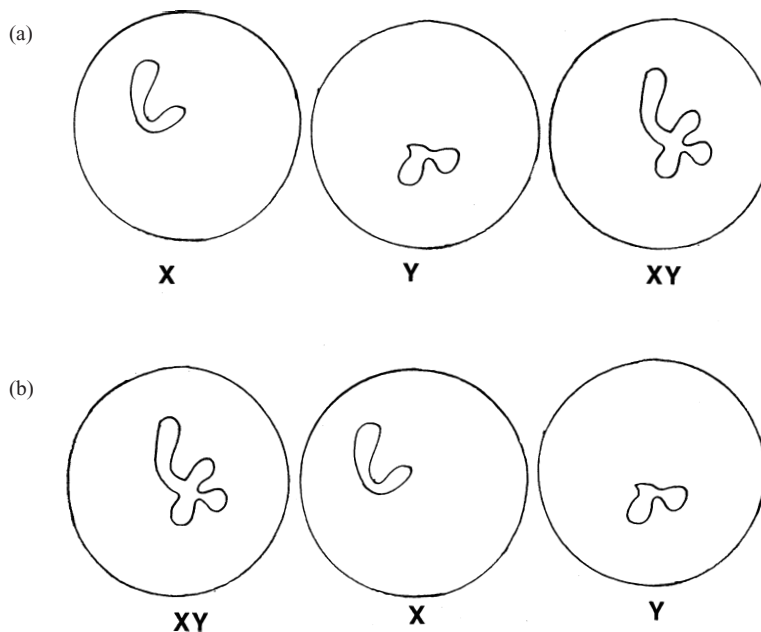


Figure 2

Schematic drawing of two different orders of novel categories learned by subjects in a study by Schyns and Rodet (1997)

example, stripes could be represented as the attribute texture. Do researchers believe that such representations have no further specification of meaning? They probably do not, as intuition strongly suggests that there is something more to such representations. Instead, such a feature representation may be a shorthand term for the commonalities shared by its referents (which researchers usually leave unspecified). For example, one possible commonality among the stripes of entities is that they have a color which contrasts with the background color of the entity. Thus, researchers generally assume that a feature has a similar representation across the entities to which it applies (i.e., a global representation).

However, evidence suggests that feature representations are also more fine-grained and specific or local to their associated entities. For example, Solomon and Barsalou (in press) tested the local representation view in a series of property verification tasks. Subjects saw an object (e.g., a bus) followed a property name (e.g., seat) and had to verify whether the property was true of the object. On a subsequent trial, verifying that property for a different object was facilitated if the property had the same form in the second object. For example, having previously verified that seat is true of bus, subjects more quickly verified that seat was true of truck. However, this benefit was not obtained when verifying that seat was true of bicycle. This finding implies that the representation of seat in truck and bus is similar, but differs from that in bicycle. As another example, many properties are relative. For instance, a large mouse and a large house are not very similar in size. A large mouse is large compared to other mice, but a large house is large compared to other houses. Thus, relative properties also have local representations.

6. Determining the Appropriate Features

As much of the discussion above implies, a central issue in feature representation is determining the appropriate features of an entity. In principle, one can describe any entity by an infinite number of features (e.g., a German shepherd dog has 4 legs, but it also has fewer than 6 legs; it has a tail, but also a tail of a particular length; it is found on the planet Earth, but not at the bottom of the ocean; it has a nose, but it also has a nose of a particular color and shape, and so on). Clearly, there must be constraints on what counts as a feature (Murphy and Medin 1985). Many of these constraints are provided by the perceptual system (see Barsalou 1999, Schyns et al. 1998, for discussion of perceptual mechanisms that extract features). For example, the perception of an object part is influenced by discontinuities and points of maximum curvature on its surface (e.g., Hoffman and Richards 1984).

Other constraints are related to the categories to which an entity belongs. For example, a small, umbrella-shaped entity growing in a forest might be

categorized as a mushroom. The representation of mushroom might contain the features stem and umbrella-shaped cap so as to distinguish it from other plants. On the other hand, a mushroom belongs to a more specific category such as Death Cap—the most poisonous mushroom. Its representation might contain the features green-tinged cap and thick stem, so as to distinguish it from other mushrooms. Thus, the features associated with a category may be partially constrained by its contrasting categories. Further, the same aspect of an entity may have multiple feature representations (e.g., cap, green-tinged cap).

7. Final Comment

Ultimately, any theory of how features are represented must address the symbol grounding problem (Harnad 1990). That is, the representations of features that are in our heads must make contact with what they represent—what is out in the world. Much of the utility of representations is in allowing us to interact successfully with the world. For example, the representation of the green-tinged cap must allow mushroom pickers to identify actual green-tinged caps of mushrooms so that they can avoid eating them. As another example, a relatively abstract feature such as ‘plays music’ for a CD player is not particularly useful unless it indicates more specifically how one can use a CD player to play music.

See also: Concept Learning and Representation: Models; Face Recognition Models; Letter and Character Recognition, Cognitive Psychology of; Neural Representations of Objects

Bibliography

- Barsalou L W 1999 Perceptual symbol systems. *Brain and Behavioral Sciences* **22**: 577–609
- Biederman I 1987 Recognition-by-components: A theory of image understanding. *Psychological Review* **94**: 115–47
- Biederman I, Shiffrar M M 1987 Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task. *Journal of Experimental Psychology: Learning, Memory and Cognition* **13**: 640–45
- Bugelski B R, Alampay D A 1961 The role of frequency in developing perceptual sets. *Canadian Journal of Psychology* **15**: 205–11
- Harnad S 1990 The symbol grounding problem. *Physica D* **42**: 335–46
- Hoffman D D, Richards W 1984 Parts of recognition. *Cognition* **18**: 65–96
- Hubel D H, Wiesel T N 1965 Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology* **28**: 229–89
- Lesgold A M, Robinson H, Feltovich P, Glaser R, Klopfer D, Wang Y 1988 Expertise in complex skill: Diagnosing X-ray pictures. In: Chi M T H, R Glaser Farr M (eds.) *The Nature of Expertise*. Erlbaum, Hillsdale, NJ, pp. 322–51
- Murphy G L, Medin D L 1985 The role of theories in conceptual coherence. *Psychological Review* **92**: 289–316

- Schyns P G, Rodet L 1997 Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory and Cognition* **23**(3): 681–96
- Schyns P G, Goldstone R L, Thibaut J P 1998 The development of features in object concepts. *Behavioral and Brain Sciences* **21**: 1–54
- Solomon K O Barsalou L W (in press) Representing properties locally: Evidence for the perceptual grounding of concepts. Manuscript submitted for publication
- Wisniewski E J, Medin D L 1994 On the interaction of theory and data in concept learning. *Cognitive Science* **18**: 221–81

E. Wisniewski

Fechnerian Psychophysics

Fechnerian psychophysics is based on the idea that an appropriately chosen measure of local discriminability in a continuous stimulus space (the degree with which an observer can discriminate a stimulus from its very close neighbors) can be used to compute ‘subjective’ distances between very close stimuli; and that by integrating these small distances appropriately along paths connecting different stimuli one can compute ‘subjective’ (Fechnerian) distances among all stimuli comprising the space. This idea, together with experimental procedures for measuring local discriminability, was originally proposed by Fechner (1860, 1877) for unidimensional stimulus continua (such as the space of fixed-frequency tones varying in intensity, or visually presented line segments varying in length). Historians often date the beginning of scientific psychology from Fechner’s work.

1. Classical Theory (Refined)

If a stimulus x changes along a single physical dimension (an interval of reals), and $F(x)$ is the *local discriminability measure* at x (positive and continuous), then the *Fechnerian distance* between any two stimuli a, b is defined as

$$G(a, b) = \left| \int_a^b F(x) dx \right| \quad (1)$$

The Fechnerian distance clearly satisfies the defining properties of a metric: $G(a, a) = 0$, $G(a, b) > 0$ for $a \neq b$, $G(a, b) = G(b, a)$, and $G(a, b) \leq G(a, x) + G(x, b)$.

Assuming that x is much greater than the *lower absolute threshold* x_{inf} of the stimulus continuum, so that the random variability in x_{inf} can be ignored, the distance $G(x, x_{\text{inf}})$ can be called the ‘Fechnerian magnitude’ of x (traditionally, ‘sensation magnitude’).

Let $\xi_x(y)$ be the probability with which y is perceived to be greater than x , and assume that $\xi_x(x) = 1/2$ (the

latter can be ensured by an appropriate recalibration of the reference stimuli x). Then the *local discriminability measure* $F(x)$ is defined as

$$F(x) = C \frac{d\xi_x(y)}{dy} \Big|_{y=x+} \quad (2)$$

where certain regularity assumptions ensure that the derivative exists and continuously changes with x ; $C > 0$ is a proportionality coefficient allowed to be different for different stimulus continua.

For $1/2 \leq p < 1$, let $W_p(x)$ denote the value of y at which $\xi_x(y) = p$; then $W_{1/2}(x) = x$. The difference $W_p(x) - x$ is called the ‘just noticeable difference’ (JND) at x on level p , or p -JND, for short. Putting $y = W_p(x)$ and using the identity $\xi_x[W_p(x)] = p$, Eqn. (2) can be written as

$$F(x) = C \frac{dp}{dW_p(x)} \Big|_{p=1/2+} = C \lim_{p \rightarrow 1/2+} \frac{p - (1/2)}{W_p(x) - x} \quad (3)$$

Since, by Eqn. (1)

$$\begin{aligned} \lim_{p \rightarrow 1/2+} \frac{G[a, W_p(a)]}{p - (1/2)} &= \frac{dG[a, W_p(a)]}{dp} \Big|_{p=1/2+} \\ &= \frac{d \int_a^{W_p(a)} F(x) dx}{dp} \Big|_{p=1/2+} \end{aligned}$$

one can use the chain rule and apply Eqn. (3) to $F[W_p(a)]|_{p=1/2+} = F(a)$ to show that

$$\lim_{p \rightarrow 1/2+} \frac{G[a, W_p(a)]}{p - (1/2)} = C \quad (4)$$

By fixing p at a value sufficiently close to $1/2$ one concludes that $G[a, W_p(a)]$ is approximately constant for all a , because of which $G(a, b)$ in Eqn. (1) is roughly proportional to the number of chained p -JNDs, $W_p(a) - a$, $W_p[W_p(a)] - W_p(a)$, etc., that fit between a and b (see Dzhafarov and Colonius 1999 for details). The p -JNDs can be estimated from empirical data by a variety of techniques (see, e. g., Falmagne 1985). The use of the methods of *limits* and *adjustments* in Fechnerian psychophysics, as these provide JNDs on unknown probability levels, is predicated on the assumption that these JNDs are roughly proportional to p -JNDs on a constant level p sufficiently close to $1/2$.

2. Psychophysical Laws

If, with some choice of the physical measure for x , $d\xi_x(y)/dy$ at $y = x$ is proportional to $1/x$ (Weber’s law), then Eqns. (1) and (2) yield $G(a, b) = C \log(a/b)$

(Fechner's psychophysical law). In spite of the enormous importance attached to Weber's and Fechner's laws in the history of psychology, the logic of Fechnerian psychophysics by no means implies them and is not contingent on their validity. Fechnerian psychophysics is equally consistent, for example, with such often found empirical approximations as $F(x) \propto 1/(x + \delta)$ (modified Weber's law) or $F(x) \propto 1/x^\alpha$ (called the near miss to Weber's law if α is just below 1, and the Fullerton-Cattell law if $\alpha \approx 0.5$). These two forms of $F(x)$ imply, respectively, $G(a, b) = C \log[(a + \delta)/(b + \delta)]$ and $G(a, b) = C'(b^{1-\alpha} - a^{1-\alpha})$ as substitutes for Fechner's law.

3. Criticisms of Classical Theory

3.1 Infinitesimals vs. Finite Differences

The account in Sect. 1 is a refined interpretation of Fechner's original theory. It is supported by Fechner's writings, but being voluminous and rather lacking in mathematical rigor, these writings lend themselves to alternative interpretations. According to one of them, Fechner believed in 'true' irreducible JNDs, $W_*(x) - x$, and simultaneously maintained that $G[x, W_*(x)] = \text{const}$ ('Fechner's postulate'), and that Eqn. (1) holds with $F(x) = C/[W_*(x) - x]$ compare with Eqn. (3). These propositions, however, are mutually inconsistent (Luce and Edwards 1958). Fechner knew of this criticism, and his 1877 rejoinder indicates that the interpretation it is based on is not correct. Whatever the historical truth, it is a sound scientific strategy to reserve the term 'Fechnerian psychophysics' for the interpretation that preserves the essence of Fechner's approach while freeing it from logical contradictions.

3.2 Fechnerian Distances and Subjective Matching

Consider four stimuli, say, tones with amplitudes and frequencies (A_1, f) , (A_2, f) , (A_1, f') , (A_2, f') , and let (A_1, f) and (A_1, f') be matched in loudness, and the same be true for (A_2, f) and (A_2, f') . Then the 'loudness-wise' Fechnerian distance between (A_1, f) and (A_2, f) should equal that between (A_1, f') and (A_2, f') , and one could expect that the numbers of chained amplitude JNDs fitting between the first two and the last two tones are roughly equal (under the assumption stipulated in Sect. 1). Riesz (1933) found this prediction wrong: for instance, the number of amplitude JNDs between two 4000-Hz tones may exceed that between the corresponding 200-Hz tones by a factor of four. This may be thought to invalidate the Fechnerian theory, as the difference seems too large to be attributed to the approximation error associated with counting finite JNDs (on unknown probability levels). The counter-argument is as follows. If the amplitudes coupled with frequency f and

those coupled with frequency f' can be treated as separate stimulus continua, then the values of C in Eqn. (3) (and hence the approximate proportionality coefficients between Fechnerian distances and JND counts) may be different for f and f' . If, on the other hand, the amplitude-frequency combinations should be treated as comprising a single two-dimensional stimulus space (see Sect. 5), then it is no longer justifiable to measure the Fechnerian distance between (A_1, f) and (A_2, f) along the straight line connecting them. For a detailed treatment of subjectively unidimensional (e.g., 'loudness-wise') discriminations among multidimensional stimuli see Dzharfarov and Colonius (1999).

3.3 Fechnerian Magnitudes vs. 'Directly' Estimated Magnitudes

For certain stimulus continua, called *prothetic* (intensity, length, duration, etc.), the Fechnerian magnitudes $G(x, x_{\text{inf}})$ computed from discrimination data typically are found nonlinearly related to magnitudes obtained by 'direct scaling' methods (Stevens 1975). For instance, if Weber's law holds on a continuum of x , implying $G(x, x_{\text{inf}}) = C \log(x, x_{\text{inf}})$, the observer's numerical estimates of x are often found closely adhering to $S(x) = Kx^\alpha = \exp[\lambda G(x, x_{\text{inf}})]$ (Stevens' psychophysical law). Many have agreed with Stevens that $S(x)$, and not $G(x, x_{\text{inf}})$, is the 'true' measure of the 'sensory magnitude' of x . As no independent definition of sensory magnitude is available, however, the argument is largely semantic. The classical Fechnerian theory does not predict that all judgments of stimuli interpretable in terms of 'how large,' 'how high,' etc., must be proportional to Fechnerian magnitudes (although this happens to be the case for some *metathetic* continua, such as frequencies of constant-amplitude tones). A meaningful question to ask within the framework of Fechnerian psychophysics is whether one can propose a plausible mechanism, process, or computation by which an observer arrives at the $S(x)$ -measures of prothetic stimuli based on their Fechnerian magnitudes $G(x, x_{\text{inf}})$. A naive but refinable example of such a proposal can be found in Ekman (1964).

4. Probability-distance Hypothesis

According to Eqns. (1) and (2), Fechnerian distances $G(a, b)$ are computed from discrimination probabilities $\xi_x(y)$ taken in arbitrarily small vicinities of $y = x$. A natural question to ask, therefore, is whether $\xi_a(b)$ (provided it is not zero or one) is determined uniquely by $G(a, b)$ for all stimulus pairs a, b . More generally: can one find a continuous internal metric $D(a, b)$ and a continuous function f such that

$$\xi_a(b) = f[D(a, b)] \quad (5)$$

and if so, what is the relationship between $D(a, b)$ and the Fechnerian metric $G(a, b)$? Metric $D(a, b)$ is called *internal* (also, *inner* or *intrinsic*) to the given unidimensional continuum if $D(a, b) + D(b, c) = D(a, c)$ for any $a < b < c$ on this continuum. Thus, the Fechnerian metric G is internal. The assumption that Eqn. (5) holds true can be called the *probability-distance hypothesis*, and the problem of finding D and f satisfying this hypothesis is referred to traditionally as ‘Fechner’s problem’ (Luce and Galanter 1963).

Primarily due to the criticism mentioned in Sect. 3.1, Fechnerian psychophysics since the 1960s has been reduced essentially to the probability-distance hypothesis. One should note, however, that the classical theory presented in Sect. 1 does not imply the truth of this hypothesis. At the same time, the implication holds in the reverse direction: Pfanzagl (1962) proved that (subject to some regularity conditions) Eqn. (5) may hold for an internal metric D only if $D(a, b) = G(a, b)$. That is, discrimination probabilities may not depend on any internal metric other than the Fechnerian one.

Falmagne (1985) proved that Eqn. (5) is equivalent to either of the following two conditions: (weak bicancellation) $\xi_a(c)$ is determined uniquely by $\xi_a(b)$ and $\xi_a(c)$, for any a, b, c ; (weak quadruple condition) $\xi_a(b) = \xi_a(b')$ if and only if $\xi_a(a') = \xi_b(b')$.

5. Multidimensional Fechnerian Psychophysics

A generalization of Fechnerian psychophysics to continuous stimulus spaces of arbitrary dimensionality (such as the CIE color space, or space of tones varying in both frequency and amplitude) was proposed by Dzhamalov and Colonius (1999). A special form of this generalization, however, can be traced back to the tradition, originated by Helmholtz and Schrödinger, of computing colors metrics from color-discrimination data (Wysocki and Stiles 1982). The discrimination of multidimensional stimuli $\mathbf{x} = (x_1, \dots, x_n)$ cannot be defined in terms of ‘greater-less,’ and the function $\xi_a(y)$ of the classical theory should be replaced with $\psi_x(y)$, the probability of y being perceived as *different* from x . Let $\psi_x(y)$ attains its minimum at $y = x$ (if necessary, after an appropriate recalibration of reference stimuli x). The local discriminability measure $F(\mathbf{x}, \mathbf{u})$ at stimulus $\mathbf{x} = (x_1, \dots, x_n)$ in direction $\mathbf{u} = (u_1, \dots, u_n)$ is defined as

$$F(\mathbf{x}, \mathbf{u}) = C \lim_{s \rightarrow 0+} \frac{\Phi[\Psi_x(\mathbf{x} + \mathbf{u}s) - \Psi_x(\mathbf{x})]}{s} \quad (6)$$

where the transformation Φ is one and the same for all $\mathbf{x}$ and $\mathbf{u}$. The assumptions about $\psi_x(y)$ underlying the theory guarantee that $F(\mathbf{x}, \mathbf{u})$ is determined by $\psi_x(y)$ uniquely, and is continuous and positive, while Φ is determined asymptotically uniquely (put roughly,

uniquely in a small vicinity of zero). The local discriminability $F(\mathbf{x}, \mathbf{u})$ can be used to compute the *psychometric length* L of any path $\mathbf{x}(t)|_a^b$, $a \leq t \leq b$, lying within the stimulus space and connecting stimuli $\mathbf{a} = \mathbf{x}(a)$ and $\mathbf{b} = \mathbf{x}(b)$:

$$L[\mathbf{x}(t)|_a^b] = \int_a^b F[\mathbf{x}(t), \dot{\mathbf{x}}(t)] dt \quad (7)$$

The Fechnerian distance between $\mathbf{a}$ and $\mathbf{b}$ is defined as

$$G(\mathbf{a}, \mathbf{b}) = \inf L[\mathbf{x}(t)|_a^b] \quad (8)$$

where the infimum is taken over all paths connecting $\mathbf{a}$ with $\mathbf{b}$. Under certain conditions, this infimum may equal the psychometric length of a certain (‘shortest’) curve connecting $\mathbf{a}$ with $\mathbf{b}$, a *Fechnerian geodesic*. Fechnerian distances are determined uniquely and are invariant with respect to all smooth one-to-one physical reparametrizations of the stimulus space. If $\psi_x(\mathbf{x}) = \text{const}$ (*constant self-similarity condition*), then Fechnerian distances are also invariant with respect to smooth monotone transformations $T[\psi_x(y)]$ (e.g., due to *response bias* changes).

The construction just described identifies the Fechnerian metric as *internal to the stimulus space*. (In general, a distance function is internal to a space if the distance between any two points in the space is the infimum of the lengths of all paths connecting these points and lying entirely within the space.) A multidimensional generalization of Pfanzagl’s theorem on the probability-distance hypothesis holds: $\psi_a(\mathbf{b})$ (different from zero and one) may be determined uniquely by an internal metric $D(\mathbf{a}, \mathbf{b})$ only if $D(\mathbf{a}, \mathbf{b}) = G(\mathbf{a}, \mathbf{b})$.

The multidimensional Fechnerian theory, when specialized to unidimensional continua, sheds additional light on some aspects of the classical theory. While the concept of a unidimensional Fechnerian distance readily generalizes to spaces of arbitrary dimensionality, no natural generalizations exist for the notions of a Fechnerian (‘sensation’) magnitude and psychophysical law, such as Fechner’s or Stevens’. This suggests that the importance traditionally attached to these notions may be unwarranted. By the same criterion (multidimensional generalizability), it is desirable to reformulate the classical theory in terms of the ‘same-different’ discrimination probabilities $\psi_x(y)$, instead of the traditional ‘greater-less’ probabilities $\xi_x(y)$, and accordingly to revise the empirical procedures for computing the local discriminability measure $F(x)$.

6. Concluding Remark

With discrimination among stimuli being arguably the most basic cognitive function, Fechnerian psychophysics is motivated by an expectation that distances

computed from discrimination probabilities should have a fundamental status among behavioral measurements, and that in the final analysis various kinds of perceptual judgments could be shown to depend prominently on Fechnerian distances among stimuli involved. The eventual status of Fechnerian psychophysics will depend on the extent to which this expectation will be confirmed by future experimental and theoretical developments.

See also: Memory Psychophysics; Psychophysics; Signal Detection Theory; Signal Detection Theory, History of

Bibliography

- Dzhafarov E D, Colonius H 1999 Fechnerian metrics in uni-dimensional and multidimensional stimulus spaces. *Psychological Bulletin and Review* **6**: 239–68
- Ekman G 1964 Is the power law a special case of Fechner's law? *Perceptual and Motor Skills* **19**: 730
- Falmagne J C 1985 *Elements of Psychophysical Theory*. Oxford University Press, Oxford, UK
- Fechner G T 1860 *Elemente der Psychophysik* [Elements of Psychophysics]. Breitkopf and Härtel, Leipzig
- Fechner G T 1877 *In Sachen der Psychophysik* [In the Matter of Psychophysics]. Breitkopf and Härtel, Leipzig
- Luce R D, Edwards W 1958 The derivation of subjective scales from just noticeable differences. *Psychological Review* **65**: 222–37
- Luce R D, Galanter E 1963 Discrimination. In: Luce R D, Bush R R, Galanter E (eds.) *Handbook of Mathematical Psychology*. Wiley, New York, Vol. 1, pp. 191–244
- Pfanzagl J 1962 Über die stochastische Fundierung des psychophysischen Gesetzes [On the Stochastic Foundations of Psychophysical Law]. *Biometrische Zeitschrift* **4**: 1–14
- Riesz R R 1933 The relationship between loudness and the minimum perceptible increment of intensity. *Journal of the Acoustical Society of America* **4**: 211–16
- Stevens S S 1975 *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. Wiley, New York
- Wyszecki G, Stiles W S 1982 *Color Science: Concepts and Methods, Quantitative Data and Formulae*. Wiley, New York

E. N. Dzhafarov

Federalism

'Federalism' refers to both an arrangement of political institutions and a philosophy of government. Federalism as a set of institutional arrangements denotes a political system in which power is shared between a central government and regional governments. Federalism as a philosophy of government holds that such federal structures are a desirable governmental form.

1. Federalism and Related Terms

The distinction between 'federalism' and related terms should be clarified. A government displaying the institutional characteristics of federalism is referred to as a 'federal system,' and the institutional structures themselves are described with the adjective 'federal.' A federal system is contrasted with a 'unitary system' of government in which all major decisions are made by the central government; some degree of administrative decentralization, however, may be present in a unitary system (see *Unitary State*). The terms 'federalism' and 'federal system' have been used as overarching terms which encompass a variety of more specific types of political arrangements, including federations, confederacies, and leagues (Riker 1975). The origins of the word 'federal' can be traced to the Latin *foedus*, denoting a league, treaty, or covenant. The term 'intergovernmental relations' refers to interactions between different governmental units within a political system. This term includes but is not limited to the study of interactions between central and regional governments within federal systems.

2. The Normative Study of Federalism

The study of federalism can be usefully divided into normative and empirical branches. The normative branch, which dates back centuries and remains very important today, includes the arguments in favor of the US constitution put forth in *The Federalist*. This normative literature consists of systematic consideration of federalism as a philosophy of government, including the enumeration of possible advantages and disadvantages of federal systems.

Proponents of federalism argue that federal systems are desirable because they may produce a number of advantageous consequences. In particular, advocates of federalism suggest that federal systems protect citizens against tyranny by dividing power among central and regional governments; expand possibilities for political participation by increasing the number of governments within a political system; allow for regional variations in public policy that are reflective of geographic differences in citizen preferences; and provide multiple governmental 'laboratories' in which innovative policy responses to public problems can be tested (Shapiro 1995).

Other scholars have argued that federal systems may produce a variety of negative consequences. For example, federal systems may encourage regional inequalities in the distribution of the costs and benefits of public policies; allow for the oppression of local minorities by local majorities within the subnational jurisdictions; and produce an undesirable 'race to the bottom,' in which regional governments attempt to attract economic enterprises by lowering regulatory standards and reducing governmental taxation and expenditure (Shapiro 1995).

3. *The Empirical Study of Federalism*

The empirical branch of the study of federalism has focused on the description and measurement of federal systems, and also on the establishment of relevant causal relationships.

3.1 *Descriptive Approaches*

In the descriptive realm, scholars have been concerned above all with identifying the core characteristics of federalism, characterizing variation in federal systems, and describing the mechanisms through which federal systems operate. At the heart of the federal idea is the division of powers and responsibilities between an overarching central government exercising authority over a broad territory, on the one hand, and a number of constituent governments each exercising authority over a particular subset of that territory, on the other. This division of power is generally enshrined in a constitution that guarantees the existence of all of the governments and provides at least a rough outline of the division of power among them.

Within these broad parameters, scholars have offered a range of more specific definitions of federalism, and no universal agreement about the term's precise definition has been reached. Different definitions produce somewhat differing classifications of political systems into federal and nonfederal categories. The result is that the term 'federal' can be applied to a variety of political systems in which power is shared between general and constituent governments. The political institutions of ancient Israel and the Greek city-states displayed federal characteristics. The US federal system, established in 1789, has served as the prototype of modern federalism (Elazar 1968). At the turn of the twenty-first century, significant federal elements existed in political systems across the globe, including Argentina, Australia, Austria, Belgium, Brazil, Canada, the European Union, Germany, India, Malaysia, Mexico, Nigeria, Pakistan, the Russian Federation, South Africa, Spain, Switzerland, the USA, and Venezuela (Elazar 1994).

Among federal systems, there is variation in the degree and character of the distribution of power between national and regional governments. This variation can be observed both in formal constitutional arrangements and in the working relationships among different levels of government within a political system. One component of the variation between federal systems, and within individual systems over time, is the extent to which powers and responsibilities rest with the central government or with the regional governments. Federal systems also vary in the extent to which these powers are either strictly divided between levels of government or shared between those levels of government.

With respect to the first of these dimensions, scholars have not yet reached a consensus about a

precise way to measure the relative centralization and decentralization of federal systems (Watts 1998). The formal distribution of powers set forth in constitutions may provide some insight into the comparative degree of centralization between systems. Such a constitutional focus has limitations, however, and is unlikely to be as useful in measuring changes over time within a federal system. Here, close attention must be paid to legislative and judicial decisions altering the balance of power between the central government and the constituent governments. One summary measure that is commonly examined is the proportion of governmental expenditures undertaken by the central government, either in total or in particular policy areas (Peterson 1995). Efforts to characterize variation in the distribution of powers between general and constituent governments in federal systems have also led to the development of a wide range of metaphors and conceptual models (e.g., 'dual federalism,' 'co-operative federalism'). These metaphors and models attempt to characterize either the degree of centralization in a federal system or the extent to which powers are strictly divided or cooperatively shared between levels of government, or both. In all, Stewart (1982) found more than 300 such models and metaphors that had been proposed in the literature.

Descriptive studies have also examined, both statically and dynamically, the mechanisms through which federal systems operate. Core areas of concern here include the constitutional, legislative, and judicial processes through which powers are allocated between levels of government, and the fiscal (e.g., grants-in-aid), regulatory, and political instruments through which governments within a federal system influence each other's actions (Wright 1988).

3.2 *Causal Approaches*

The causal branch of the empirical study of federalism addresses two major areas of concern: the causes and consequences of the establishment and dissolution of federal systems, and the causes and consequences of changes in how powers are divided between governments within federal systems.

Federal systems can be established either by joining together previously independent governments, or by decentralizing power in a previously unitary state. The formation of federal systems involves somewhat different processes and conditions in each of these two cases. In all cases, however, federal systems emerge under political conditions that contain strong pressures for both political unity (in the form of a national government) and political autonomy (in the form of regional governments with meaningful powers).

The pressures for unity most commonly result from a desire for military or economic advantage that constituent governments feel cannot be achieved independently. When federal systems are formed from previously independent governments, the reluctance

of these governments to cede too much power to a newly formed central government creates pressure for regional autonomy. More generally, pressures for autonomy most commonly involve protecting constituent governments' diverse interests, whether based on nationality, language, ethnicity, religion, economy, or other variables. Against the backdrop of these general dynamics, scholars have debated what conditions are necessary and sufficient to produce a federal system, with particular discussion of cultural, economic, political, social, and historical variables (Riker 1975).

Scholars have sought to understand not only why federal systems are formed, but also what the consequences of federal systems are. For example, federal systems can have important consequences for the management of ethnic conflict, either mitigating or exacerbating ethnic tensions, depending on the particular institutional arrangements in place.

The late twentieth century witnessed the disintegration of some political systems that were federal in principle or practice, including the Soviet Union, Yugoslavia, and Czechoslovakia. The survival of federal systems is imperiled when institutional arrangements reinforce rather than crosscut existing lines of division within the system, or when one group or constituent government representing a supermajority dominates the overall system (Watts 1998, Elazar 1968).

At a less macroscopic level, the causal study of federalism focuses to a significant extent on the evolution of federal systems: why the distribution of powers changes in federal systems, and with what consequences. There are numerous studies of the proximate impact of legislative, judicial, and executive decisions on the distribution of powers in federal systems. Other analysts have focused on the role of underlying forces such as changes in technology, culture, ideology, society, and economy (Anton 1984).

In all, there is significant dissatisfaction among scholars with the current state of our ability to construct robust theories of why federal systems evolve in the ways that they do (Beam et al. 1983). Explanations focusing on proximate causes have been criticized for being atheoretical and for too often relying on single case studies. Analyses focusing on underlying causes have not produced a clear consensus (Anton 1984). Understanding the causes of centralization and decentralization in federal systems will remain an important area of investigation in the decades ahead. Scholars may increasingly attempt to link explanations focusing on proximate and underlying causes, constructing theories that specify how underlying forces such as culture or economy might in turn produce centralizing or decentralizing actions by proximate institutions such as legislatures or courts.

Scholars have also discussed the consequences of centralization and decentralization within federal systems. For example, there is evidence that decentral-

ized systems make decentralized political party structures more likely (Riker 1975). In another realm, some scholars have argued that decentralization of redistributive and regulatory functions may lead to a competitive spiral in which subnational governments reduce their interventions in these areas (Peterson 1995).

4. Future Directions

Several areas of the study of federalism will likely be of particular importance in the first decades of the twenty-first century, as the literature continues to respond to important political developments. Scholars will be particularly interested in: the interaction between economic globalization and federal structures (see *Globalization: Political Aspects*); the emergence and evolution of supranational federal arrangements, including the European Union (see *European Integration; Regional Integration*); and the relationship between federal structures and the management of conflict in multiethnic political systems.

See also: Electoral Systems; Globalization: Political Aspects; Political Representation; Political Thought, History of; Regional Government; Representation: History of the Problem

Bibliography

- Anton T J 1984 Intergovernmental change in the United States: An assessment of the literature. In: Miller T (ed.) *Public Sector Performance: A Conceptual Turning Point*. Johns Hopkins University, Baltimore, MD, pp. 15–64
- Beam D R, Conlan T J, Walker D B 1983 Federalism: The challenge of conflicting theories and contemporary practice. In: Finifter A (ed.) *Political Science: The State of the Discipline*. American Political Science Association, Washington, DC, pp. 247–79
- Elazar D J 1968 Federalism. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York, Vol. 5, pp. 353–67
- Elazar D J (ed.) 1994 *Federal Systems of the World*, 2nd edn. Longman, Harlow, UK
- Peterson P E 1995 *The Price of Federalism*. The Twentieth Century Fund, New York
- Riker W H 1975 Federalism. In: Polsby N W, Greenstein F I (eds.) *Handbook of Political Science*. Addison-Wesley, Reading, MA, Vol. 5, pp. 93–172
- Shapiro D L 1995 *Federalism: A Dialogue*. Northwestern University, Evanston, IL
- Stewart W H 1982 Metaphors, models, and the development of federal theory. *Publius: The Journal of Federalism* 12: 5–24
- Watts R L 1998 Federalism, federal political systems, and federations. *Annual Review of Political Science* 1: 117–37
- Wright D S 1988 *Understanding Intergovernmental Relations*, 3rd edn. Brooks/Cole, Pacific Grove, CA

A. Burris

Female Excision: Cultural Concerns

1. Overview

Female excision, also known as female circumcision, is a euphemism for a variety of genital surgical procedures. It is a time-honored cultural tradition that has recently come under intense international scrutiny. Practices of genital alteration have existed in recent times in Australia, Asia, Latin America, America, and Europe. In fact, some Western surgeons claimed to have invented similar procedures as treatment for sexual dysfunction. Presently, excision is most common in Africa. Thus, this article focuses on African experiences. Given the discrepancies in existing studies, it is difficult to obtain accurate information about the exact prevalence of the practice. However, it is roughly estimated that about 80 to 110 million African women and children are affected.

There are at least four major forms of excision. In the mildest form, the clitoris is barely nicked or pricked to shed a few drops of blood. The second kind which is often called *sunna* involves the removal of the clitoral prepuce, hood, or outer skin. The third gradation, clitoridectomy, more accurately describes the process of excision. Here, the clitoral glans and some of the *nympha* or *labia minora*—the narrow lip-like enclosures of the genital anatomy, are severed. Infibulation is the least popular and most extreme procedure; it entails scraping the *labia majora*—the two rounded folds of tissue that contour the external boundaries of the vulva, and stitching the remaining raw edges together in a manner that reduces the vulva opening after the surgery heals. Different forms of the operation affect the genitalia differently.

2. Rationales

Nobody can unequivocally identify where, when, or why excision originated. The practice can mean different things to different actors in different milieus. Embedded in an intricate web of habits, attitudes, and values, excision has both functional and symbolic connotations. It is validated and undergirded by a spectrum of temporal and spiritual beliefs. While some of the justifications are at odds with ample medical knowledge, efforts to counter the practice by illuminating the medical contradictions have not been very successful in eroding commitment to it. This is mainly because, for adherents, the attraction and imperative of the practice has little to do with medical science. Recurring themes such as sexuality and fertility signify and affirm paramount indigenous values like solidarity, public recognition of lifecycle change, and procreation for social continuity. That it deals with the genitalia and is suffused with gender constructs does not denote unequivocally exclusive or primary concern with gender and sexuality.

Among the Sande for example, it has been reported that initiates spend months in seclusion being instructed on topics as broad as the cultural secrets of successful living and traditional medicine. With the Chagga, excision is said to test courage by simulating birth labor pains. To the Mende and Sherbro, it fosters cooperation among women. The Bambara and Mossi assume it reduces mortality and morbidity. Some Bantu and Sudanese use it for purification. In other areas, excision is employed as a means of self-definition, self-assertion, and expression of personhood and identity; it has also been characterized as the equivalent of an ornamental tattoo, or a kind of cosmetic surgery akin to ear-piercing. On a continuum with these worldviews are a competing set of beliefs which tend to emphasize the dimensions of sexuality and gender. But even these beliefs are tempered and nuanced in the manner in which they inform everyday realities.

The Mandingo, Massai, and Swahili believe that excision renders the body fertile and fit for the social order. Among some Yoruba it is considered a contraceptive device. In some other groups, excision reflects a fear of a primal hermaphroditic human nature and an attempt to differentiate the sexes. Still others justify it on grounds of aesthetics; to these, unaltered female genitalia are unsightly, and occluded or attenuated body orifices are more preferable. Some communities subscribe to the belief that a girl is ripe for sexual relations only after being ushered into womanhood by circumcision. Finally, there are communities that, to enforce moral values and ethical standards, discourage premarital sex. These groups believe that circumcision attenuates sexual desire and this mindset underscores its incorporation as a rite of passage. Regardless of the prevalence of this mindset in select groups, excision does not necessarily encode absolute prenuptial chastity nor does it *prima facie* signal a distortion of women's sexual experiences. From all indications, the practice is not homogenous; it has divergent contexts, meanings, and consequences. However, some opponents of the practice insistently explain it monolithically to privilege connotations of sexual politics and patriarchal control.

3. Emerging Trends

For various reasons, many communities where excision was once an aspect of elaborate rites of passage celebrations with minimum age requirements now excise children who are too young to comprehend the ritual implications of the experience. A corollary of this development is that some persons undergo the operation independent of accompanying ceremonies. In contradistinction to the evolving tendency to cut the genitalia without performing complementary ritual ceremonies, a few communities are embracing alternative rites that replace physical cutting with

symbolic enactments. Here initiation modules only entail health education, transmission of traditional knowledge, and festivity.

Patterns of change are also manifesting in more decidedly oppositional guises. Often, fierce opposition is a catalyst for other forms of transformation. Accordingly, it is not unusual to find the circumcision of the under-age and/or unceremonious cutting as forms of resistance to intense expressions of anti-circumcision impulses. These conflicts mutually reinforce each other.

An abiding bone of contention in the controversy about excision is the question of its health effects. Increasingly, the veracity of extant claims about health hazards and the premium that is placed on it as an arsenal for the anti-circumcision campaign, are being attacked. Searching analyses of available research suggest that advocates of abolition are quicker abstractly to chronicle an array of hazards than they are to produce objective evidence to back up their claims about the severity and frequency of health complications. Where these claims cannot be substantiated empirically, they operate as mere propaganda that remain unrepresentative of the lived experiences of circumcised persons. Although existing data are deficient, there is sufficient evidence to demonstrate that the health complications that are attributable to excision vary with the extent and circumstances of the surgery. The health hazards of the operation typically are compounded by the crude circumstances under which it usually occurs, including the lack of anesthesia and the use of unsterile instruments.

Drastic forms of excision such as clitoridectomy are associated with an increased risk of bleeding and infection; the complications of infibulation include serious bleeding, infections, urinary problems, infertility, labor and delivery difficulties, and adhesions and obstructions. There is no objective evidence that directly links excision with mortality. The World Health Organization categorizes a range of possible negative health consequences as follows: short-term complications which include pain, injury to adjacent tissues, potentially fatal hemorrhage and shock, urinary retention, and acute and chronic infections; long-term complications that range from difficulty in passing urine, urinary tract infections, pelvic infections, infertility, keloid scars, abscesses and cysts, to menstrual difficulties, dyspareunia and sexual dysfunctions, and problems in pregnancy and childbirth as well as sexual, mental, and social consequences.

This list is more concise and less exaggerated than the inventories of deleterious sequelae in other literature; nevertheless, it is subject to caution as it derives from cross-sectional studies that are not always able to establish direct causation between the operation and the putative complication. This list is also comparable to other studies in the sense that it offers little guidance for inferring the rates of complications. Recent systematic reviews of relevant materials show that

consistent and conclusive evidence of far-reaching complications are rare. These observations call for the qualification of the claims at the heart of the anti-circumcision campaign strategy. Notwithstanding the observations, the anti-circumcision alliance has done a remarkable job of mobilizing international consensus against the practice by emphasizing its harmful effects.

3.1 Eradication

Systematic efforts to abolish excision date back to the colonial era. In 1946, the British colonial administration in Sudan amended its Penal Code to make illegal all forms of genital surgeries, except *sunna*. In Kenya, interference by missionaries condemned the practice to secrecy; as a clandestine custom, it was particularly difficult to assess the incidence of casualties. In material respects, the intervention also politicized the significance of the operation as some began to appropriate it as a symbol of nationalism. According to Kenyatta (1938), the controversy came to a head in 1929–30 when families who interpreted the colonist measures as an affront to their freedom, social customs, and cultural beliefs hastened to circumcise their daughters, even at the hands of unskilled persons with inevitable tragic consequences.

The backlash recurred in 1989 when the incumbent President of Kenya, Arap Moi, denounced the custom and scores of women and girls volunteered for the procedure in protest of the official sanction. More recently, some grassroots communities are reinforcing, instead of resisting heightening global pressure against circumcision. To this end, several villages have adopted renunciation oaths and declarations that they complement with active learning and socioeconomic empowerment programs. Some eradication strategies train former excisors in grassroots management skills and facilitate their involvement in business enterprises. In lieu of material benefits, certain competing models—especially those that reconfigure the rite to eliminate the ritual of cutting—incorporate creative ways of enlisting the support of traditional excisors.

Several sovereign powers have prescribed an assortment of steps to stem excision. These range from economic sanctions and extradition to prosecution and punishment. In 1991, France became the first country to establish that the threat of ‘genital mutilation’ is sufficient grounds to grant women refugee status. Subsequently, Canada pioneered guidelines expanding the basis of refugee claims to include gender-related persecution which was construed to encompass circumcision. More recently in the United States, a few African women have succeeded in obtaining political asylum or staying deportation by conceptualizing circumcision as persecution.

A series of human rights instruments and provisions can also be read as prohibiting excision. These range

from regulations concerning equal protection and nondiscrimination to those pertaining to the rights of children, the right of sexual and corporal integrity, the right against torture, the right to life, and the right to health. The World Health Organization officially condemned surgical alterations of female genitalia in 1982. The World Medical Association, the International Federation of Gynaecology and Obstetrics, and the Population Fund issued warnings about the risks of the practice and oppose the participation of health professionals in it. In 1990, the Committee responsible for the implementation of the Convention for the Elimination of Discrimination against Women recommended the eradication of the practice.

Although a few African countries have instituted various initiatives to address excision, it appears that the force of culture and the poverty of resources inhibit meaningful implementation of these initiatives. A wide range of official declarations against excision have been made in Benin, Nigeria, Ghana, Burkina Faso, and Senegal. Some European countries have also promulgated relevant laws and policies. Several interventions promote prevention through information, education, and communication campaigns. However, the predominant trend proscribes the practice as a crime that is subject to prosecution. As of the summer of 1994, France was the only country in which the practice of circumcision has actually resulted in criminal trials. In 1982, the Royal College of Obstetricians and Gynecologists found that excision was a very infrequent occurrence in Britain. By 1985 however, a scare of waves of occurrences prompted the promulgation of the British Prohibition of Circumcision Act. Section 1 of the Act makes it an 'offense for any person (a) to excise, infibulate or otherwise mutilate the whole or any part of the labia majora or labia minora or clitoris of another person; or (b) to aid, abet, counsel or procure the performance by another person of any of those acts on that other person's own body.' Section 2 of the Act indicates that genital surgeries are not absolutely prohibited; its precise wording seems to accommodate exceptions under certain circumstances. This section provides that 'Subsection 1(a) of Section 1 shall not render unlawful the performance of a surgical operation if that operation (a) is necessary for the physical or mental health of the person on whom it is performed and is performed by a registered medical practitioner.' This provision allows the continuation of 'trimming' surgeries that typically are performed on women who are dissatisfied and psychologically disturbed by the shape or size of their external genitalia. Plausibly, the mental health exemption can be construed to support psychologically-grounded claims for circumcision. However, section 2 (2) of the Act precludes the consideration of psychological effects induced by any belief 'that the operation is required as a matter of custom or ritual.'

In 1996, the US Congress made the excision of a

person who has not attained the age of 18 a felony punishable by up to five years imprisonment. This statute equally exempts a surgical procedure performed by a licensed medical practitioner, which is either necessary for the health of the woman on whom it is performed, or is performed prenatally or postnatally on a woman. The US Congress further authorized the imposition of economic sanctions against any countries that have a cultural custom of female circumcision and have not implemented educational programs to counteract the practice. Resorting to this course of action as an antidote to circumcision is not without limitations. Withholding or conditioning aid as a mechanism for controlling the practice subordinates humanitarian considerations to political expedience. While there obviously are no simple solutions, it is important to recognize that when deeply embedded cultural attitudes are reinforced by complex economic, social, and political variables, a more nuanced strategy may be more effective than essentially top-down or carrot-and-stick approaches.

3.2 Medicalization

Empirical evidence of the range and frequency of adverse health outcomes posit a strong correlation between the probability of harm and the degree of cutting, the nonhygienic and anesthetic conditions under which it is performed, and the limited skills of the excisor. Drawing on these findings, some authorities recommend a spectrum of medical interventions as a way of containing potential harm. In one initiative, nurses dispense prophylactic antibiotics, anti-tetanus injections, and sterile razors to girls who are later cut by traditional circumcisors. In another model, traditional birth attendants are trained in septic and precautionary procedures for genital cutting. In urban parts of Sudan and in Egypt, the governments have attempted to curtail the adverse health consequences that can be traced to excision by clinicalizing mild forms of the procedure. In the Netherlands, a Welfare, Health and Cultural Ministry report and officials at the Harborview Medical Center in Seattle in the US recommended a similar course of action.

Apparently some African immigrants in Holland circumcised themselves under circumstances that compound the potential harms of the procedure. Their dilemma arose partly because, while they could not afford to be alienated for failure to comply with the cultural mandate for circumcision, the Dutch medical and social workers denied them assistance in executing the operation. To address the needs of these women, the Ministry, on the basis of the findings of a study it sponsored among 500 Somali refugee women, recommended a distinction between mutilating or tissue-impairing circumcision and nonmutilating ritual incisions. The report further argued for official stipulation of conditions under which doctors might

be allowed to perform simple incisions or ritual perforations of the clitoral covering as an alternative to infibulation. The report drew a storm of protests and eventually was superseded in 1992.

In 1996, a similar controversy erupted in Seattle, Washington, when in the interest of some Somalian immigrants, pediatricians at Harborview Medical Center took under advisement the possibility of performing 'symbolic blood-letting' procedures on girls. The proposed procedure would only involve a negligible nicking of the prepuce far less invasive than what is allowed in the circumcision of infant males. Beseiged by outraged opponents, Harborview decided to abandon the prospective accommodation.

The Dutch and Seattle experiences illustrate the extent to which sworn opponents of the practice favor prohibition and are vehemently opposed to medicalizing any variant of the practice—even as an interim measure to arrest potentially dire consequences.

One reason proffered for this position is that the surgical compromise ultimately signals complicity in patriarchal control and diminution of the right of women to sexual and bodily integrity. Some ridicule operations performed under medical conditions and/or supervision as an empty caricature devoid of traditional connotations. Others who envisage a scenario where traditional midwives are displaced by forum change seek to enhance their fundamental opposition to the practice by coupling it with an ancillary aversion to the implicit loss of income for the midwives. Another concern is that medicalization may unduly burden and deplete public funds which could be deployed for less controversial purposes. It is also contended that the compromise creates a slippery slope that can be abused to camouflage severe forms of the operation.

Whatever the validity of these concerns, it is worth ruminating over the merits of changing the forum and form of the practice. The options may be problematic, but as already suggested, there are no exact solutions. Criminalization, which is overwhelmingly championed by the anticircumcision coalition, is not a panacea. Indeed, where not being excized is perceived as a transgression of social mores, attempting an undifferentiated assault may jeopardize the prospects for enduring reform. The rhetoric of the opposition is premised largely on claims about the health hazards of excision. However, a recent examination of how much available research supports widely publicized allegations about the prevalence and harmful effects of female genital surgeries found that most of what are purported to be universal 'facts' vis-à-vis the deleterious consequences are drawn from studies of the Sudan where most women are infibulated. The review confirmed that obtaining irrefutable epidemiological data is limited by logistical and ethical constraints involved in monitoring over time the experience of individuals who undergo the surgeries.

The fact that adverse complications are exacerbated

by the poor skills, anatomical knowledge, sanitary conditions, and surgical equipment that characterize excision in traditional settings reinforces arguments for medicalization. By the same token, instituting a change of form such as symbolic pricking is inconsequential from a health standpoint. Without a probable risk of harm, it is difficult to make a compelling case against the accommodation of consensual genital marking. In this light, medicalization reconciles the need to protect women with a respect for embodied sociocultural identities. Some would even argue that the virtue of tolerance and the principle of self-determination allow accommodation of more extensive genital alterations, as long as the procedure is demonstrably benign. Given the current wave of anticircumcision sentiments, this view is subject to dispute and far from endorsement. Nevertheless, it raises issues that deserve serious attention.

3.3 Education

Several authorities hail education as the single most important factor in combating the practice. For one, law alone seldom changes behavior. Effective legal engineering can hardly occur without favorable public disposition, which in turn can be stimulated by fostering meaningful interactions between legal and educational strategies. In this vein, it is noteworthy that gains achieved through the medium of education are often contingent and enhanced by concurrent poverty alleviation. The emphasis on education and connection with poverty, however, can only go so far. The substantial prevalence rate of the genital operations among very educated and privileged women in many ethnic groups suggests that advances in literacy and socioeconomic status do not necessarily decrease allegiance to circumcision. Undoubtedly, good education is an effective antidote for challenging erroneous beliefs. However, many women undergo circumcision on the bases of rationales that cannot be readily refuted by objective evidence. Ritual thought and action belong to specific symbolic frames of reference and are best understood within the totality of their natural context. Attempting to falsify indigenous explanations of circumcision by regarding them as assertions of facts that can be disconfirmed through formally logical discursive categories obscure valid psychosocial processes.

See also: Female Genital Mutilation; Gender, Feminism, and Sexuality in Archaeological Studies; Gender Ideology: Cross-cultural Aspects; Gynecological Health: Psychosocial Aspects; Health: Anthropological Aspects; Health in Developing Countries: Cultural Concerns; Human Rights, Anthropology of; Human Rights, History of; Ritual; Ritual and Symbolism, Archaeology of; Sexuality and Gender

Bibliography

- Association of African Women for Research and Development 1983 A statement on genital mutilation. In: Davies M (ed.) *Third World: Second Sex*. Zed Books, London
- Boddy J P 1989 *Wombs and Alien Spirits: Women, Men and the Zar Cult in Northern Sudan*. University of Wisconsin Press, Madison, WI
- Case Western Law Review 1996 Colloquium: Bridging society, culture, and law: The issue of female circumcision. *Case Western Law Review* 47: 263
- Coleman D L 1998 The Seattle compromise: Multicultural sensitivity and Americanization. *Duke Law Journal* 47: 717
- Daly M 1978 African genital mutilation: The unspeakable atrocities. In: *Gyn/ecology: The Metaethics of Radical Feminism*. Beacon Press, Boston
- Dorkenoo E, Elworthy S 1992 *Female Genital Mutilation: Proposals for Change*. Minority Rights Group, London
- Giorgis B W 1981 *Female Circumcision in Africa*. United Nations Economic Commission for Africa, Addis Ababa
- Kenyatta J 1938 *Facing Mount Kenya; The Tribal Life of Gikuyu*. Secker and Warburg, London
- Lightfoot-Klein H 1989 *Prisoners of Ritual: An Odyssey into Female Genital Circumcision in Africa*. Harrington Park Press, New York
- Lionnet F 1992 Women's rights, bodies and identities: The limits of universalism and the legal debate around excision in France. *Inscriptions* 6: 98
- Obermeyer C M 1999 Female genital surgeries: The known, the unknown, and the unknowable. *Medical Anthropology Quarterly* 13(1): 79–106
- Saadawi N 1980 *The Hidden Faces of Eve: Women in the Arab World*.
- Shell-Duncan B, Herlund Y in press *Female Circumcision in Africa: Culture, Change and Controversy*. Lynne Rienner, Boulder, CO
- Shweder R in press What about FGM? And why understanding culture matters in the first place. In: Shweder R, Markus H, Minow M (eds.) *The Free Exercise of Culture: How Free Is It? How Free Ought It To Be?* Russel Sage, New York
- Toubia N 1994 Female circumcision as a public health issue. *New England Journal of Medicine* 331: 712–16
- Van Gennep A 1960 *The Rites of Passage* (trans. Vizedom M, Caffee G). University of Chicago Press, Chicago
- Walker A 1992 *Possessing the Secret of Joy*. Vintage Books, London
- Wilson B R 1970 *Rationality*. Blackwell, Oxford
- World Health Organization 1982 *Seminar on Traditional Practices Affecting the Health of Women and Children*.
- World Health Organization 1996 *Female Genital Mutilation*. WHO, Geneva

L. A. Obiora

Female Genital Mutilation

Female genital mutilation is a collective term for various surgeries performed on young girls, often done as part of initiation rituals into adulthood. Other terms commonly used are female genital cutting,

female circumcision, clitoridectomy, infibulation or Pharaonic circumcision, excision, and *sunna*. Except for female genital cutting and female circumcision which also are general terms and used interchangeably with female genital mutilation, the other terms are more specific, indicating the particular type of operations being performed.

1. Extent and Distribution

Sources estimate that perhaps as many as 130 million women in the world today have been mutilated genitally (Toubia 1995). Since relatively little extensive statistical research has been conducted on the topic, prevailing figures quoted tend to vary. Figures usually are derived from approximate population censuses from areas where mutilation is known to be practiced.

As a cultural practice, female genital surgeries are particularly widespread on the African continent and to some extent on the Saudi Arabian peninsula, but are only sporadically documented from the rest of the world. Instances of genital mutilation are reported from selected communities in the Middle East and from Muslim Malaysia. In Europe and the USA, clitoridectomy was until the 1930s medically administered in diagnosis of excessive masturbation, nymphomania, and hysteria (cf., e.g., Roheim 1932, Bonaparte 1953). Cosmetic surgery of enlarged clitorises and labia minora has been reported recently both from the USA and Europe (cf. Walley 1997). In Europe and the USA, female genital mutilation was primarily a single instance phenomenon and not a social convention. During the last few decades of the twentieth century, however, large-scale immigration to Europe, the USA, and elsewhere from African countries where genital operations are mandatory, has extended the geographical location of female genital mutilation (cf. Lionnet 1992).

In Africa, the practice of genital surgeries is contained to certain regions only, most notably to the northeastern and eastern parts of the continent, from Egypt in the north to Tanzania in the south and across the continent along the Sahelian countries from Ethiopia to Senegal. It is practiced in some 28 countries and by a wide variety and number of ethnic groups. The extent and type of surgical intervention as well as its cultural meaning and social context vary considerably making it difficult to generalize across the societies in which female genital surgeries are the norm. Female genital mutilation occurs among Muslims, Christians (particularly among Copts), as well as adherents of traditional systems of belief, among agricultural and nomadic peoples alike and in rural as well as urban communities. The highest regional concentration, that is, where a majority of women in a particular area are mutilated, is in Northeast Africa; in the Horn (Somalia, Djibouti, and Eritrea), across Ethiopia, and into northern Sudan

and Egypt. In Somalia and northern Sudan female genital mutilation is almost universal (Gruenbaum 1996). Elsewhere in Africa, genital surgeries as a cultural practice are scattered and alternate with ethnic affiliation, religious faith, economic class, and level of education. Within the same community, practices may even vary between subgroups and individual families, causing a situation where mutilated women live side by side with nonmutilated females.

As a general rule, where female genital mutilation is performed, male circumcision is also the norm. People themselves often draw a symbolic and social parallel between boys' and girls' circumcision. This is, among other things, reflected in a common term for the interventions in the local vernacular. While a number of societies in Africa and beyond practice male circumcision without cutting their women, the opposite, however, is never the case.

2. Operation Practices

Mutilation practices vary greatly. In some areas, such as Amharic Ethiopia, girls are operated upon within 40 days after birth, while in neighboring Somalia the operation usually is performed when the girls are between 6 and 8 years old (Hicks 1993, Talle 1993). In Kenya and Tanzania the girls are cut routinely just before puberty and often in association with elaborate initiation ceremonies (cf. Murray 1974). In a few, relatively exceptional cases reported among others from West Africa, the surgery is performed after puberty. It also has been reported that married women who have defied traditions and remained unmutilated, in cases of barrenness have undergone the surgery after being married in order to promote fertility. Not only the girl's age, but also the surgical intervention, the performance context, and ritual elaboration exhibit substantial variations. In places in Egypt and among devoted Muslims in towns in the Sudan, for instance, only the prepuce or the tip of the clitoris is excised. This operation is referred to as *sunna* ('duty' in Arabic, i.e., to follow tradition). Surgically, *sunna* is the mildest form of mutilation and physiologically comparable to the removal of the foreskin at male circumcision. Although encouraged by religious leaders, health personnel, and opinion makers as a commendable mutilation form, it is still uncommon. What people label *sunna*, however, in practice is often really a clitoridectomy.

The most widespread form of female genital mutilation in fact is clitoridectomy, an intervention that implies removal of the whole or part of the clitoris, sometimes together with the *labia minora*. In the latter case the operation is also referred to as excision. Various forms of clitoridectomy-cum-excision are practiced from Senegal and Mauritania in West Africa, across Sahel to Egypt in the northeast, and southwards to Tanzania in East Africa. It is carried

out across the whole area where female genital surgeries occur.

The surgically most severe form of operation is called infibulation (from Latin *fibulae*, clip) or Pharaonic circumcision. This is a form of surgery where the *labia minora* and parts of *labia majora* are excised wholly or partly, and often the clitoris; the sides of the vulva are then stitched together. A tiny orifice is left open serving as an outlet for urine and menstrual blood. The surgery transforms the genitals of a woman into a flattened, smoothed, and closed vulva with an oblong scar and a miniscule opening at the lower part. The practice of infibulation is particularly prevalent in the Horn of Africa, in the Sudan (Northern), and in southern (Nubian) parts of Egypt. It also occurs widely in Mali, in northern Nigeria, and among ethnic Somalis in other East African countries. The infibulated women are opened (defibulated) at marriage, either surgically—often by the practitioner who did the closing—or by the husband at penetration.

In medical and anthropological literature it is common to differentiate between four types of genital surgeries: *sunna*, clitoridectomy, excision, and infibulation. The types are classified by a gradient of surgical severity. Variations and modifications of the different types may occur over time as well in individual cases within the same geographical location. In the Sudan, for instance, an intermediate type, a compromise between excision and infibulation, sprung up during the colonial area as a local response to legislation against the practice imposed from outside (El Dareer 1982). The skills and techniques of the practitioners as well as the preferences of individual families may contribute to variation and differences in mutilation styles.

Genital operations usually are performed in the home of the girl and, as a rule, undertaken by an elderly woman skilled in the task. Often these women are local birth attendants as well. In Somalia, the practitioners belonged traditionally to a ritually ambiguous and despised subclan within the community. In other places, and these days even in Somalia, any adult woman with the necessary courage and skills may engage in the practice.

The operation normally is performed without any extensive form of anaesthetic or antiseptic precautions. Different herbs, animal fat, dung, or other substances believed to have homeostatic and cicatrizing properties are applied to the wound after the operation to prevent excessive bleeding and infection. Depending upon the routine, skills, and preference of the practitioner, the instruments used for the surgery are either a small knife prepared for that kind of work, or a razorblade. In the case of infibulation, the severed parts are stitched with thorns of the acacia species, or recently, with catgut. In urban areas, medically trained personnel or paramedics commonly perform 'circumcision' on both boys and girls. In addition to giving prophylactic antibiotics and tetanus injections, the

urban practitioners normally apply local anesthetics when operating. Many parents prefer these modern operations as they tend to reduce the pain and minimize adverse effects, such as hemorrhaging or infection from the operation. With the medicalization of the genital surgeries, men have also become involved in operating on girls. Traditionally, men operated only on boys, while women operated on girls. The same-sex pair of circumcizer and circumcized reflected the thinking of male circumcision as an analogy of female circumcision.

Among peoples in Kenya and Tanzania and in large parts of West Africa, the surgical intervention is associated with elaborate rituals of initiation into adulthood (e.g., Murray 1974). In northeastern Africa, where infibulation is widespread, the operation, however, is not normally followed by any extensive ceremonial celebration. The operated girl is usually given a few presents and the family mark the day with some extra food and drinks, but the event is not a ceremony of great societal worth. In cases of physical cutting being part of some more significant rites of passage into adulthood, the waning of such traditions has not necessarily led to a decrease in the surgery itself. Instead of being part of a larger event, genital mutilation is now often a single act performed on girls at almost any age.

3. *Historical Roots*

The existence of female genital mutilation dates back to pre-Islamic and pre-Christian time. Although the practice as such does not stem from religious affiliation, many people make an association between circumcision and religious faith. This is particularly the case in West Africa where genital surgeries both in men and women are more common in Muslim than Christian communities (Koso-Thomas 1987).

The term Pharaonic circumcision originates from the folk legend that the pharaohs of ancient Egypt performed similar operations on their women. Whether the pharaohs actually infibulated their women or not will probably never be substantiated, but the term nonetheless indicates that we are dealing with a cultural phenomenon of great antiquity.

Greek merchants and travelers who crossed the Mediterranean in classical time mention that female genital mutilation occurred among Egyptians as well as among the peoples along the Red Sea. The sparse descriptions emanating from these travelers, however, do not allow us to ascertain whether it was infibulation or clitoridectomy, or both, people practiced in earlier times.

European travelers visiting Egypt much later reported on the occurrence of clitoridectomy (Browne 1799, Seligmann 1913). However, female slaves from the south (perhaps Nubians from the Sudan or southern Egypt), whom the wealthy Egyptians favored

as mistresses, were infibulated (Widstrand 1964). Some authors held that these girls were mutilated upon arriving in the north by the slave merchants or by their masters to avoid loose living or unwanted pregnancies. This explanation is disputed as early reports suggest that infibulation occurred for some time among the peoples of the upper Nile in northern Sudan. Burckhardt who visited these areas early in the nineteenth century claimed that infibulated girls ('virgins') received a higher price on the Egyptian slave market than noninfibulated ones (Burckhardt 1819).

The beginning of female genital mutilation remains obscure. It has not been possible to determine whether mutilation originated in one area and from there spread to other areas or whether it originated independently in various places. With reference to the African continent, however, it is not unlikely that the practice arose in a culturally, commercially, central area and from there reached and influenced other parts of the continent. This assumption is based upon the fact that the practice of female genital mutilation appears to have had varying historical origins depending when people came into contact with each other (cf., e.g., Murray 1974). Historical and linguistic sources point to the Chusitic and Semitic-speaking peoples living along the Red Sea Coast as a core area of female genital mutilation (Seligmann 1913, Hicks 1993).

Recently, the practice of female genital mutilation is reported to have spread from one ethnic group to another in countries such as Chad, the Sudan, and Tanzania (cf. Leonard 2000, Gruenbaum 1996). In the Sudan, lower hierarchy groups often assimilate the practice of mutilation in order to be able to marry those of higher rank. Young girls in Chad are reported to perform clitoridectomy as a mark of modernity; since around 1980 the operation has become a kind of fashion in at least one reported area (Leonard 2000). Studies from places in West Africa suggest that the expansion of Islam is conducive to the spread of female genital mutilation. Among the ethnic Meru in Northern Tanzania, clitoridectomy has been revitalized as a sign of cultural pride and resistance against rapid modernization. These sketchy, contemporary examples show that the meaning of female genital surgeries may assume various values following shifts in social reality. To know its origin does not explain its present existence (Gruenbaum 1996). One distinguishable trait in the spread of female genital mutilation, however, appears to be the copying of cultural practices of higher status groups by people of lower rank (cf. Mackie 1995).

4. *The Cultural Explanation*

The practice of female genital mutilation is embedded deeply in cultural and moral preferences, gender identity and person forming, perceptions of body

aesthetics, and ethnic marking. The unsuccessful attempts made by the British colonial administration during the 1930s and 1940s to abolish or modify the most severe types of female genital mutilation in the Sudan and in Kenya testify to the social importance of the practice (cf. Murray 1974, El Dareer 1982). Anthropologists and others often associate the tradition of female genital mutilation with a patriarchal social structure, practically and symbolically privileging female reproductive powers over female sexuality. (cf. Boddy 1982, Talle 1994).

People themselves give many explanations to defend the practice. For instance, the pastoral Maasai of Kenya and Tanzania claim that a girl can neither be married nor give birth to healthy children if she is not clitoridectomized. Should she deliver before she is operated upon, the child will die or be born disabled (Talle 1994). In the Sudan and Egypt, uncut female genitals are considered ugly and dirty. Men are reluctant to marry such girls because by not being 'purified' by genital excision, they are held to be immoral in the widest sense. For the Bantu-speaking Kikuyu in Kenya a girl must have her clitoris removed to be considered an adult and responsible person. And so on. People present a plethora of meanings and motives for upholding the practice. Although seldom elaborated on explicitly, the control of female sexuality is at stake. References to moral looseness, unmarriagability, and fertility failures speak strongly to the societal value of female chastity and of male control of female fertility and sexuality (cf., e.g., El Sadaawi 1980, Gordon 1991, Talle 1993).

In Somalia, where the practice of infibulation is particularly widespread, the operation is performed in two stages: the excision and the stitching. First, the clitoris is excised. This is considered an act of purification (*xalaalays*; from the Arabic *xalaal*, allowed by religion). The clitoris (*kintir*) is a part of the body which is associated with dirt and childhood. A girl who has not been operated on is referred to as *kintirleey* (childish, i.e., having clitoris) and teased by peers. To adult Somali women the word *kintirleey* has a very negative and somewhat embarrassing ring. Another Somali synonym for *kintir* is *awle* which means approximately 'with (or having) father.' This term suggests that the Somalis, as a number of other peoples performing genital surgeries, see the clitoris as a male part in the female body and thus held to be unclean. When excised, women become more feminine (Talle 1993).

Where infibulation is practiced, virginity is not held to be a natural condition in women but has to be forcibly implanted in girls while they are still young. The hymen of a woman, which attracts so much interest and attention in many other Muslim societies and evokes strong sanctions if broken before marriage, has little social or cultural significance among infibulation practicing people. To them a chaste girl means a 'sewn' girl. A tight infibulation is a sign of

distinction and moral excellence. Symbolically, the closing is more important than the excision as the cutting of the flesh is irrevocable while the closing of the vulva can, in principle, be executed over and over again. Although never as tight as the first time, later surgeries (after childbirth and at divorce) return the woman's body to its previous state by remaking her virginity (El Dareer 1987). Virginity under such circumstances is not an anatomical facticity but a social category (Hayes 1975).

The closing of her vulva is looked upon as a protection for the girl against indecency and immorality. A reason often given, for instance by Somalis, as to why originally they began to infibulate their girls was to prevent sexual assault of young girls while out herding alone. The infibulation, or the creation of the artificial virginity, is seen as a safeguard against unwanted pregnancies and unidentified paternity. Both in northern Sudan and in Somalia a nubile, 'virgin' daughter upholds and marks the moral standing of the patrilineal group (Hayes 1975, Kennedy 1980, Talle 1993). To leave a daughter or sister 'uncut' is held to be unnatural; the vision of such a woman is beyond comprehension, something which cannot even be expressed in words.

5. A Global Concern

During the feminist wave of the 1970s in Europe and the USA, female genital surgeries became a politicized issue. At the Copenhagen conference in 1980 during the United Nations Decade of Women (1975–85) the issue was brought to widespread international attention. Opponents of the practice asserted that female genital mutilation was a severe abuse of women's and children's rights and demanded legislation against it. The term 'female circumcision' was substituted by 'female genital mutilation' (Hosken 1982). This change of name from circumcision which is used for genital operations on men, was meant to draw attention to the severity of intervention, primarily in terms of health risks, on girls and women as compared with that of men (cf., e.g., Toubia 1995).

Legislation against female genital mutilation is on the agenda of international organizations such as Save the Children, Amnesty International, WHO, UNICEF, and other UN agencies. To prevent the spread of such surgeries into their countries, European governments are passing laws criminalizing all forms of genital mutilation (cf. Lionnet 1992). African countries, Egypt, the Sudan, and Kenya among others, have also passed legislation against female genital surgeries.

The global discourse on female genital mutilation has often been moralizing, racially prejudiced, and without sufficient understanding of the social and cultural context of the practice. Many African women have reacted to the paternalistic attitude that often

permeates the global debate. The term female genital mutilation, which has received widespread acclaim, but which holds the implicit assumption that parents and kin deliberately intend to harm children, is contested (cf. Walley 1997). A more neutral term proposed recently is 'female genital cutting' (Shell-Duncan and Hernlund 2000). African women—themselves frequently victimized by genital surgery—have, however, also been among the strongest opponents to the practice (e.g., Abdalla 1982). The latest voice in a string of personal testimonies of an excruciatingly painful intervention with lifelong suffering comes from an internationally celebrated Somali model, now a UN ambassador against female genital mutilation (Dirie and Miller 1998).

Although opposition is growing everywhere, women continue to favor the practice. They do it not necessarily because they like it, but because they see no other alternative. Africans in exile, exposed to a majority society where mutilation is neither a norm nor an ideal and, additionally, relieved of the social pressure to perform it, may spearhead the eventual eradication of female genital mutilation.

See also: Cultural Relativism, Anthropology of; Female Excision: Cultural Concerns; Gender and Feminist Studies in Anthropology; Gender Ideology: Cross-cultural Aspects; Group Decision Making, Social Psychology of; Islam and Gender; Male Dominance; Near Middle East/North African Studies: Gender; Reproductive Rights in Developing Nations; Ritual; Sexuality and Gender

Bibliography

- Abdalla R M D 1982 *Sisters in Affliction: Circumcision and Infibulation of Women in Africa*. Zed Press, London
- Boddy J 1982 Womb as oasis, the symbolic context of Pharaonic circumcision in rural northern Sudan. *American Ethnologist* 9(4): 682–98
- Bonepart M 1953 *Female Sexuality*. International Universities Press, New York
- Burckhardt J L 1819 *Travels in Nubia*. Darf Publishers, London
- Dirie W, Miller C 1998 *Desert Flower: The Extraordinary Journey of a Desert Nomad*. William Morrow, Location
- Dareer A 1982 *Women, Why Do You Weep? Circumcision and its Consequences*. Zed Press, London
- Sadaawi N 1980 *The Hidden Face of Eve: Women in the Arab World*. Zed Press, London
- Gordon D 1991 Female circumcision and genital operations in Egypt and the Sudan: A dilemma for medical anthropology. *Medical Anthropology Quarterly* 5(1): 3–14
- Gruenbaum E 1996 The cultural debate over female circumcision: The Sudanese are arguing this one out for themselves. *Medical Anthropology Quarterly* 10(4): 455–75
- Hansen H H 1972/73 Clitoridectomy: Female circumcision in Egypt. *FOLK* 14/15: 15–26
- Hayes R O 1975 Female genital mutilation, fertility control, women's roles, and the patrilineage in modern Sudan: A functional analysis. *American Ethnologist* 2(4): 627–37

- Hicks E K 1993 *Infibulation: Female Mutilation in Islamic Northeastern Africa*. Transaction Publishers, New Brunswick and London
- Hosken J P 1982 *The Hosken Report: Genital and Sexual Mutilation of Females*. Women's International News Network, Lexington, MA
- Koso-Thomas O 1987 *The Circumcision of Women: A Strategy for Eradication*. Zed Press, London
- Kennedy J G 1970 Circumcision and excision in Egyptian Nubia. *Man* 5: 175–91
- Leonard L 2000 'We did it for pleasure only': Hearing alternative tales of female circumcision. *Qualitative Inquiry* 6(2): 212–28
- Lionnet F 1992 Identity, sexuality, and criminality: 'Universal rights' and the debate around the practice of female excision in France. *Contemporary French Civilization* 16(2): 294–307
- Mackie G 1996 Ending footbinding and infibulation: A convention account. *American Sociological Review* 61(6): 999–1017
- Murray J M 1974 *The Kikuyu Female Circumcision Controversy with Special Reference to the Church Missionary Society's 'Sphere of Influence'*. Ph.D., University of California, Los Angeles. Available from Silverprint-Order Entry
- Roheim G 1932 Psycho-analysis of primitive cultural types. *International Journal of Psycho-analysis* XIII: 1–221
- Seligmann C G 1913 Some aspects of the Hamitic problem in the Anglo-Egyptian Sudan. *Journal of the Anthropological Institute* XLIII: 593–705
- Shell-Duncan B, Hernlund Y 2000 Female 'circumcision' in Africa: Dimensions of the practice and the debates. In: Shell-Duncan B, Hernlund Y (eds.) *Female 'Circumcision' in Africa: Culture, Controversy, and Change*. Lynne Rienner Publishers, Boulder, CO
- Talle A 1993 Transforming women into 'pure' agnates: Aspects of female infibulation in Somalia. In: Broch-Due V, Rudie I, Bleie T (eds.) *Carved Flesh/Cast Selves: Gendered Symbols and Social Practices*. Berg, Oxford/Providence
- Talle A 1994 The making of female fertility: Anthropological perspective on a bodily issue. *Acta Obstetrica et Gynecologica Scandinavica* 73: 280–3
- Toubia N 1995 *Female Genital Mutilation. A Call for Global Action*. Women Inc., New York
- Walley C J 1997 Searching for 'voices': Feminism, anthropology, and the global debate over female genital operations. *Cultural Anthropology* 12(3): 405–38
- Widstrand C G 1964 Female infibulation. *Studia Ethnografica Upsaliensia* XX: 95–124

A. Talle

Feminist Economics

This article analyzes the consequences of the underrepresentation of women in the foundation, developments, and applications of economic science.

The use of the term 'feminist economics' appeared in the 1990s with the emergence of a new, extensive, and evolving literature challenging economic theory and research by applying a feminist perspective in all fields of economics. This research, focusing on the relationship between gender and economics, and exploring the extent to which gender bias has influenced economics in all its fields, is based on the idea

that conventional economics is neither universal nor impartial, but that its concepts are built on masculine concerns.

Pluralism and diversity are the key characteristics of feminist economics and they appear in every aspect of the subject. The three main approaches to research in feminist economics are as follows:

(a) one group of studies is devoted to the analysis of the under-representation of women in the economics profession and applies discrimination theories to the profession of economics, stressing the consequences of the absence of women;

(b) a second approach, sometimes described as the 'add women and stir' approach, consists of studies in which a gender component is simply added into classical economics in order to make it more realistic, but without calling into question the fundamental hypothesis and the tools of analysis; and

(c) a third approach comprises studies that fundamentally re-examine the bases and hypotheses of so-called 'universal' and 'gender neutral' mainstream economic theories.

1. Feminism in Economics

The contemporary literature on women in economics developed during the course of the 1960s. Its development was linked to a reassessment of the micro-economic analysis of the behavior of households to adapt to the considerable change in family structures that took place and was due partly to the sustained and continued growth of female labor supply. Within couples, two-earner households have in many countries become the majority. An increase in the number of lone-parent families, mainly women raising their children by themselves, has also been noted.

The use of the label 'feminist economics' appeared more recently with the emergence of a new school of economic thought still marginal within the traditional discipline of economics.

Over the last decade, there has been an increase in research focusing on the relation between gender and economics, which is due largely to the dynamism of the International Association for Feminist Economics (IAFFE), established in 1992 in the USA to provide a feminist perspective in the economics profession. IAFFE annual summer conferences, held since 1992, and the publication of a scholarly journal, *Feminist Economics*, first published in 1995, have contributed considerably to the development of a fruitful challenge of economic theory and research, applying a feminist perspective in all areas of economics. The Association now has over 600 members based in 38 countries.

At the same time, and in a similar vein, the first international conference on feminist economics, entitled 'Out of the margin: Feminist perspectives on economic theory,' was held in Amsterdam in 1993 with over 300 participants from more than 20 nations.

This proved to be a major event, bringing together a series of groundbreaking papers focusing on the relation of gender to economic science. In 1995, a research network working on issues pertaining to gender and work, called 'MAGE,' was created in France, fostering comparative work and cross-disciplinary discussions.

The 1990s also saw the publication of the first anthologies of feminist economics, by Ferber and Nelson (1993), Humphries (1995), Kuiper and Sap (1995), Hewiston (1999) and Peterson and Lewis (1999).

The establishment of the IAFFE in the USA and of MAGE in France and their success testify to the inability of gender-neutral associations to reflect the questions and interests of feminist economists: in these associations, the feminist perspective is anything but mainstream; its sessions generally have a limited audience. The scant interest shown by economists in gender aspects is reflected in the findings of Randy Albelda's (1995) survey of 400 US-based members of the American Economic Association. The salient results (reported by Hewiston (1999, p. 29)) showed that the surveyed economists believed that feminism had made virtually no impact on the methodology of economics or on economic theory (exceptions: labor market analysis 12.3 percent, and household behavior and family economics 11.1 percent); less than one-fifth believed that more space should be devoted in economic journals to 'feminist perspectives on economic analysis'; more than 70 percent either strongly disagreed with or had no opinion on the statement that 'mainstream economics would be enriched if it incorporated more feminist analysis.'

Practitioners of the male sex heavily dominate economic science, and this dominance has heavily influenced the working hypotheses, theories, and methods developed and used. Whereas the discourse associated with research of a high standard is that of objectivity and universalism, the choice of themes covered and methods of analysis would appear to derive inevitably from social constructs which, in the case of economics, are a reflection of tradition: 'Victorian ideology is at the root of the treatment of women in the neoclassical paradigm' (Pujol 1995, p. 17). Economic science is not neutral, it is not a universal science, and it is permeated with sexist hypotheses. The pivotal hypothesis of the rational economic man is not an abstract, sexless concept: the rational economic man appeared to be a male, white, and Western person.

The androcentrism typical of theoretical and applied economic research has engendered bias, particularly in the analysis and design of economic policies, by ignoring the fact that assumptions about gender influence virtually all aspects of economic reasoning. It is this bias, evident in all fields of economic science, which is nowadays being highlighted in publications by feminist economists.

2. Under-representation of Women in Economics

The founding fathers of economics generally regarded women as separate beings, confined to household tasks and devoid of rational behavior. By the same token, the contributions of female economists to the history of economic thought have often been played down or ignored. Even today, female economists are under-represented in positions of responsibility, despite the sizeable growth in the number of female economics students, whose examination results often surpass those of their male peers.

2.1 Women in Economic History

In a forceful book, Pujol (1992) conducted an extensive historical study of the treatment of women by the founding fathers of classical and early neoclassical economics. In a later essay (Pujol 1995, p. 18) she analyzed five elements that characterize early neo-classical views of women based on the writings of the founding fathers, Marshall, Pigou, Edgeworth and Jevons:

- (a) All women are married, or if not yet, they will be. Similarly, all women have or will have children.
- (b) All women are (and ought to be) economically dependent on a male relative: father or husband.
- (c) Women are (and ought to be) housewives; their reproductive capacities specialize for that function.
- (d) Women are unproductive (whether absolutely or relative to men is not always clear) in the industrial workforce.
- (e) Women are irrational, they are unfit as economic agents, and they cannot be trusted to make the right economic decisions.

These elements, which fell within the unstated and undiscussed assumptions, justified the poor position of women in the labor market or their exclusion from it. Pujol (1995, p. 29) argued that 'approaches to women in the neoclassical paradigm have not changed much since the founding of the paradigm in the last decade of the nineteenth century.'

Pujol (1992) also underlined another important aspect of the under-representation of women in economics, namely the underestimation of the contribution by female economists to the history of economic thought. She showed the influence of Harriet Taylor on the work of John Stuart Mill as well as her personal contribution, and presented various works on the condition of women by Bodichon, Millicent Garret Fawcett, Adair Heather-Bigg, Eleanor Rathbone and Beatrice Potter Webb that were ignored by historians of economics.

One example, among many others, of the underestimation of women's contributions to economics is the scant attention paid to the writings of Margaret Reid (1896–1991), who worked on households, and of Charlotte Perkins Gilman (1860–1935), who focused specifically on gender in economic relationships.

'Reid's colleagues at the University of Chicago, Theodore Schultz and Gary Becker, incorporated Reid's idea of household production into their 'new home economics' (Woolley 1999, p. 335). 'Reid's contribution has been largely ignored in the histories of the 'new home economics,' which Becker is credited with "fathering"' (Lewis 1999, p. 437).

2.2 Women in the Profession

As in most professions in Europe and the USA, there is vertical segregation among economists: women are under-represented in positions of responsibility, in all sectors of activity.

In academic circles they do not occupy the most senior positions and are under-represented in the most prestigious institutions. They do not win prizes: not one woman has so far received the Nobel Prize for Economics; only exceptionally do they chair associations of economists, both in the USA and in Europe; likewise, their representation on the boards of journals is poor.

In the publicity leaflet for *Who is Who in Economics* (Blaug 1999), only two out of 19 photographs are of women: Joan Robinson and the newcomer Deirdre McCloskey.

This segregation is also to be found in the field of public administration as well as private companies. The glass ceiling, that invisible barrier hindering women's access to top jobs, is operating efficiently in economics.

Whereas there is everywhere a growing number of female students and graduates (first degrees) in economics, whose examination results are in no way inferior to those of their male peers, fewer women proceed to Ph.D.s and pursue academic careers. The factors put forward in an attempt to explain these behavioral differences include the following:

(a) The absence of women in the classical economics textbooks and the choice of male-oriented or even phallographic examples to illustrate theories: the white Western male is the norm in economics textbooks, but he is concealed behind a gender-neutral, universal façade. Similarly, few pages deal with gender and inequality: women are excluded as objects of investigation.

(b) A classroom environment unfriendly to women, the shortage of female lecturers to act as role models, the examination process, the intensive use of mathematical tools and abstract modeling have also been mentioned; however, the superior examination results of young women in mathematics do not corroborate this last explanation.

Moreover, in the same way as women occupy a traditional role in civil society, if female economists are to be acknowledged and valued by their male colleagues they must confine themselves to narrowly circumscribed fields: home production, voluntary

work, discrimination. Women economists have often written about so-called 'women's issues' such as women's labor force activity and wages, which seemed of little interest to male economists and benefit from a lower level of reward than other fields.'

The Massachusetts Institute of Technology established a committee on women faculty to understand and ameliorate the long-term effects of discrimination in academia: their key conclusion is that (Massachusetts Institute of Technology 1999, p. 3) 'gender discrimination is subtle but pervasive and stems largely from unconscious ways of thinking that have been socialized into all of us, men and women alike. ... Our first instinct is to deny that a problem exists, or to blame it on the pipeline or the circumstances and choices of individual women.' These conclusions could probably be extended to most faculties and universities in the world.

3. Consequences of Women's Under-representation in Economics

Women's under-representation engenders bias and error in the foundations, development and applications of economic science. The androcentric bias in basic neoclassical hypotheses is substantial: historically the entire theory of utility is based on the rational behavior of a *Homo economicus* who may choose to allocate his time to paid work or leisure. Only in recent economic research have male economists realized that life does not boil down to only a choice between paid work and leisure.

Women and the family are invisible in the economic analysis of human behavior: even when gender is central to the subject under investigation it is often ignored, as evidenced by studies of poverty.

3.1 Sex Discrimination

The question of inequality between the sexes has not been the subject of any key works of normative economics; instead, discrimination between men and women is either included in the broader issue of social inequality or ignored.

Discrimination has mainly been the subject of writings in applied and comparative economics, which have quantified it and measured its development but have left extensive shady areas in respect of its causes and dynamic. Two approaches have been followed in formulating the models appraised: human capital theory, which explains the differences evident between men and women on the labor market through differences in the investment in human capital—education, training, experience, health—and discrimination theories, which analyze the differences between individuals whose production capacity is identical—discrimination exists when persons of equal capacity

and with the same skills are treated differently because of their sex (Blau and Ferber 1992).

The factors explaining the segregation of men and women are to be found in the theories of labor supply and demand, institutional theories, and transaction cost theories (Jonung 1998, pp. 36–71). Labor-supply theories highlight the differences between the preferences and aptitudes of men and women, the socialization of roles and the domestic division of labor, all of which affect decision making; demand theories explain segregation in terms of the differing investment in human capital and the notion of discrimination that colors the attitude of employers towards women. Transaction-cost theories point to the incomplete information available to employers, which leads them to adopt discriminatory attitudes based on observations of average behavior; and finally, institutional theories stress the legal and social constraints, the traditions that influence individual decision making.

3.2 Household Economics

For feminist economists, existing economic theory and analysis are unsatisfactory when it comes to studying intrafamily relationships and the distribution of resources within the family unit. According to the neoclassicists, the household or family is a black box: it matters little who produces or who consumes resources; this share-out is a technical matter or one of preference, something of no concern to economic analysis.

The situation of households is characterized by global earnings or global consumption managed by an altruistic head of household who looks after the interests of his loved ones and behaves as a rational economic agent.

This simplistic view can be found in most European-level comparative studies on poverty or earnings inequality. The hypothesis is that resources are shared equally among the members of a household: poor individuals belong to poor households. Households are classified according to the age, occupation, and sector of activity and educational level of the head of the household. Rarely does the composition of the household feature, other than through the simplistic application of a scale of equivalence.

However, certain studies have underlined the importance of power in the distribution of earnings and consumption within households, not merely of preferences and abilities, the sole elements to impinge on the reasoning of economists. Recent empirical research has confirmed that the allocation of resources among family members depends on age and gender. Studies conducted in Asia and Africa have emphasized the high rates of female mortality due to the systematic deprivation of resources suffered by women within households (not necessarily poor ones; see Sen (1989)).

In one of the few studies on distribution within American households, Lazear and Michael (1988) reveal that the division of consumption between parents and children alters radically with the number of adults in the household: as the number of adults rises, the children's share declines proportionately. Furthermore, access to paid work considerably changes the share-out of power within the household. On average, women earning an income have more decision-making power within the family, and this share of the power increases in step with their earnings (England and Kilbourne 1990).

Findlay and Wright (1996, pp. 335–351) have assessed, for Italy and the USA, the way in which an unequal distribution of resources within households affected the measurement of poverty. They conclude that the conventional methods of gauging poverty within households lead to a systematic underestimation of female poverty and an overestimation of male poverty.

The criticisms of feminist economists have addressed the lack of analysis of the household as an economic entity, but they have also addressed the problematic analysis of the household in the new home economics: 'For feminists, the minimum requirement is a model of the household that allows for the conflict of interests and patriarchal power relations within and outside the family' (MacDonald 1995, p. 185). New home economics, an example of the 'add women and stir' type of studies, is based on an implicit hypothesis of women's natural superiority in nursing and housekeeping which justifies, from the point of view of efficiency, the division of labor within the family. It used to treat the sexual division of labor in the family as if the differences were biologically predetermined and not socially constructed. The neoclassical model or child support model disadvantages women by obscuring and justifying men's neglect of their parental responsibilities.

Bargaining theories and game theory are also becoming increasingly popular in dealing with intra-household allocations: the game theory model of household decision making is an interesting approach to overcome the restrictive assumptions of household intrafamily consent and formally develop a more realistic family behavior pattern. One application is the treatment of 'fertility as a prisoner's dilemma' (Ott 1995, p. 89). However, formal modeling limits the reality of analysis and the 'cooperative conflict' approach developed by Sen (1990) and Agarwal (1994) which incorporate qualitative, historical, and cultural variables provides an interesting alternative to game-theoretic models.

3.3 Labor Economics

Labor economics, along with the new home economics, is the first field to incorporate gender into the neoclassical framework. Gender has been included in

the models without challenging the theoretical foundations; it was merely a matter of arriving at a more accurate analysis of observable reality.

The mainstream understanding of women's employment in labor economics is 'men as breadwinners and women as second earners.' The concentration of women in low-wage occupations and atypical employment appears to be a result of their own choice.

In fact, there is circular reasoning at work here: women choose less well-paid posts which enable them to take on the domestic chores for which they have a preference and are made in the knowledge that the performance of domestic chores reduces the earnings which they could command from employers.

Numerous econometric studies (Silvera and Sonnac 1996) have attempted to explain the persisting wage gap between women and men: even when controlling for differences in working time, educational attainment, training, tenure, occupations, sectors, etc., an unexplained gap amounting to 11–14 percent of women's wages remains. This gap, a measure of pure discrimination, exists even in the countries where protective laws concerning wage equality are in application. However, and contrary to neoclassical principles, women's lower wages are not associated with lower unemployment rates: in nearly all European Union countries women's unemployment rates are considerably higher than men's.

3.4 Data and Methods in Empirical Economics

The same bias is evident in the development of statistical tools: to include a feminist perspective in empirical economic analysis, new kinds of data and indicators have to be constructed, and there is a need for more qualitative data and for panel data sets.

The available data concerning the labor market, for example, are constructed around standard male practice: full-time paid work without interruptions. In most cases these are individual data with no information as to the individuals' household or lifestyle.

The data available from national statistical agencies to study women on the labor market are inadequate in many ways (MacDonald 1995, p. 188):

(a) there is a preponderance of cross-sectional data, and very few panel data;

(b) the level of aggregation makes it impossible to study issues of segmentation within occupations;

(c) labor force data are collected on the individual, with information on the family context either missing or inadequate;

(d) the use of categories such as 'personal or family reasons' versus 'economic reasons' in labor force surveys, to classify reasons for not looking for work or for working part-time, embed gender biases in the data;

(e) skill measures have been shown to be gender biased (Boyd 1990); and

(f) it is difficult to measure changes in the quality of jobs.

National accounting ignores all non-monetary aspects: women's traditional contribution to growth—domestic work and child rearing—are never quantified. To develop alternative indicators of economic and social welfare is an important component of feminist economic research.

Bias in measuring economic growth, ignoring domestic work, leads to wrong political decisions insofar as no account is taken of cost when policies negatively affect home production. The fact that no value is attributed to home production, and that it is statistically overlooked, leads to an inferior allocation of public expenditure.

3.5 Policies

One example of the adverse effects of mainstream economics on economic policy is the structure of social protection systems. These often refer back to the neoclassical patriarchal model: a single main income provider per household. This reference engenders discriminatory practices, which can adversely affect the female supply of labor and can keep women in a situation of dependence. The existence of discriminatory measures, the trend in family structures, and the contradictions between the labor market and social security systems are three elements that lie behind the issue of individualizing social security entitlements.

As they operate at present, Europe's social security systems exert a discriminatory effect on women by two means. First, these systems do not ensure adequate protection for workers in atypical circumstances, among whom women are over-represented: such persons are often deprived of proper protection, mainly with respect to unemployment and pensions. Second, the absence of individual entitlements and the proliferation of derived rights forces women into relationships of dependence tinged with insecurity, which discourage them from entering the official labor market and results in inequalities between the benefits received according to family circumstances. Europe's taxation and social security systems contain various 'inactivity traps' which were deliberately set and still exist: derived rights in social protection systems are an incentive for women to remain at home or to engage in undeclared work, since marriage automatically confers on a spouse the healthcare and pension benefits enjoyed by the partner. The advantages granted to the home-based spouse by the tax system, namely due to dependents' allowances, as well as all the allowances linked to inactivity, such as career breaks, constitute inactivity traps that affect a far larger number of persons of working age than the unintentional traps that are currently being singled out for corrective action (Jepsen et al. 1997, pp. 81–98).

Whereas the Welfare State gave men their inde-

pendence in the market, it also reinforced the economic dependence of women on men by ignoring the sexual division of labor.

4. The Future: a Reconstruction of Concepts

Ferber and Nelson (1993, p. 8) identified five, not necessarily mutually exclusive, responses to perceived inadequacies in the academic discipline:

(a) Affirmative action: based on the idea that the central problem is the under-representation of women and acting to correct this by monitoring but not criticizing the discipline itself.

(b) Feminist empiricism: not calling into question the tools but rather the way in which they are used and the manner in which the issues are addressed: 'It is not the theory that is patriarchal, but the questions male economists have asked and the conclusions they have drawn and particularly the policy implications based on the research' Gustafsson (1990, p. 6); Ferber and Nelson (1993, p. 8) noted that 'Casual conversation suggests that most feminist economists currently adhere to this view.'

(c) Feminist difference: the replacement of male bias by a feminist bias, not a quest for objectivity but subjectivity, emotion, holistic approach (less well represented).

(d) Feminist postmodernism: to deconstruct traditional conceptions, the key point: is gender a meaningful categorization?

(e) Feminist constructionism: analyzing the social production of gender identity.

Whatever the classification, the reconstruction of economic concepts underpins the work of feminist economists: the concepts used for analysis and the way concepts are linked together in the theory may be better suited to explain the activities of some groups than others: they are culturally biased and show a tendency to treat male values and practices as universal. A gendered view applied to all parts of economics questions the hypothesis, methods, and data used.

The aim is also to take greater account of complex social constructs that have been ignored or inadequately incorporated into economic modeling. This means taking account of societal variables often neglected by economists, and implies a reconstruction of traditional databases that are unsatisfactory for gender mainstreaming.

Feminist economics also considers gender as a central category for analysis: a feminist reconstruction of economic theory can improve all the different fields; this new economics will not just address feminist issues, but will be better economics.

See also: Economic Development and Women; Economic Globalization and Gender; Economics, History of; Economics: Overview; Feminist Epistemology;

Feminist Political Ecology; Feminist Theory; Gender and Feminist Studies in Economics; Gender and Feminist Studies in Sociology

Bibliography

- Agarwal B 1994 *A Field of One's Own: Gender and Land Rights in South Asia*. Cambridge University Press, Cambridge, UK
- Albelda R 1995 The impact of feminism in economics—beyond the pale? A discussion and survey results. *Journal of Economic Education* 26: 253–73
- Barker D K 1999 Gender. In: Peterson J, Lewis M (eds.) *The Elgar Companion to Feminist Economics*. Edward Elgar, Cheltenham, UK, pp. 393
- Blau F D, Ferber M A 1992 *The Economics of Women, Men, and Work*. Prentice Hall, Englewood Cliffs, NJ
- Blaug M (ed.) 1999 *Who is Who in Economics*. Edward Elgar, Cheltenham, UK
- Boyd M. 1990 Sex differences in occupational skill: Canada 1961–1986. *Canadian Review of Sociology and Anthropology* 27: 287–315
- England P, Kilbourne B S 1990 Marriages, markets, and other mates: The problem of power. In: Friedland R, Robertson A F (eds.) *Beyond the Market Place: Rethinking Economy and Society*. Walter de Gruyter, New York, pp. 163–8
- Ferber M A, Nelson J A (eds.) 1993 *Beyond Economic Man—Feminist Theory and Economics*. University of Chicago Press, Chicago
- Findlay J Wright R E 1996 Gender, poverty and the intra household distribution of resources. *Review of Income and Wealth*, Series 42, No. 3, September, pp. 335–351
- Gustafsson S 1990 Half the power, half the incomes and half the glory: The use of microeconomic theory in women's emancipation research. Inaugural Lecture at the Official Assumption of the Chair of Labor Market Issues with Special Attention to Women's Emancipation. University of Amsterdam
- Hewitson G J 1999 *Feminist Economics. Interrogating the Masculinity of Rational Economic Man*. Edward Elgar, Cheltenham, UK
- Humphries J (ed.) 1995 *Gender and Economics*. Edward Elgar, Aldershot, UK
- Jepsen M, Meulders D, Plasman O 1997 Protection sociale: le rôle des droits dérivés. *Les Cahiers du Mage* No. 3–4, pp. 81–98
- Jonung C 1998 Occupational segregation by sex and change over time. In: Persson I, Jonung C (eds.) *Women's Work and Wages*. Routledge, London, pp. 36–71
- Kuiper E, Sap J (eds.) 1995 *Out of the Margin*. Routledge, London
- Lazear E P, Michael R T 1988 *Allocation of Income Within the Household*. University of Chicago Press, Chicago
- Lewis M 1999 History of economic thought. In: Peterson J, Lewis M (eds.) *The Elgar Companion to Feminist Economics*. Edward Elgar, Cheltenham, UK, pp. 433–43
- MacDonald M 1995 The empirical challenges of feminist economics. In: Kuiper E, Sap J (eds.) *Out of the Margin*. Routledge, London, pp. 175–97
- Massachusetts Institute of Technology 1999 *A Study on the Status of Women Faculty in Science at MIT*. Massachusetts Institute of Technology, Boston, mimeo
- Ott N 1995 Fertility and division of work in the family. In: Kuiper E, Sap J (eds.) *Out of the Margin*. Routledge, London, pp. 80–99

- Persson I, Jonung C (eds.) 1998 *Women's Work and Wages*. Routledge, London
- Peterson J, Lewis M (eds.) 1999 *The Elgar Companion to Feminist Economics*. Edward Elgar, Cheltenham, UK
- Pujol M A 1992 *Feminism and Anti-feminism in Early Economic Thought*. Edward Elgar, Cheltenham, UK
- Pujol M A 1995 Into the margin. In: Kuiper E, Sap J (eds.) *Out of the Margin*. Routledge, London, pp. 17–34
- Sen A 1989 Women's survival as a development problem. *Bulletin of the American Academy of Sciences* 43: November
- Sen A 1990 Gender and cooperative conflicts. In: Tincker I (ed.) *Persistent Inequalities: Women and World Development*. Oxford University Press, New York, pp. 123–49
- Silvera R, Sonnac N 1996 *Le Salaire des Femmes: Toutes Choses Inégales par Ailleurs*. La Documentation Française, Paris
- Woolley F 1999 Family, economics of. In: Peterson J, Lewis M (eds.) *The Elgar Companion to Feminist Economics*. Edward Elgar, Cheltenham, UK, pp. 328–35

D. Meulders

Feminist Epistemology

Reflecting on feminist epistemology, namely the claim that there are distinctively feminist perspectives on the theory of knowledge, initially proposed at the end of the 1970s, is a little like recalling the English Chartist movement with its utopian program of radical political reform. Within what historians would regard as rather little time, what was seen as an absurd and impossible political project, not least the demand for universal suffrage, was to become an everyday reality. The utopian dreamers turned out to be more practical than the dismissive establishment. The equally utopian proposal for a distinctively feminist epistemology was placed firmly on the theoretical agenda of the second wave of feminism by the end of the 1970s. Within just two further decades, substantial aspects of the feminist epistemology debate began to find acceptance, not just among the diversity of feminists, but much more widely within culture and society.

That hot debate among the scholars and activists of the women's movement, which gave birth to the claims for feminist epistemology, took rather little notice of disciplinary boundaries. The new field they were in the process of creating, 'Women's Studies,' was self-consciously not only highly innovative but was also transgressive. Having found that the academic gaze was blind to the bodies and lives of women, to say nothing of sex–gender relations, there seemed little purpose in deferring to the canon of the old exclusionary disciplines. Thus the new epistemology drew in theorists whose disciplinary formation lay in philosophy, psychoanalysis, biology, history, physics, and sociology, but whose cultural products often showed little trace of their formation. These feminist epistemology debates, without quite ever settling down and forging an unequivocal consensus, have entered

and helped foster profound changes in many academic disciplines, particularly where the mode of production of knowledge requires rather modest capitalization. The so-called 'Little Sciences' have proved much more susceptible to change than the 'Big Sciences.' Thus the most power charged techno-sciences of molecular biology and informatics are proving to be remarkably resistant to the kinds of internal changes that primatology or sociology have experienced.

1. *The Birth of Academic Feminism*

Despite the power of late twentieth century feminism to touch women worldwide, the development of academic feminist theorizing required financial support. During this key early period support given to Women's Studies by both the Ford and Rockefeller foundations enabled US academic feminism to develop more strongly than its counterpart within Europe, not least in the development of feminist approaches to the theory of knowledge. By contrast, although Scandinavian research policy was swift to support Women's Studies, it did so as part of policy-oriented, rather than as basic research, while in Western Europe, with the exception of The Netherlands, support from either the state or from foundations was slow to arrive. Thus, while there was a rich outpouring of articles and books published by feminists from many countries criticizing existing theories of knowledge and proposing alternatives, the confident assertion that there were distinctively feminist perspectives on epistemology had to wait until the publication of *Discovering Reality* (Harding and Hintikka 1983). The editors, the US philosophers of science, Sandra Harding and Merrill Hintikka, described their theoretical project.

During the 1990s, feminist thinkers provided brilliant critiques of the political and social beliefs and practices of patriarchal cultures. But less attention has been given to the underlying theories of knowledge, and to the metaphysics which mirror and support patriarchal belief and practice. Are there—can there be 'distinctive feminist perspectives on epistemology, metaphysics, methodology, and philosophy of science' (p. 9)?

The US contributors, divided between deconstructionist accounts of the values embedded within organized knowledge or scientia, and those which try to set out the conditions for rebuilding these as feminist knowledges. A key target was the natural sciences, which had, with immense success, claimed to be above culture. In setting the natural sciences as their target, the feminist critics of science shared more than a little in common with the new, post-Mertonian, sociology of science, although there was little visible connection between the two. However, the very title of the book indicates that these were daughters, albeit dissident daughters, of the enlightenment. As feminist critics of

science, they sought to replace what they saw as flawed accounts of reality with more reliable accounts free from ideologies and relations of domination. For this generation of theorists, the fact that reality was both real and knowable was not significantly at issue; the nominalism which was to sweep the 1980s and 1990s, was not, at that moment, a powerful current. Recognizing the power of foundationalist thought, the authors set out to analyze the patriarchal constructions of those cultural giants who had deeply shaped Western culture from antiquity, to the early theorists of modernity to the late modern figures of Darwin, Marx, and Freud. Within the natural sciences it was not by chance that the focus was Darwin and the life sciences, and not, say Einstein and physics, for biology was the discourse which claimed cultural authority to define women's nature.

The definition of women's nature was an old battleground, with pioneer nineteenth century feminists arguing that in Darwinian theory, only men evolved. Unquestionably, scientific sexism along with scientific racism demanded critical attention; clearing away the culture of biology-as-destiny would help create a political space in which new human subjects could claim agency. Where in the past such detailed critiques had been mounted predominantly by progressive men biologists, as women entered the laboratories, their consciousness raised by the second women's movement, they too had the skills to enter the lists in their own right. The pioneers among these feminist biologists include Ruth Hubbard, Marian Lowe, Ethel Tobach, Betty Roscoff, Anne Fausto Stirling, Ruth Bleier, and Lynda Birke. These insisted that the androcentricity which produced and reflected a patriarchal culture also produced poor biology, and that a more faithful, more reliable account of women's nature could only be produced through feminist science resisting the relations of domination. It was the pioneering work of Carolyn Merchant (1980) as a historian of science that linked the problem of gender domination with the domination of nature. While this link by the feminist historians opened a continuing connection with the burgeoning environmentalist movement, the feminist life scientists made a powerful connection with the women's health movement. Both had, and have, epistemological and political implications.

Their deconstructionist critiques (without Derrida) constituted an immense methodological challenge to contemporary biomedical research. They charged successfully that biomedicine had taken the body of the male of European bio-geographic ancestry, or 'race,' and fetishized it as the universal human body. They demonstrated forcefully the ways through which women's bodies, apart from their problematic reproductive systems and their troublesome psyches, along with the bodies and minds of other Others, were subsumed within a demonstrably false universalism. Under the impact of this epistemological questioning,

together with the increasing presence of historically excluded groups within the laboratories, the discourse of biomedical research has gradually but dramatically changed, above all in the study of the pre-eminent lethal diseases of the West. Thus, heart research now routinely explores difference without hierarchy, where previously the study of the heart of the pale male was equated with some universal human organ. Researching cancer also underwent a sea change. It was not simply the immense pressure to address breast cancer seriously but, almost ironically, this attention to sex–gender difference has led to greater attention being paid to the specific problems of men, not least prostate cancer. Even mental health began to address gender and ‘race’ difference without automatic notions of domination. This was feminist epistemology in action.

2. *The New Feminist Epistemology*

This successful critique of biomedical science as skewed and inadequate, bad for women and bad for science, fostered an initial comfortable solidarity between the pioneer feminist science critics, embracing both the feminist biologists and those, typically from the social sciences and humanities, engaged in setting out the theoretical project of a distinctively feminist epistemology. One of the most influential theorists of the new feminist epistemology was political scientist Nancy Hartsock (1983), who formulated and named feminist standpoint theory. Her socialist feminist project revisioned Marx’s theory of knowledge. Taking the project of a proletarian science, she runs it through the sieve of feminist analyses of the specificities of women’s reproductive labor. She argues that it is a feminist consciousness that gives rise to the distinctively feminist standpoint. This approach was shared by Harding in the same volume, and by sociologist Hilary Rose in her *Signs* article (1978). In the important earlier natural science issue of *Signs*, the historian Donna Haraway (1978) had already spelled out the need to rebuild the life sciences outside the relations of dominance. In this project of a successor science these feminists reflected the strong Hegelian and Marxist influences in their work. The heretical claim that nature and not just society could be more reliably ‘seen from below’ challenged the hegemony of the claims of the natural sciences to be a culture of no culture.

Feminist standpoint theory, particularly when it was directed to the study of nature, represented a theoretical development from the ‘two sciences thesis,’ articulated initially as the struggle between proletarian and bourgeois science by the social relations of the science movement of the 1930s and 1940s, lost in the Stalinist period, but subsequently re-articulated as a rainbow politics of ‘science for the people’ by the radical science movement of the 1960s and 1970s. Other influential feminist science theorists came at the

epistemology question from very different theoretical and political trajectories. Thus, philosopher Helen Longino, while concerned to define a feminist science, did so by specifically repudiating the Hegelian master/slave argument; for her the view from below carried no epistemic privilege. Evelyn Fox Keller, by contrast, took a psycho-historical approach, focusing on the gendered origin stories of science. She saw masculinity as embedded in that mechanical philosophy which had triumphed in those founding years of both modern science and capitalism.

Not only did early feminist epistemology theorizing show little deference to the existing canon, initially it paid little formal attention to the paradigm-breakers emerging from the crisis of French social thought in the wake of the defeat of 1968. Few of these early feminist epistemological texts paid attention to the theorists of post-structuralism and post-modernism. As noted above, self-proclaimed feminist deconstructionists paid little attention to Derrida or to the discourse analysis of Foucault, or to Lyotard’s post-modernistic project, even though debates were raging around these in the mainstream disciplinary discourses. Despite the parallels which have led to some over facile assertions such as ‘feminism is part of post-modernism,’ these new currents were relatively slowly taken into feminist theoretical production. They first entered the humanities and then, with continuing resistance, the social sciences.

3. *Disciplining Feminism?*

Feminism’s successful arrival within the academy by the end of the 1980s brought with it substantial epistemological implications, not least the restoration of disciplinarity within feminist discourse. Feminists began to address the specific canon and current debates of their own disciplines; the ethical and political commitment to address the feminist movement both in and out of the academy increasingly weakened. The language of feminist theorizing soon changed. The early commitment to transparency and accessibility ceded to a new complexity, even obscurantism. There was a sudden interest in post-modern theorists, whom an early, more politicized, generation of feminists had set aside for their anti-feminism; earlier theorists such as Nietzsche became widely discussed despite an evident anti-semitism and misogyny. While disciplinary boundaries appeared, the claims of theorists became grandiose. Feminists in the humanities often spoke of ‘feminist theory,’ by which they meant more precisely feminist humanities theory, but tacitly let the term carry with it the aggrandizing claim that this had universal power over the diversity of fields. Even feminist sociology, something of a bastion of critical realism, entered a period of self-questioning. When the differences between women, whether of class, ethnicity, and bio-

geographic ancestry were so very visible, how did it make sense to speak of a single feminist standpoint? Theoretical developments responding to these challenges by feminist sociologists such as Canadian Dorothy Smith, German Frigga Haug, and African American Patricia Hill Collins gave standpoint theory a new and enriched life.

4. Feminist Science Wars?

Among feminist science studies scholars it was the historian Donna Haraway who embraced post-structuralism decisively, simultaneously throwing into question the possibility of universal reliable knowledge. In a key paper (which echoed silently philosopher David Bloor's methodological stance of refusing to privilege the truth claims of the natural sciences), Haraway, despite her respect for Hubbard as a pioneering feminist biologist, pointed to her uneven treatment of truth claims. She argues that for Hubbard, good science remained privileged; only bad science was to be deconstructed. This epistemological difference between the feminist scientist and the feminist historian foreshadowed, with some strained courtesy, the divisions among the feminist approaches to science theory which were to develop.

However, the tensions between feminists science theorists were never characterized by the savagery of those within mainstream epistemological fights, above all those of the Science Wars. Science theory feminists, despite their theoretical differences, frequently shared a desire to build solidarity between those in science studies and those in science. Thus it was matter of common recognition that natural scientists, particularly biologists, had to be realists. It was equally clear that, for science theorists, more relativistic positions were entirely possible, whether they were historicized as by materialism or full-blown as by nominalism. Feminist science theorists routinely drew back from stark 'either/or' choices, not least those dichotomously opposing relativism to realism. Thus, where the early Harding had shared Rose's concern at the hyper-reflexivity of the strong program of science, and the reduction of science just to social relations proposed by Robert Young and the *Radical Science Journal*, her subsequent publication suggested a new theoretical ambiguity. Certainly reviewers who were outside science theory of *The Science Question in Feminism* (Harding 1986) diversely claimed the book both for standpoint theory and as support for the more relativistic approach of post-structuralism. The ambiguity was not accidental but typical of the feminist science theorists as they fought to avoid an unwelcome dichotomy.

Later, Haraway made a sophisticated attempt to spring the 'either/or' trap: unwilling to lose all claim to objectivity she proposes a feminist objectivity built from 'partial critical knowledge located in multiple

embodied sites sharing a web of connections called solidarity in politics.' Harding too made an attempt to go beyond the 'either/or' dichotomy in her advocacy of 'strong objectivity,' an argument which has much in common with the earlier externalist analysis of Marxism.

5. Global Technoscience and Global Feminisms

In a quite extraordinary way, the feminist epistemology debates have paralleled and offered an entirely different face to the widespread demands for equal opportunities within the labor market, including that of science. This, whether claimed for reasons of justice or for economic efficiency, has become subtly linked to arguments about changing the character of cultural production, typically side-stepping whether the claim entails an appeal to social, cultural, or essentialist difference. Today, innovative corporations concerned for their global markets sound remarkably like dedicated multiculturalists in their arguments that products aimed at a socially diverse market are better if designed and produced by a matching diversity of producers. That the former tend to equate better with 'sells more' and the latter equate better with 'is more democratic' and 'will produce a more sustainable science and technology' is almost incidental. Both are responding to the feminist epistemological challenge that who gets to produce knowledge, and who is excluded, matters.

But it has been the advent of what social theorists have termed the risk society, that is, where the development of science and technology have become associated intrinsically with major environmental risks, which has also fostered an immense turn towards democratic control to protect the entire socio-ecosystem. This immense cultural and political move has paralleled the struggle within the sciences. Working from outside science, a mixture of NGOs and mass popular movements from both the North and the South have confronted the technosciences and the corporations. As we have seen with the fate of the Kyoto agreement, individual governments and global corporations continue to mobilize the old constructions of science as outside culture and capable of producing certainty, in order defend themselves against the need to take socio-ecological risk seriously. Nonetheless, although it would be a mistake to announce the environmentalist cause won, it has become less and less easy for the corporations and supportive governments to pursue their commercial and technological objectives without any sense of global responsibility and concern for either human or green nature. Where they display such indifference they find themselves confronted by new alliances capable of mobilizing immense popular criticism. The more sophisticated recognize the need for sustainable development and are more open to new more socially

inclusive forms of governance. How far these are serious or tokenistic is too soon to judge.

It is only possible in a brief article to do little other than hint at the links between the feminist epistemology debates and these huge cultural and political developments. Sometimes, as with the mass people's science movement in India and the work of the feminist science theorist and activist Vandana Shiva, the links are more evident. But it will be historians looking back at this extraordinary shift in the cultural status of science and technology at the cusp of the twentieth and twenty-first centuries, who will have the pleasure of disentangling the cultural and political contribution of the feminist epistemology debate to these sea changes.

See also: Feminist Ethics; Feminist Political Theory and Political Science; Feminist Theory; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Gender and Feminist Studies; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Rationality and Feminist Thought; Situated Knowledge: Feminist and Science and Technology Studies Perspectives

Bibliography

- Bleier R 1986 *Feminist Approaches to Science*. Pergamon, Oxford, UK
- Haraway D 1990 *Primate Visions: Gender, Race and Nature in the World of Modern Science*. Routledge, New York
- Harding S, Hintikka M (eds.) 1983 *Discovering Reality: Feminist Perspectives on Epistemology, Metaphysics, Methodology and Philosophy of Science*. Reidel, Dordrecht, The Netherlands
- Harding S 1986 *The Science Question in Feminism*. Open University Press, Milton Keynes, UK
- Keller EF 1985 *Reflections on Gender and Science*. Yale University Press, New Haven, CT
- Longino H 1990 *Science as Social Knowledge, Values, and Objectivity in Scientific Enquiry*. Princeton University Press, Princeton, NJ
- Rose H 1978 Signs: Special issue on women, science and society. *Journal of Women in Culture and Society* 4(1)
- Rose H 1989 *Staying Alive: Women, Ecology and Development*. Zed Press, London
- Rose H 1994 *Love, Power and Knowledge: Towards a Feminist Transformation of the Sciences*. Indiana University Press, Bloomington, IN

H. Rose

Feminist Ethics

Feminist ethics has its roots in the more general feminist theory and politics, going back to Mary Wollstonecraft's Declaration of the Rights of

Women, and on through the work of Jane Addams and Charlotte Perkins Gilman early in the twentieth century, work that took women's moral visions seriously. As an intellectual discipline in Anglophone nations, it was born in the 'second wave' of the Women's movement in the mid- to late-1960s, taking a few clues from de Beauvoir (1948). The early feminist theories in this 'wave' were written by activists in the movement as position papers to provide conceptual and theoretical bases for charting strategy. The basic concerns and projects of feminist ethics today have grown from the political roots.

As time passed, the feminist activists entered professions across the disciplines. They established journals and courses and eventually the discipline of women's studies itself was legitimized and institutionalized, with its own subspecialties that include feminist ethics. Although feminist intellectual work has always been multidisciplinary, its institutionalization has taken place in the traditional context of departments, course assignments, disciplinary specialties, etc. One outcome has been that feminist ethics, like traditional ethics, has tended to fall within the territory of the humanities, despite the fact that the social sciences also have essential contributions to make.

Intellectually, feminist ethics was shaped by criticisms of the traditional moral theories that were developed to suit life in secular, Western capitalist democracies. Moral subjects and agents were defined as autonomous individuals who were equal under morality and the law—a stereotypically male subject. The traditional theories incorporate, for the most part, a variety of dualisms in which one side is valued over the other, underwriting an ethics that supports dominance and subordination of many sorts. The dualisms fundamentally value reason over emotion, with men identified with reason and mind and women (and the racially subordinated) identified with emotion, body, and nature. Out of this soil a variety of feminist moral theories have grown, ranging from modified, liberal theories of equal opportunity and rights; to theories of mothering, care, and relationship; to lesbian ethics; liberatory ethics; standpoint approaches; third world and minority approaches; and the feminist ecological ethics that is now being developed.

From the beginning, feminists worked in 'applied ethics,' for example, on moral issues of abortion, violence, prostitution, AIDS, clitoral cutting, the environment, peace, immigration, and many others. All of these theorists insist on taking women's experience seriously. In addition, some feminists focus on issues of power and dominance, insisting that feminist ethics must be 'liberatory' and provide guides for action. The liberatory focus can be seen as asking, 'What would an ethics be like that takes patriarchy, or racism, or anthropocentrism seriously as a social and political structure?'

The dominant public ethics in the Western democracies is an abstract moral theory of impartial reasoning, under which autonomous individuals freely chose principles of rights and justice and make their decisions accordingly. It has been used to good effect in some places as women argue for equal opportunity and rights. The impartial ethics is implicitly radical because it is supposed to apply to all individuals regardless of gender, race, age, ethnicity, relationship, etc. In practice, the radical extension to all individuals was politically won during nineteenth and twentieth century struggles to extend the suffrage, change property laws, and so on. In the process, the ethics itself was modified so as to include women (and others) among the autonomous individuals—leading to the intellectual discovery that the original moral agents of the theory were conceived to suit a stereotypically Western, white, higher class, male model. A family of feminist moral theories sprang up to remedy the problem.

The dominant Western ethics of justice, rights, and autonomy relies heavily on the reason–emotion dualism. One feminist alternative has been to remedy the imbalance by emphasizing women’s experience of care and responsibility. Long ago, Addams (1907) called for a ‘social ethic’ that would offer a guide to women’s action by expanding women’s morality in home and family into the public sphere—the Settlement House movement and the women’s peace movement were outcomes. This alternative puts the focus on care and relationship rather than on autonomous agents acting individually according to their interests, self-chosen principles, etc., and it has provided an important direction for feminist ethics. Diana T. Meyers, for example, developed an account of moral autonomy in terms of competency that could be exercised either under an ethics of impartial reason or under an ethics of responsibility and care—using studies in social psychology to support her work. Gilligan (1982) distinguished two ‘voices’ in moral reasoning, and two developmental modes, one of them charted by Lawrence Kohlberg’s scale, the other by a scale she developed through interviewing women about their abortion decisions.

Any moral theory presupposes a social theory and a theory of human nature—often implicitly. Kohlberg (1981) was explicit: his view on human rationality relied on Kant, and so did his social theory, traced back to Kant through Piaget and Durkheim. Like the latter, Kohlberg supposed there is an intimate connection between social organization and development of moral reasoning. Differences in the rate of moral development, he said, are due to differences in opportunities for social role taking—and for the highest stage of development, opportunities in the ‘public sphere’ institutions of government, law, and economy are required (historically ‘men’s sphere’), in contrast to the ‘domestic or personal sphere’ of family, etc. Women in general tested

out at the lower stages, and they were diagnosed as confusing moral problems with issues of interpersonal relations.

Kohlberg measured development in terms of subject responses to abstract moral dilemmas. Gilligan developed a different scale for measuring development by asking women about their actual abortion decisions. The basic concepts of her theory were those of care and responsibility in relationships—in contrast to individual rights and justice. She constructed the ethics in the mid-1970s, using interviews with women, but her later work indicates that both men and women use care and responsibility reasoning and develop along her scale.

Gilligan was one of the early theorists of the care and relationship approach, but there were many others, particularly those who focused on mothering as an alternative paradigm to the ‘public man’ that operated in a society of autonomous moral equals. These ‘maternal ethics’ concern care and responsibility among those who are not equal in power or ability—those who are cared for and those who give care, those who are dependent and those on whom they depend. Ruddick’s (1980) work on mothering was widely influential.

To a degree, these ‘care’ ethics accept the dualisms of reason/emotion and public/personal, but they argue for the distinctive importance of women’s experience in developing an ethics for both men and women. However, there are political implications of accepting the dualisms. Tronto (1993) connects the traditional separation of ethics and politics with the division into public and domestic (or personal) spheres, marking it as a boundary that has excluded women from power. She develops a positive alternative of an ‘ethico-politics’ of care.

The ethics of care and responsibility makes a major contribution to philosophical ethics by analyzing moral relations among persons who are not equals. For example, Baier’s (1995) work on trust considers relations among unequals in a way that is applicable in the ‘public sphere’—investigating trust between the governed and the governors. Baier also mined the history of philosophy to find alternatives to the ‘rights and justice’ approach, taking David Hume as a candidate for ‘the women’s theorist’ because he took sentiment and relationship as basic. Other feminists have broadened the ethics to other personal relationships, such as friendship or love, while some have developed an ethics of relationship within lesbian communities (Hoagland 1988, Mohin 1996).

The ‘care and responsibility’ approach has sometimes been labeled ‘feminine,’ rather than ‘feminist,’ because it emphasized women’s traditional place and feminine virtues. Proponents of what is labeled ‘feminist ethics’ claim that any feminist theory must be liberatory, or at least must provide a basis for arguing issues of injustice and domination in the public sphere. For this reason, many feminists have worked to inte-

grate the care and justice approaches (see Held 1995). They insist not only that women's experience be taken seriously, but that patriarchy, racism, class, and other oppressive social structures be taken seriously. Jaggar (1995) has offered an agenda for any ethics that is called feminist: the ethics must underwrite critiques of actions and practices that perpetuate women's subordination; it should prescribe morally justified ways of resisting those practices; and it should envision morally desirable alternative promises of women's emancipation. Feminist ethics should offer a guide to feminist action.

Traditional ethics was criticized because the ideal of a moral agent was modeled on certain men's experience, excluding the experience of women. Both feminine and feminist ethics have themselves been criticized for using an ideal of Western white, middle-class women's experience as the experience of all women—excluding the experiences of many women worldwide. They have been accused of 'essentialism'—assuming women everywhere share certain essential experiences of reproduction and of subordination under patriarchy. This is a very serious accusation, because it questions whether women share a common experience across race and class, in rich and poor nations, in radically different personal and public relations. Because feminists claim to take women's experience seriously, a 'common experience' is needed to make a general 'ethico-politics' that works for all women—whether it is classified as feminine or feminist.

Many critics argue that 'women's experience' is heterogeneous, and it varies along the significant lines of dominance that mark racial, economic, and political oppression. Women of color, lesbians, and working-class women have raised objections—with the result that the varieties of women's experience have been brought into the light. Collins's (1991) work on black feminist thought has offered both criticism and positive alternatives. Narayan (1995) has written on colonialism and the ethics of rights and care, and the place of 'the other' in feminist theory. The early and influential feminist theorist, Flax (1993), now calls for 'an ethics of multiplicity.' Those working in both feminine and feminist ethics agree to the need, but it is by no means clear how the ethics might offer either a general theory of care and responsibility or a general guide to feminist action. As Sandoval (2000) writes, 'US third-world women' have their own strategies of action which suit their own needs in their own situations.

Feminists are committed to respecting the experiences of all women—and avoiding the imperialism of traditional Western approaches. This commitment may be impossible to fulfill, given the multiplicity of ways of life in all the many places of the earth. Globally, women's situations and their moral and political agency differ radically. It is not clear how a secular, Western feminist ethics would apply in a

situation where Islam is important to the lives, identities, and politics of women, for example. In other communities, women have objected to the ethics of women's individual autonomy that is grounded in the ethics. Western feminists have themselves been accused of using Western standards and meanings to define third-world subjects (Ong 1988). 'Postmodern' feminists have argued that feminist ethico-politics presupposes women as the subject of action, and as moral agents, when in fact gender and agency are not givens but are political constructs, enacted in particular times and places (Butler 1990).

Other interesting developments in feminist ethics have been made by the 'ecological feminists.' These feminists offer general criticisms of hierarchies involved in dominance and subordination. They insist that the dualisms of the traditional ethics are incorporated into the social order so as to support dominance in many forms—with domination by men over women and man over nature being central. Ecological feminists share some of their ethics with nonfeminists in the environmental movement, but they insist that understanding patriarchy is fundamental to understanding the hierarchies involved in domination of nature (Warren 1995). One of the earlier books that had an impact was Merchant's *The Death of Nature* (1980), that presents a history of the change in the relation of man and nature from one of respect for a living being (Mother Earth) to a resource to be exploited. Feminist spirituality and 'third world' women have also been important influences in ecological feminism (Ruether 1996).

Within feminist ethics proper, there is criticism of the androcentrism or male centeredness of traditional ethics, but ecological feminists also criticize its anthropocentrism—putting humans at the center of the value universe in a way that permits exploitation of nonhumans. Human interests and human life are the sole moral issues, not only for traditional ethics but for most feminist ethics as well—adding women's interest to men's interest doesn't change that fact. The most interesting of the ecological feminist theories require respect for nonhumans, and they insist on including them within the range of ethics.

Over the years, feminist ethics has come to have a wide influence in other fields. It has made major contributions to biomedical ethics, to business, engineering, and other professional ethics (Leidka 1996, Tong 1997, Whitbeck 1997). The work on care in relationship and the discussions of trust have been particularly important. In this way, feminist ethics has offered practical, new 'guides to action' in some significant arenas of the public world.

See also: Cultural Critique: Anthropological; Cultural Relativism, Anthropology of; Ethics and Values; Feminist Epistemology; Feminist Political Ecology; Feminist Theology; Feminist Theory; Feminist Theory and Women of Color; Feminist Theory;

Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Theory: Sociological History of

Warren K J (ed.) 1996 *Ecological Feminist Philosophies*. Indiana University Press, Bloomington, IN
Whitbeck C 1997 *Understanding Moral Problems in Engineering*. Cambridge University Press, Cambridge, UK

K. Addelson

Bibliography

- Addams J 1907 *Democracy and Social Ethics*. MacMillan, New York
Baier A C 1995 *Moral Prejudices: Essays on Ethics*. Harvard University Press, Cambridge, MA
de Beauvoir S 1948 *The Ethics of Ambiguity* (trans. Frechtman B). Philosophical Library, New York
Collins P H 1991 *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Routledge, New York
Flax J 1993 *Disputed Subjects*. Routledge, New York
Gilligan C 1982 *In a Different Voice: Psychological Theory and Women's Development*. Harvard University Press, Cambridge, MA
Gilman C P 1966 *Women and Economics*. Harper & Row, New York
Gilman C P 1979 *Herland*. Pantheon, New York
Held V (ed.) 1995 *Justice and Care*. Westview, Boulder, CO
Hoagland S 1988 *Lesbian Ethics*. Institute of Lesbian Ethics, Palo Alto, CA
Jaggar A 1995 Caring as a feminist practice of moral reason. In: Held V (ed.) *Justice and Care*. Westview, Boulder, CO
Jakobsen J 1998 *Working Alliances and the Politics of Difference: Diversity and Feminist Ethics*. Indiana University Press, Bloomington, IN
Kohlberg L 1981 *The Philosophy of Moral Development*, 1st edn. Harper & Row, San Francisco, CA
Leidka J 1996 Feminism, morality and competitive reality: A role for an ethic of care? *Business Ethics Quarterly* 6(2): 179–200
Merchant C 1980 *The Death of Nature: Women, Ecology, and the Scientific Revolution*. Harper & Row, San Francisco, CA
Meyers D T 1989 *Self, Society, and Personal Choice*. Columbia University Press, New York
Mohin L (ed.) 1996 *An Intimacy of Equals: Lesbian Feminist Ethics*. Harrington Park Press, New York
Ong A 1988 Colonialism and modernity: Feminist representations of women in non-western societies. *Inscriptions* 3(4): 79–93
Ruddick S 1989 *Maternal Thinking: Towards a Politics of Peace*. Beacon Press, Boston
Ruether R R 1996 *Women Healing Earth: Third World Women on Ecology, Feminism, and Religion*. Orbis Books, New York
Sandoval C 2000 *Methodology of the Oppressed*. University of Minnesota Press, Minnesota
Tong R 1997 *Feminist Approaches to Bioethics*. Westview, Boulder, CO
Tronto J C 1993 *Moral Boundaries: A Political Argument for an Ethics of Care*. Routledge, New York
Walker M U 1998 *Moral Understandings: A Feminist Study in Ethics*. Routledge, New York

Feminist Legal Theory

1. Background

Although it was only as recently as the 1970s that feminist legal theory or feminist jurisprudence were first heard as names for the diversity of gender-focused currents in legal scholarship and argumentation, these strands of contemporary thought have their roots in the 'rights revolution' that took place in the political thought of seventeenth and eighteenth century Europe and the Americas. The foundations of early feminist thought can be traced to explicit efforts to appropriate the emancipatory liberal ideals for women.

The extension of the right to vote was central to the nineteenth century feminist legal agenda. Informed by the social contract theory, early feminists believed that women's political enfranchisement would facilitate their emancipation from historical oppression. The women's rights movement was closely linked to the struggle for the abolition of slavery. In the United States, the campaign for women's vote is associated with the 1848 Seneca Falls convention and the Declaration of Sentiments that emerged from the meeting. It was at Seneca Falls that the former slave, Sojourner Truth, gave her celebrated 'Ain't I a woman?' speech, which questioned the parallel that is often drawn between the legal status of white women and that of slaves in the rhetoric of early nineteenth century feminism.

Despite the significance of voting as the emblematic act of citizenship and of full participation in the *polis*, gender-based discrimination was no less pervasive in other spheres of life. Under the common law doctrine of coverture, a woman lived under the tutelage of her father or other male guardian, until through the sacramental covenant of marriage, her civil legal identity was subsumed by that of her husband under whose cover she derived protection. A woman was typically incapable of making contracts or testamentary dispositions, owning and alienating real property, testifying as a witness in court against her husband, suing or being sued in her own name, or controlling her earnings. She could not even protect her own physical integrity—her husband had the right to chastise her, although only with a switch no thicker than his thumb (see *Gender and the Law*).

The legal fiction of the unity of spouses as well as the rise of separate spheres ideology exempted the private

sphere of the family from state intervention, and rendered women vulnerable to an array of abuses, including rape by their husbands. Beginning in the middle of the nineteenth century, a series of statutes—generically called the Married Women's Property Acts—removed some of the most severe civil disabilities pertaining to women, and acknowledged their personhood for the purposes of the law. Ultimately, struggles for the right to vote, for property rights, for protective labor legislation, for access to education and professional careers, and for related entitlements helped transform the legal landscape. Over the past few decades, feminist strategies have continued to effect change in consciousness and in laws. Legal efforts have focused, *inter alia*, on expanding reproductive freedom, deterring sexual violence, and eliminating sex-based discrimination in employment, education, family, welfare, and related contexts.

2. Definitions

Feminist legal theory is a legacy both of the nineteenth century campaigns for women's emancipation and of the renaissance of feminist activism in the United States and elsewhere since the 1960s. Although there is much debate about its definition, feminist legal theory can be understood as the manifestation in the legal academy of a range of efforts to understand the nature of women's subordination—especially the ways in which it is rationalized and naturalized—and to propose remedies for this subordinate condition. Drawing upon cross-cutting academic and activist traditions, feminist inquiry approaches law both as an ensemble of potential tactics for reform and, more generally, as a site of struggle for a more general political transformation. Most strands of contemporary feminist legal theory have conceptualized law as a historically and socially constructed artifact embedded in particular social relations. Feminist legal theory elucidates the ways in which law facilitates patriarchal dominance and control. Explicating how the doctrines, discourses, institutions, and culture of law perpetuate gender hierarchy, feminist legal theory has provided the impetus for reconceptualizing tightly bonded concepts and premises of legal thought.

Distinctions are often made among various strands of contemporary feminist legal theories. Liberal feminists emphasize the similarities between men and women as rights-bearing autonomous human beings (see e.g., Williams 1982, Kay, 1985, Littleton 1987). Radical feminists argue that society is premised on relations of male domination and female subordination, and that the very epistemological preconditions of law and politics are profoundly gendered (MacKinnon 1982). Postmodern legal feminists show how identities are constituted through multiple structures and discourses that in various ways overlap,

intersect, and contradict each other (Frug 1992, Williams 1991). Cultural feminists emphasize the differences in experience and perspective between men and women, challenging the reification of instrumental capacities of men and denigration of the affective capacities of women which prefer the connection of the individual self to others (see e.g., Menkel-Meadow 1985, West 1987).

The various strands identified above are not an exhaustive inventory of the competing perspectives in feminist legal thoughts, and each has been criticized for its limitation. For example, the cultural feminist perspective is sharply countered by critics who view it as simply an atavistic, throwback to nineteenth century stereotypes, which reified and naturalized the separate-spheres moralizing that entrenched gender asymmetry (Dubois et al. 1985). By the same token, the radical view is criticized for presupposing a unique female point-of-view that does not allow for social differentiation (Cornell 1990). Whatever their drawbacks, it is important to note that these strands are not at all mutually exclusive; their overlap is particularly evident in their commitment to make the world better for women by reconstituting legal practices that have excluded, devalued, or undermined women's concerns.

3. Legal Education

The emergence of the feminist legal theory critique is no doubt linked to the unprecedented entry of women into the legal academy over the past few decades. It was not until 1972 that all bar-accredited law schools in the United States admitted female students. In other countries, the participation of women in legal education has also not been without resistance (see e.g., Dahl 1987, Sachs and Wilson 1978). Feminist theorists have argued that legal education is a microcosm that reflects, reinforces, and reproduces the asymmetric gender relations of the society at large (Obiora 1996). The entire structure of law—its hierarchical organization, its adversarial orientation, and its privileging of formal logic—is considered evidence of its gendered historical and ideological foundations. Studies suggest that the standard law school curriculum marginalizes, distorts, and omits issues of particular concern and interest to women (see e.g., Charlesworth et al. 1991). A recurrent criticism is that the behavioral norms elicited and rewarded by dominant pedagogy are at fundamental variance with socio-cultural norms of female behavior and cause high levels of alienation and role conflict for women in law school (O'Brien and McIntyre 1986).

The idea that gender differences matter is by no means new in the legal terrain. Western philosophical and legal thought is informed and characterized by dichotomous paradigms that may be traced to ancient assumptions of male superiority. Such assumptions found particular expression in systematic sex-role

constraints that allocated public power and authority to men and increasingly relegated the activities of women to the private sphere. One of the theories justifying the exclusion of women from the public sphere posited that there was a discrepancy between womanhood and adulthood; the qualities deemed necessary for adulthood—the capacity for autonomous thinking, rational decision making, and logical action are associated with masculinity, but considered undesirable attributes of the feminine self. The opinion of Justice Bradley in the 1873 United States Supreme Court decision in *Bradwell vs. Illinois* (83 U.S. 130), which affirmed the Illinois bar's refusal to admit a woman, epitomized the operation of the ideological generalization about women's intellectual deficiency and providential ordination for domestic concerns. As he put it, '[t]he paramount destiny and nature of woman is to fulfill the noble and benign offices of wife and mother.'

4. Reform Initiatives

Feminist legal theory has had diverse targets of inquiry, and has sought to answer many difficult, sometimes intractable, questions about law and society. At the methodological level, it has foregrounded gender as an analytic category. Consonant with techniques of consciousness raising which made visible the personal consequences of gender inequities and the commonalities amongst women's life stories, feminist legal method draws and builds upon analyses of women's experiences. Through the critical scrutiny that it has brought to bear on the form and content of orthodox legal theory and practice, feminist methodology has exposed the ambiguities, contradictions, deficiencies, and politics of law. Interrogating the assumptions, arguments, and conclusions which underlie the patterns of constraint that pervade and define the everyday lives of women, it has illuminated how gender hierarchies have been constructed and legitimated in the legal arena.

Contemporary feminist legal theory has provided a framework for the questioning of numerous gendered social practices, such as prostitution and pornography, and made possible the recognition of sexual harassment and various forms of rape as legal categories (see e.g., MacKinnon 1989, Estrich 1986, Smart 1995, Strossen 1993, Shrage 1994). Feminist theorizing has also shown how the private sphere, as exemplified by the family, which had been treated in law as an autonomous realm deserving of protection from state intervention, incorporates myriad forms of gender subordination. Insisting that the 'personal is political,' feminist scholars have exposed the gendered consequences of domestic violence and of the allocation and valuation of 'invisible' domestic labor, such as child care and housekeeping (see e.g., Schneider 1991, Pateman 1988, Olsen 1983). They have also reinvig-

orated debates concerning family law issues, ranging from the consequences for women of traditional marriage, divorce, custody, and child support to issues relating to the control of reproduction (see e.g., Fineman 1991). Exposing how women's unpaid domestic labor and socialization patterns determine the extent to which they can enter the marketplace of paid employment and the terms on which they do so, feminist inquiry has illuminated the connection between domestic arrangements and the structure of the workplace, the regulation of motherhood, the availability of welfare, etc. (Finley 1986, Rose 1992, Radin 1987, Roberts 1991, Abrams 1995, Nedelsky 1990).

Just as beliefs about differences between men and women have rationalized gender inequality, other notions of sexual difference have inspired feminist theorizing. Given the centrality of the principle of equality from the Declaration of Independence to present day political discourse, it is not surprising that feminist legal theory began with an attack on discrimination. Criticisms of restrictive governmental categories were framed in terms of equality of treatment and opportunity. Within this framework, feminists have successfully challenged the assumptions and policies underlying many laws, such as those of female dependency and unsuitability for certain jobs, both in courts and in legislatures. Condemning policies or practices whereby women are denied opportunities simply because of their gender, the liberal feminist enterprise has stressed the similarities between men and women and sought to minimize their differences. Especially during the 1960s and 1970s, feminist jurists relied on a formal equality model which required sameness of treatment for similarly-situated persons and disparate treatments for persons who are not alike as a strategy to trigger the judicial invalidation of many sex-based discriminatory classifications and practices (see e.g., Franke 1995).

To survive judicial scrutiny in the United States, sex-based classifications must bear a substantial relationship to an important governmental purpose. Employing this standard of review, the Supreme Court has insisted on gender-neutral military, social security, welfare, and workers compensation benefits. Under the rule of gender neutrality, in many jurisdictions, the law of custody and divorce has been transformed, giving men (in principle) equal rights to the custody of their children and to spousal support. By and large, the objective of challenging unequal opportunities and the ideology that legitimated them has resulted in a certain undeniable progress. Women have increasingly attained equality in formal rights, but they still experience substantial inequity in social, economic, and political status. The persistence of such disparities is in part attributable to the continuing devaluation of women's capabilities and women's work. The ultimate defeat of the proposed Constitutional Equal Rights Amendment (ERA) in the

USA during the 1980s underscored the extent to which 'traditional' views of gender roles have continued to frame debates. Even in places such as Nigeria, where the mechanism for challenging gender-based discrimination is enshrined in the Constitution, efforts to implement the law remain constrained by deeply embedded customs and tradition.

Faith in formal legal equality obscured its political and practical implications (see *Discrimination*). Equality-based arguments were necessarily value-laden to the extent that they affirmed the status quo in general and accepted of the view of law as rational, objective, neutral, and apolitical. Such equality for women came to mean equality with men—in the United States, with white, middle-class men. In this fashion, the social attributes of men set the norm against which women often appeared deficient. Where there was no man to set the norm (as in pregnancy), it was assumed that there was no discrimination and differential treatment was not condemned. Discourse regarding pregnancy discrimination, for example, ignored women's reproductive health needs. The achievement of formal equality in some settings has aggravated the real world inequality of women in others.

The substantive approach to equality modifies the formalist argument by focusing on the outcome of actual differences between men and women. Proponents of substantive equality demand that rules address disparate outcomes resulting from sex differences. A variant of this model attempts to remedy past discrimination through measures such as affirmative action and comparable worth schemes. An alternative model of substantive equality invokes sex-based biological differences as a rationale for sex-specific protections. A more comprehensive strategy advocates special accommodations to counteract the costs of social norms and practices that disadvantage women (see *Affirmative Action, Sociology of*).

The debate over differential treatment strategies caused a rift in the women's movement. For some, emphasizing women's unique characteristics has exacerbated their economic disadvantage and magnified underlying stereotypes of their frailty. With time, equality has been reconceptualized as an anti-subordination principle to counter discrimination, not just because the classical focus on formal equality implied unfair treatment of specific individuals, but also because it perpetuated gender hierarchy. This perspective, also known as dominance theory, focuses attention on the imbalance of power between men and women (MacKinnon 1989). Increasingly, feminist legal theory has been influenced by a view of the individual as being constituted of multiple, intersecting, and contradictory institutional and ideological forces. This emerging view has been directed not only outward, against conventional legal principles and values, but also inward against feminist legal theory itself (see e.g., Minow 1988).

5. Critique of Identity

Feminist legal theory has been faulted for its failure to condemn the other dimensions of injustice that also mediate gender identity. The preoccupation with differences between men and women has obscured the differences among women and the way that other factors such as race, class, age, disability and sexuality mediate these differences (see *Critical Race Theory; Sexual Orientation and the Law; Law and People with Disabilities*). At the same time that women's choices have been limited, the role of some women—white women of the privileged classes—has been idealized and unreflectively generalized to the situations of other less fortunate women. Consistent with the recognition that the perspective from which one speaks affects the meaning of what is said, some feminists have criticized the liberal mainstream tendency to homogenize and universalize a tacitly normative concept of 'woman'—typically a white, heterosexual, bourgeois, able-bodied, female citizen of an advanced industrial democracy (see e.g., Harris 1990, Crenshaw 1989, Kline 1989, Williams 1991).

Influenced by the post-modern acknowledgment of the plurality of situated realities, these feminists have insisted that the analytic focus of the feminist paradigm take account of social context, structure, and history. Arguing that the category 'woman' has no universal explanatory force, they reject any foundationalist view of identity based on gender distinction alone. They challenge the very notion that there could exist a prototypical woman who can be described in ways that reflect and have meaning for the lives of many different women living in very different geographical, economic, political, and social settings (see e.g., Obiora 1996).

6. Global Dimensions

Gender relations in many non-Western societies are complicated by an intermingling of custom and religion in the law of personal status that governs gender relations. In nineteenth century British debates around numerous gendered practices in colonized regions, the agenda of rescuing non-white women from the barbarity of the culture into which they had the misfortune of being born played an important role in justifying the imperial project. With the intensification of global cosmopolitanism from the 1980s onward, local gender relations have increasingly been required to justify themselves in dialog with alternative modes of being. Gender equity norms symbolize the advance of the human rights agenda.

The most eloquent expression of the prohibition of gender discrimination is codified in the United Nations Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW). The imperatives that it articulates have been consolidated in subsequent documents and conferences such as

the 1993 Vienna Conference on Human Rights. Notwithstanding the force of the appeal of a global norm of gender parity, there is reason to question how well it has traveled across radically different histories, cultures, and socioeconomic structures. Amid the enthusiasm for spreading human rights ideas across the globe, it is easy to lose sight of the fact that abstract formal rights are seldom a panacea. Gender reform initiatives, laudable as they may be, show clearly that meaningful sociolegal acts of transformation must be sensitive to context and culture.

See also: Fatherhood; Feminist Epistemology; Feminist Movements; Feminist Political Theory and Political Science; Feminist Theory; Feminist Theory and Women of Color; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gender and Feminist Studies; Gender and Feminist Studies in Political Science; Gender and the Law; Rationality and Feminist Thought

Bibliography

- Abrams K 1995 Sex wars redux: Agency and coercion in feminist legal theory. *Columbia Law Review* **95**: 304
- Babcock, Freedman, Ross, Williams, Copelon, Rhode, Taub 1996 *Sex Discrimination and the Law: History, Practice and Theory*. Little Brown, Boston
- Bartlett K T 1994 'Gender law'. *Duke Journal of Gender Law and Policy* **1**: 1
- Bartlett K, Harris A 1998 *Gender and Law: Theory, Doctrine, Commentary*. Aspen Law
- Barnett H 1998 *Introduction to Feminist Jurisprudence*. Cavendish, London
- Beauvoir S de 1949 *The Second Sex*. Picador
- Becker, Bowman, Torrey 1994 *Feminist Jurisprudence: Taking Women Seriously*. West
- Charlesworth H, Chinkin C, Wright S 1991 Feminist approaches to international law. *American Journal of International Law* **85**: 613
- Cornell D 1990 Sexual difference, the feminine and equivalency: a critique of MacKinnon's. *Toward a Feminist Theory of the State*, *Yale Law Journal* **100**: 2257
- Crenshaw K 1989 'Demarginalizing the intersection of Race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics'. *The University of Chicago Legal Forum*: 139-67
- Dahl T S 1987 *Women's Law: An Introduction to Feminist Jurisprudence*
- Dubois E C, Dunlap M C, Gilligan C J, MacKinnon C A, Menkel-Meadow C 1985 Feminist discourse, moral values and the law. *Buffalo Law Review* **34**: 11
- Estrich S 1986 *Rape*
- Fineman M 1991 *The Illusion of Equality: The Rhetoric and Reality of Divorce Reform*
- Fineman M, Thomadsen N 1985 *At the Boundaries of Law: Feminism and Legal Theory*. Routledge, London
- Finley L 1986 Transcending the equality theory: A way out of the meternity and workplace debate. *Columbia Law Review* **86**: 1118
- Fiss O What is Feminism? *Arizona State Law Journal* **26**: 413
- Franke K 1995 The central mistake of sex discrimination law: the disaggregation of sex from gender. *U. Pa. L. Rev.* **144**: 1
- Friedan B 1963 *The Feminine Mystique*
- Frug M J 1992 *Postmodern Legal Feminism*
- Gilligan C 1982 *In a Different Voice*. Harvard University Press, Cambridge, MA
- Harris A 1990 Race and Essentialism in feminist legal theory. *Stanford Law Review* **42**: 581
- Kay H H 1985 Equality and difference: the case of pregnancy. *Berkeley Women's Law Journal* **1**: 1
- Kay H, West M 1996 *Sex-Based Discrimination*. West
- Kline M 1989 Race, racism and feminist legal theory. *Harvard Women's Law Journal* **12**: 115
- Larson J 1993 Can feminists use the law to effect social change in the 1990s? *Northwestern Law Review* 1252
- Littleton C 1987 Reconstructing sexual equality. *California Law Review* **75**: 1279
- MacKinnon C 1982 Feminism, Marxism, method and the state: an agenda for theory. *Signs: Journal of Women in Culture and Society* **1**(3): 515-44
- MacKinnon C 1989 *Toward a Feminist Theory of the State*. Harvard University Press, Cambridge, MA
- MacPherson C A B 1962 *The Political Theory of Possessive Individualism: Hobbes to Locke*. Oxford University Press, Oxford, UK
- Menkel-Meadow C 1985 Portia in a different voice: speculations on a lawyering process. *Berkeley Women's Law Journal* **1**: 39
- Mill J S 1869 *The Subjection of Women*. Cambridge
- Minow M 1988 Feminist reason getting it and losing it. *Journal of Legal Education* **38**: 47
- Nedelsky J 1990 *Law Boundaries and the Bounded Self*
- Obiora L A 1996 Neither here nor there: of the female in American legal education. *Law and Social Inquiry* **21**: 355
- Obiora L 1997 Feminism, globalization and culture: after Beijing. *Journal of Global Legal Studies* **4**: 355
- O'Brien M, McIntyre S 1986 Patriarchal hegemony and legal education. *Canadian Journal of Women and the Law* **2**: 69
- Olsen F 1983 The family and the market: a study of ideology and legal reform. *Harvard Law Review* **6**: 9
- Olsen F 1995 *Feminist Legal Theory*. New York University, New York
- Pateman C 1988 *The Sexual Contract*. Stanford University Press, Stanford, NJ
- Radin M 1987 Market-inalienability. *Harvard* **100**: 1849
- Rhode D 1989 *Justice and Gender*. Harvard University Press, Cambridge, MA
- Rhode D 1991 The 'no-problem' problem: feminist challenges and cultural changes. *Yale Law Review* **100**: 1731
- Roberts D 1991 Punishing drug addicts who have babies: women of color, equality and the right of privacy. *Harvard Law Review* **104**: 1419
- Rose C M 1992 Women and property: gaining and losing ground. *Va L. R.* **78**: 421
- Rossi A (ed.) 1989 *Essays on Sex Equality*. University of Chicago, Chicago
- Sachs A, Wilson J 1978 *Sexism and the Law*
- Schneider E The violence of privacy. *Connecticut Law Review* **23**: 973
- Shrage L 1994 *Moral Dilemmas of Feminism: Prostitution, Adultery and Abortion*
- Smart C 1995 *Law, Crime, and Sexuality: Essays in Feminism*. Sage, London

- Smith P 1993 *Feminist Jurisprudence*. Oxford University Press
- Strossen N 1993 A feminist critique of 'the' feminist critique of pornography. *Val. L. Rev.* 79: 1099
- United Nations 1980 *Convention on the Elimination of All Forms of Discrimination Against Women*. U.N. Document A/RES/34/46
- Weisberg D K 1993 *Feminist Legal Theory*. Temple University Press, Philadelphia, PA
- West R 1987 The difference in women's hedonic lives: A phenomenological critique of feminist legal theory. *Wisconsin Women's Law Journal* 3: 71–145
- William J 1991 Dissolving the sameness/difference debate: a critical post-modern path beyond essentialism in feminist and critical race theory. *Duke Law Journal* 296
- Williams P 1991 *Alchemy of Race and Rights: Diary of a Law Professor*
- Williams W 1982 The Equality Crisis: Some Reflections on Culture, Courts and Feminism. *Women's Rights Law Report* 7: 175
- Wollstonecraft M 1967 *Vindication of the Rights of Women*. Norton

L. Obiora and R. Perry

Feminist Movements

Feminist movements are, in the broadest sense, collective efforts to improve the situation of women. Yet much scholarly debate has focused on the question of what movements qualify as 'feminist.' The problem is one of both time and place. 'Feminism' is a term of relatively recent origin, and it was first used in the countries of Europe and the 'neo-Europes,' places such as the United States, Canada, and Australia where European settler colonies flourished. Furthermore, the meaning of feminism has changed over time. Scholars have proposed different definitions of feminism and different ways of categorizing feminist movements. This article considers these various approaches, explores the origins and development of feminist movements in different times and places, and ends with the coming together of women in international feminist movements.

1. Defining Feminism

The term feminism originated as the French word *féminisme* at the end of the nineteenth century and then spread to other languages and parts of the world (Offen 1988). It referred in general to women's emancipation but quickly developed multiple and even conflicting meanings. In the international women's movement in the years before World War I, the term feminism had come to connote, quite specifically, support for legal equality and equal opportunity in the labor market for women and men, particularly opposition to special—what was known as 'protective'—

labor legislation for women (Wikander 1992). Since then a raft of definitions and associations have developed, along with a multitude of qualifying adjectives preceding the word feminism. For some scholars, a feminist movement must specifically embrace that identity, while for others, any activism by women can qualify as part of a feminist movement. The former position means that there can be no feminist movements before the late nineteenth century, while the latter can claim such actions as food riots as feminist. Lerner (1993) locates the origins of the creation of feminist consciousness in medieval Europe, defining feminist consciousness as:

the awareness of women that they belong to a subordinate group; that they have suffered wrongs as a group; that their condition of subordination is not natural, but is societally determined; that they must join with other women to remedy these wrongs; and finally, that they must and can provide an alternate vision of societal organization in which women as well as men will enjoy autonomy and self-determination.

Cott (1989) comes down on the opposite side of the question, arguing that 'not even all activities undertaken by women who claim to have "women's interests" or "women's needs" at heart' are feminist. She proposes, instead, that we distinguish feminist consciousness from 'female consciousness,' that stems from women's shared sense of obligation to preserve and nurture life, and 'communal consciousness,' a sense of solidarity between women and men of any class, ethnic, racial, or other group.

One way scholars have attempted to deal with the complexity of feminism is by distinguishing among different varieties.

2. Categorizing Feminist Movements

The most influential scheme for categorizing feminism and feminist movements has been Offen's (1988) distinction between individualist and relational feminism. Arguing that the anomalous Anglo-American brand of individualist feminism that emphasizes human rights and personal autonomy has been taken as the yardstick to determine what was and was not feminist, Offen asserts that relational feminism, which advocates equity for women in their traditional familial roles as wives and mothers, has been more characteristic throughout time and in different locations. Critics have questioned whether relational feminism is really feminist.

Other schemes for categorizing feminist movements in the past include 'social feminism' and 'hard-core feminism' (Cott 1989) or 'social feminism' and 'equity feminism' (Black 1989). These pairs essentially mirror Offen's distinction. At bottom, the basic divide in these categories is between those who emphasize the sameness of women and men (individualist, hard-core, and equity feminism) and those who emphasize dif-

ference (relational and social, as well as 'domestic,' 'maternal,' and 'cultural' feminism). This ideological distinction has its counterpart in the goals of movements, some of which work to win identical treatment for women and men and some of which advocate different treatment as the right route to equality. In fact, there is no neat fit between an ideological tendency to sameness and the goal of identical treatment, since some feminists have believed that women and men should have identical rights precisely so that women can bring their different (and better) qualities into public life.

For more contemporary movements, scholars have distinguished among liberal, socialist, and radical feminism. Liberal feminism seeks equality for women within the existing social structure, holding that women lack power because as a group they are denied equal opportunity to compete and succeed in the male-dominated economic and political arenas. Thus, liberal feminist movements advocate equal access and equal opportunity. Socialist feminism identifies both class and gender as equally important in creating inequality and calls for a transformation of both capitalism and patriarchy. Radical feminism, in contrast, sees women as a 'sex class and views gender as the primary foundation for the unequal distribution of rewards and privileges. Radical feminist movements, which arose in the industrialized West in the late 1960s and 1970s, like socialist feminist movements seek more than access to existing privileges. Radical feminists call for a fundamental transformation of society based on the superiority of the 'female values' of caring, nurturing, egalitarianism, pacifism, and cooperation.

The categories of feminism have proliferated as research on different kinds of feminist activism has expanded. Scholars have identified Nazi feminists, Islamic feminists, lesbian feminists, and imperial feminists, to name just a few. Increasingly the plural, 'feminisms,' is used to capture the diversity of these varieties of feminism. From a social movement perspective, Rupp and Taylor (1999) have proposed to shift focus away from the ideology of feminist movements to the collective identities that are created and sustained within movements. Such identities exist at three levels: organizational, movement, and solidary. Women develop a sense of themselves as members of discrete organizations, as part of a movement, and, at the broadest level, as women. By exploring the construction of these overlapping identities within specific contexts, it is possible to see how women with very different understandings and strategies and goals might all claim the label 'feminist.'

3. The Origins of Feminist Movements

Social movement scholars point out that everywhere and at every time, women have had sufficient grievances to stimulate the rise of feminist movements. Yet

collective activity on the part of women aimed at improving their own status in society has blossomed primarily during times of generalized social upheaval. In addition, scholars have identified structural conditions that underlie the emergence of feminist protest, at least in the West: changes in patterns of female labor force participation, increased access to education, and shifts in fertility rates and reproductive roles. Chafetz and Dworkin (1986) argue that industrialization and urbanization play a critical role in the emergence of women's movements. In industrializing and urbanizing societies, a large middle class develops, women gain access to education, married women enter the labor force in larger numbers, and the resulting role expansion and conflict results in the development of gender-based and ultimately feminist consciousness. The levels of industrialization and urbanization, the size of the middle class, the extent of women's access to education, and the numbers of married women in the labor force, in combination with the attitude of governments to women's demands, will determine the size of women's movements and whether they qualify as feminist. Despite Chafetz's and Dworkin's attention to movements throughout the globe, their model privileges the western pattern of movement emergence. Feminist movements, scholars have argued, have emerged in other parts of the world in conjunction with nationalist movements, not industrialization and urbanization (Jayawardena 1986).

Although the conditions that lead to the emergence of feminist movements may vary in different parts of the world, there do seem to be broad commonalities in the ways such movements originate and develop. A first stage often begins with the writings of elite women (and men) and demands for greater access to education. In Egypt, e.g., educated, elite women in urban harems began to call for educational opportunity in the context of Islamic modernism at the end of the nineteenth century (Badran 1995). In Argentina, Uruguay, Chile, Brazil, Mexico, and Cuba, women schoolteachers took the lead in the late nineteenth century in calling for expanded education for women (Miller 1991). Chinese women, many educated in Japan, began in the early twentieth century to relate China's weakness to the lack of education for women (Beahan 1975). Typically, women who have gained some educational advantage begin to spread their ideas, often through their own publications. Throughout the Middle East, Asia, and Latin America, women's journals played a central role in introducing feminist ideas.

A next stage was often the formation of organizations devoted to various kinds of social reform. In the United States in the mid-nineteenth century, these included radical projects such as abolitionist groups as well as more societally acceptable activities such as literary societies and church groups. In many cases, demands that originally sprang from women's interest in other causes—that women be allowed to speak in

public or participate in the political process in order to bring about some change in society—eventually came to be voiced as a basic human right. During the national revolution of 1919–22, Egyptian feminists who had earlier formed philanthropic groups worked actively to free Egypt from British domination but experienced disappointment when they found themselves deprived of the vote in the new Egyptian state. This was when the word feminist (in French) first came into use, claimed by the Egyptian Feminist Union, founded in 1923 (Badran 1995). The move from participation in nationalist or anti-imperialist movements to demands for basic political rights and for social freedoms was a common one throughout the colonized and dependent countries.

As feminist movements developed in different places, they often split along ideological, political, religious, or other lines. In countries such as Germany, e.g., the bourgeois women's movement and socialist women's movement had little to do with each other. In England, and to a lesser extent in the United States, the major divide was between moderate and militant groups. The French women's movement splintered along political lines into Republican, Catholic, and socialist organizations. Because feminism as an ideology and movement emerged earlier in the United States and Europe, feminist movements in the rest of the world often give rise to charges of imperialist influence. Despite feminism's origins as an indigenous movement in Egypt, the fact that the upper and middle-class urban women who claimed the label tended to wear western dress and speak French associated feminism with foreign and imperialist forces (Badran 1995). Korean New Women in the 1920s made use of Japanese, British, and Swedish feminist discourses for their own purposes in developing a critique of traditional Korean marriage and sexuality, giving rise to the notion that they were simply dupes of the West (Kwon 1998).

4. *The Two-wave Model of Feminist Movements*

Most scholarship on the development of women's movements conceptualizes two waves of activism, the first in the mid-nineteenth century to the early twentieth century and the second since the 1960s. Evans (1977) explores a variety of feminist movements in Europe, the United States, and Australasia during the first wave, charting the shifting aims of the movements from economic and educational to moral and then political. Evans concludes that the feminist movements of the first wave declined along with the liberalism with which they were so closely associated by the early twentieth century. But there is more continuity in the existence of women's movements than this model suggests. In fact, movements grew outside the Euroamerican arena precisely when scholars think of the decline of feminism in the aftermath of World War I. It may still be useful to

think of two peaks or waves of feminist activism, as long as we remember that not all movements followed the same trajectory and that feminist movements did not die out between the two waves (Rupp and Taylor 1987). Nevertheless, it is clear that feminist movements burst forth with renewed vigor in the 1960s and 1970s in association with civil rights, student and left-wing movements, and movements of national liberation. Often divided on questions of ideology, strategy, goals, and membership, these late twentieth-century movements mobilized around a broad range of issues, from reproductive rights to sexual and economic exploitation to violence against women. Feminist movements both conceptualized women as sharing basic experiences of oppression, asserting that 'sisterhood is global,' and increasingly came to emphasize differences among women on the basis of race, class, ethnicity, nation, and so on. This was true both within national feminist movements and even more so on the international stage.

In the United States, scholars identify two wings of the second-wave women's movement: the women's rights branch, consisting of older professional women organized in bureaucratic national groups working for change within the existing system, and the women's liberation branch, made up of younger women associated with the civil rights and New Left movements who organized small collective groups on the local level and worked for a radical transformation of society (Freeman 1975; Ferree and Hess 1994). The women's rights branch, dominated by the National Organization for Women, founded in 1966, originally worked for legal and public policy reform, while the women's liberation groups pioneered consciousness-raising as a strategy for change and engaged in dramatic 'zap' actions to protest such things as the institution of marriage and beauty pageants and the stock market. Yet over time the distinctions between the two wings dissipated.

The political contexts of different countries in the late twentieth century have been critical in shaping feminist movements and in facilitating or constraining success in moving toward feminist goals. In Western Europe and the United States, feminist movements have had the most success in policymaking where the Left is either very strong, as in Sweden, or very weak, as in the United States (Katzenstein and Mueller 1987). Likewise, in India, the women's movement in Calcutta, which is dominated by the communist party and has a strong left culture, functions more as a political party, while the movement in Bombay, where there is more political competition, takes on more autonomous forms of organizing, including the formation of groups receptive to Western ideas (Ray 1999).

5. *International Feminist Movements*

Since the 1970s, regional and international organizing among women has blossomed, although the origins of

international women's organizations date back to the closing decades of the nineteenth century. In fact, women from the long-established international groups played a vital role in bringing the issue of women's equality to the founding conferences of the United Nations in the aftermath of World War II (Rupp 1997). It was the United Nations-sponsored International Women's Year in 1975, the Decade for Women conferences in 1975, 1980, and 1985, and the Beijing conference of 1995, with their attendant meetings of non-governmental organizations, that nurtured global feminist dialogue (Miles 1996). Although the official conferences and the affiliated forums have proven extremely contentious, struggling over questions of politics, religion, sexuality, and social custom, they have grappled with such basic issues as women's worldwide economic disadvantage, women's lack of reproductive freedom, and different forms of violence against women. Despite variations in feminist movements around the world, international conferences have diffused feminist ideology and practice throughout the globe.

See also: Civil Rights Movement, The; Feminist Theory; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Radical Lesbian; Gay/Lesbian Movements; Right-wing Movements in the United States: Women and Gender; Social Movements, History of: General

Bibliography

- Badran M 1995 *Feminists, Islam, and Nation: Gender and the Making of Modern Egypt*. Princeton University Press, Princeton, NJ
- Beahan C 1975 Feminism and nationalism in the Chinese women's press, 1902–1911. *Modern China* 1: 379–416
- Black N 1989 *Social Feminism*. Cornell University Press, Ithaca, NY
- Chafetz J S, Dworkin A G 1986 *Female Revolt*. Rowman & Allenheld, Totowa, NJ
- Cott N F 1989 What's in a name? The limits of 'social feminism': Or, expanding the vocabulary of women's history. *Journal of American History* 76: 809–29
- Evans R 1977 *The Feminists*. Croom Helm, London
- Ferree M M, Hess B B 1994 *Controversy and Coalition: The New Feminist Movement across Three Decades of Change*, rev. edn. Twayne, New York
- Freeman J 1975 *The Politics of Women's Liberation*. McKay, New York
- Jayawardena K 1986 *Feminism and Nationalism in the Third World*. Zed Books, London
- Kwon I 1998 The new women's movement in 1920s Korea: Rethinking the relationship between imperialism and women. *Gender & History* 10: 381–405
- Lerner G 1993 *The Creation of Feminist Consciousness: From the Middle Ages to Eighteen-seventy*. Oxford University Press, New York
- Miles A R 1996 *Integrative Feminisms: Building Global Visions, 1960s–1990s*. Routledge, New York

- Miller F 1991 *Latin American Women and the Search for Social Justice*. University Press of New England, Hanover, NH
- Offen K 1988 Defining feminism: A comparative historical approach. *Signs* 14: 119–57
- Ray R 1999 *Fields of Protest: Women's Movements in India*. University of Minnesota Press, Minneapolis
- Rupp L J 1997 *Worlds of Women: The Making of an International Women's Movement*. Princeton University Press, Princeton
- Rupp L J, Taylor V 1987 *Survival in the Doldrums: The American Women's Rights Movement, 1945 to the 1960s*. Oxford University Press, New York
- Wikander U 1992 International women's congresses, 1878–1914: The controversy over equality and special labour legislation. In: Edwards M L et al (eds.) *Rethinking Change: Current Swedish Feminist Research*. Humanistisk-samhällsvetenskapliga forskningsrådet, Uppsala, Sweden, pp. 11–36

L. J. Rupp

Feminist Political Ecology

Feminist political ecology offers explanations and analyses of human–environment relations by uniting feminism with political ecology, a subfield prominent within geography and anthropology. Political ecology emphasizes how power relations and politics shape the dynamics of economic development, environmental transformation, and social change across geographic scales of analysis from the local to the global. This approach lies at the interdisciplinary intersection of critical development studies, feminism, and environmental studies. Feminist political ecology analyzes identity, difference, and meanings in their relation to sites of environmental change, degradation, and struggle (Rocheleau et al. 1996). This subfield emphasizes policy relevant natural resource management issues and environmental and social justice concerns in both urban and rural places around the world. Feminist political ecology explains how gender identities and social relations shape, and are shaped by, power relations and social inequality, and how gendered inequalities are linked to questions of liberation struggles, environmental change, and degradation. This article examines the changing dimensions of these cultural, political and ecological intersections through an overview of the emergence of feminist political ecology since the 1980s.

1. Feminism and Environmentalism

Feminism's link with environmental issues and concerns embraces both social and ecological theory and political activism that is rooted in the rise of environmental and feminist activism and intellectual activity of the late 1960s and early 1970s. Ecofeminism, environmental feminism(s), and feminist political ecology all refer to the linkages of people–gender–culture–

environment-nature with varying emphases and frameworks (Merchant 1980, Plumwood 1993, Rocheleau et al. 1996).

Since the 1980s, activists' concerns focused upon nuclear warfare and militarism, toxic waste pollution, and deforestation among others. Ecofeminist work, which spanned academic, popular, and literary texts, emphasized that women were both biologically and socially linked to nature. Women's status and subordination was thus linked to environmental degradation and change so that the oppression and degradation of nature and the environment paralleled the oppression and degradation of women (Merchant 1980). However, when viewed from an international perspective, gendered human perceptions of nature as well as gendered material and sociocultural links to the environment varied over place and time and among cultures. In numerous societies around the world, social scientists have documented and explained how changing cosmologies and geographic histories construct and reproduce key binaries: man/woman; culture/nature; self/other. Structuralism and semiotics influenced much of this initial work. Women are commonly, but not necessarily, associated with nature, and landscape, the fertility of the earth, the waxing and waning of the moon, with passivity, and the hearth (Merchant 1980, Kolodny 1984). These cultural metaphors shaped individual lives as well as public discourse through images of Mother Nature/Mother Earth as contrasted with Spaceship Earth, for example. Early ecofeminist work provided insight as to the ways in which nature and the environment are gendered, and then employed this insight to analyze environmental transformation or celebrate the connection in its spiritual dimension. Ecofeminists defined masculinity and femininity as socially constructed categories which are detrimental to people and the environment and are linked to larger social processes such as colonialism and development. Subsequent critiques pointed to the monolithic and essentialist foundations of earlier ecofeminist work. These critiques argued against a monolithic view of woman that was unmediated by other social categories such as race, ethnicity, gender, age, sexual orientation, and citizenship and pointed to the necessity of site specific empirical case studies (Agarwal 1992).

2. *Critical Ecofeminism*

Critical ecological feminism involves a conscious decision to align human and nature in a move beyond the binary dualism embedded epistemology and ideology in the West. In feminist poststructural critiques, masculinity and femininity as modes of identity are examined in relation to our store of knowledge about the human link to nature and the ways in which nature and environment are gendered and socially constructed. Feminist political ecology rejects the dual-

isms of man/nature and emphasizes multiplicity, diversity, and the complex interconnections among ecologies, societies, and politics in questions of environmental change of both short and long-term duration. This insight is linked to the critical examination of development policies and practices and increasing social inequality. A key question is how identity and difference shape resource control, access, and distribution in multiple historical and ecological contexts initially in Africa, Asia, and Latin America. Research examined how the marginalization and exploitation of local communities and ecologies is linked to larger social and economic networks, institutions, and processes.

3. *Gender, Environment, and Development*

Critiques of mainstream economic development and modernization theory at the international scale revealed how women were invisible in modernization initiatives and within development theory and policy more generally. For example, the majority of women in sub-Saharan Africa are food producers who are responsible for provisioning their households. However, throughout Africa and elsewhere, women were, and still are, unable to gain access to the social, financial, technical, and natural resources that would enable them to provide adequately for their households. As women were incorporated slowly and unevenly into the modernist project through economic development programs aimed at their inclusion, subsequent research revealed how gender bias in development and nesting patriarchal ideologies could exacerbate as well as alleviate their particular vulnerabilities. In addition, this research demonstrated how technology transfer deskilled or did away with working class women's work, and how modernization and development policy and practices in Africa, Asia, and Latin America transformed rural environments and indigenous social and cultural networks, often reproducing or reinforcing gender stereotypes and inequalities. These critiques of mainstream development theory and practice drew from structuralism and Marxian political economy. Studies of the nexus of capitalism, development, populations, and ecologies stressed how oppression and injustice are linked to environmental transformation and change, the industrial geographies of resource extraction and capitalism's creations of surplus (Di Leonardo 1991, Seager 1993).

Examining diverse dimensions of women's work and access to natural resources was critical in revealing the ways in which women were largely disadvantaged by mainstream modernization efforts and policies. Intersections of imperial and indigenous forms of patriarchal power and control could constrict access to land, credit, technology, and education and shape human reproduction. Environmental justice move-

ments and feminist research linked poverty and environmental degradation such as tropical deforestation, irrigation, and pesticide poisoning to gendered relations within the colonialism and to development-induced scarcity. Household vulnerability, changing individual, household, regional and national entitlements, and a lack of democracy are linked to the persistence of poverty and to environmental degradation. Gendered identities and relations crosscut racism, class, age, marital status, ethnicity, and nationality.

Alternatives to mainstream development explore indigenous agroecologies and ecological knowledge, the organization and endurance of social movements linked to environmental concerns, and examine popular struggles and resistance by women and by marginalized and indigenous peoples. The ecofeminist project is one of inclusion and interconnectedness. Feminist political ecology refuses to accept binary dualisms linking gender relations and activities to nature and valuing women and/or nature as inferior or unequal (Plumwood 1993). Thus, ecofeminism and feminist political ecology concern struggles for liberation.

4. Human Ecology Politicized

Political ecology emerged from the subfields of human/cultural ecology prominent in geography, anthropology, and sociology. The next section examines the emergence of political ecology and feminist political ecology from these subfields.

Cultural ecology examines society–environment relations and uses the concepts of adaptation and equilibrium and organic metaphors to explain these relations over time and place. Cultural ecologists focused upon case studies involving extractive economies and peasant societies in Africa, Asia, and Latin America. Critics of cultural ecology pointed out its tendencies to explanations, which were functionalist, teleological and devoid of considerations of politics and power relations. Subsequent critiques questioned whether organic metaphors were appropriate for defining human societies and also indicated that research in ecology had moved beyond conceptualizing environments in terms of equilibrium states and were defining them in terms of dynamism, impermanence, and constant change. Women's access and use of natural resources within a cultural and ecological frame progressed in this period as research focused upon issues of women as resource managers and farmers.

5. Gendering Cultural/Social/Political Ecologies

Since the 1980s, political ecology has examined the political and social dimensions of environmental transformation and change primarily in rural areas of

Africa, Asia, and Latin America. It emerged as a critique of cultural ecology's inattention to power relations and considerations of the politics of resource distribution, access, and transformation. Political ecology combines a broadly defined political economy of resource development and change centering upon the role of social relations and processes to environmental change and degradation, resource distribution, access, and control and the social constructions of nature. In its critique of adaptation and equilibrium, political ecology emphasized agrarian and environmental transformation and change throughout the colonial period up to the present day by examining the ways in which capitalism transformed societies, environments, and the relation between them.

Poverty and environmental degradation are proximate outcomes of the workings of capitalism and world market integration. Peet and Watts (1996) define four contemporary directions in political ecology. First, there are attempts to connect explicitly the dynamics of capitalist growth and environmental outcomes at varying levels and scales. Second, is the integration of power relations and politics into examinations of social and environmental change at varying levels of scale from the global to the micro-politics of the household. Third is the focus examining the roles of ideology and social institutions and organizations. This involves probing the roles of knowledge construction and legitimization as well as state–society relations in terms of the politics of environmental development and change. Fourth, concerns the myriad perceptions of environmental and resource problems and the ways in which these problems are struggled over.

Notably, political ecology is also prominent in its critique of neo-Malthusian explanations of environmental degradation and change. These explanations avoid considerations of social structures and processes and gender identities and power relations within specific historical and geographic contexts as ways to understand the linkage between environment and population growth.

Bryant and Bailey (1997) outline four major research themes in political ecology. The first centers on a particular environmental problem or set of problems such as soil erosion, deforestation, or water. The second studies concepts that are linked to political ecology such as critiques of sustainable development. A third examines political and ecological problems by means of a specific regional case study. The final approach which is explicitly linked to feminism emphasizes individual and group social characteristics such as class, age, ethnicity, gender, and race.

Over roughly the same period, feminist studies in international political economy had taken up considerations of gender and environments, gendered resource access and control, and activism and social movements as they involve the environmental change and degradation ranging across geographic scale.

6. Research Directions in Feminist Political Ecology

Contemporary research directions in feminist political ecology concern how knowledge about society, environments, spaces, and places are gendered. Studies take up questions of local, gendered knowledge about ecologies and agrarian, rural, and urban landscapes. Gendered knowledges refer to particular bodies of knowledge such as that of plant species, agricultural techniques, and food preparation and processing which are frequently the provinces of women or indigenous groups. A second strand of research examines changing environmental rights and responsibilities as they pertain to resource access, distribution, and control. Gendered rights and responsibilities may involve differential access to natural resources such as education, technology, land and trees, and the impacts upon female headed households, the poor, or ethnic minorities. Third, feminist political ecology examines the social struggles that fuse social and environmental concerns and contribute to progressive social change.

Environmental degradation is linked to survival and the quality of life. Survival pivots upon questions of entitlement within rural and urban spaces and is linked in by a diversity of production systems and subject to collective struggles. These microlevel questions and struggles are linked necessarily to broader processes of globalization and ecological dynamics across geographic scales.

See also: Cultural Geography; Ecology, Cultural; Ecology, Deep; Ecology, Political; Feminist Epistemology; Feminist Theory; Ecofeminist and Cultural Feminist; Gender and Environment; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Gender and Place; Gender Ideology: Cross-cultural Aspects; Human–Environment Relationships

Bibliography

- Agarwal B 1992 The gender and environment debate: lessons from India. *Feminist Studies* 128(1): 119–55
- Bryant R L, Bailey S 1997 *Third World Political Ecology*. Routledge, London
- Di Leonardo M 1991 *Gender at the Crossroads of Knowledge*. University of California Press, Berkeley, CA
- Kolodny A 1984 *The Land Before Her*. University of North Carolina Press, Chapel Hill, NC
- Merchant C 1980 *The Death of Nature*. Harper & Row, San Francisco
- Peet R, Watts M 1996 *Liberation Ecologies*. Routledge, London
- Plumwood V 1993 *Feminism and the Mastery of Nature*. Routledge, London
- Rocheleau D, Thomas-Slayter B, Wangari E 1996 *Feminist Political Ecology*. Routledge, London

Seager J 1993 *Earth Follies: Coming to Feminist Terms With the Global Environmental Crisis*. Routledge, New York

L. Jarosz

Feminist Political Theory and Political Science

In popular discourse the term feminism is normally used to mean support for women's rights, while a feminist is someone whose central concern is with the position of women and their struggle for emancipation (Delmar 1986). The definitions are contested by feminist scholars and subject to change. In the 1970s feminism was frequently defined as opposition to the dominance or subjugation of women. By the end of the 1990s it was impossible to define feminism without reference to the concept of gender and it was unusual to refer to feminism in the singular. Feminisms are therefore standpoints. They are also political practices (most often social movements), and sets of ideas and theories. To complete the definition it is necessary to add the concept of gender. Gender is a category that was developed to explore what counts as 'woman' and what counts as 'man' (Squires 1999, p. 54). Thus, one useful definition of feminisms declares them to be oppositions to gender hierarchy, capturing the different situations and interests of different women in the pluralization of 'feminism,' while the social construction of masculinity and femininity is signaled by the term 'gender.'

1. Gender Studies and Feminist Scholarship

Challenging the approaches and preoccupations of mainstream scholarship has proved to be a continuing focus of feminist politics. Originally mainly concerned with 'women's studies,' the attention of feminist scholarship gradually shifted to the study of gender and especially to gender relations and the nature of masculinity and femininity. Feminist writings include critiques of most areas of contemporary scholarship calling into question their sensitivity to issues of gender. The shift had important implications for the social sciences including the study of politics. In the social sciences an important feminist concern is to establish the category of gender as a fundamental social division. The feminist critiques of political science are part of this tendency, addressing the standard themes of research on politics, questioning narrow definitions of the 'political,' and also drawing upon research and thinking about gender in other disciplines.

Feminist political science is therefore interdisciplinary and radical, offering a broad definition of the

field. Feminist political studies take three main forms: (a) feminist political theory which is predicated on a critique of liberalism and challenges the conventions of political theory, (b) feminist political science which is a critique of political institutions and the ways they have been understood, and (c) feminist accounts of gender and politics which combine explorations of gender as a category, as a hierarchy, and as performance in studies of how gender relations are implicated in power. Each area is work in progress, complicated and contested. Similar patterns of development have characterized each of the three kinds of discussion.

2. Feminist Political Theory

The women's liberation movement of the 1960s and 1970s gave a new impetus to feminist politics. This was at first new social movement politics, suspicious of formal hierarchical organizations which were regarded not only as undemocratic but also as sexist. However, feminism was, and is, profoundly political. The early slogan 'the personal is political,' questioned the nature of power and challenged the conventional restriction of politics to the 'public sphere.' Initially, feminist political theory was a matter of unmasking sexist assumptions and the neglect of women by Western Political Theory. Carol Pateman argued that the subjugation of women was a central underpinning of modern political thought and that gender difference made a public sphere possible at the expense of women. She contended that the distinction between state and civil society is only visible if the family is ignored. If civil society is labeled private then family is forgotten in political discussion (Pateman 1988). Hence the addition of women to the concerns of political theory brought the central assumptions of liberalism into question, notably the separation between public and private life.

Recognizing that writing gender differences into political theory required a major reconsideration of political ideas, feminist theorists addressed issues of equality, justice and representation, initiating debates that grew increasingly sophisticated. Originally theoretical arguments were built around binary oppositions or dualisms such as equality/difference, nature/nurture, justice/care, inclusion/exclusion, all thought to parallel male/female dichotomies. Arguments in the equality/difference debates considered whether men and women should be treated the same or whether differences should be taken into account in order to treat both fairly. The justice/care debates addressed a perceived difference between an ethic of justice (in the liberal distributive sense) and an ethic of care (in a maternalist sense), argued to be male and female, respectively.

Representation arguments considered whether women need women to represent them, reasoning that

the identity of the speaker matters to the interpretation of what is said (Phillips 1995, Young 1990). Discussions of justice, equality, and representation tended to converge into arguments about citizenship in which feminists seek to redefine politics in terms that integrate public and private concerns, breaking through the public/private divide (Dietz 1985, Mouffe 1995, Lister 1997). Since the 1980s a tendency to contest the oppositions has become more prevalent as theorists seek to transcend or displace the hierarchies, often by redefining issues and concepts (Squires 1999, pp. 226–33).

3. Feminist Political Science

Although there were very few links with feminist political theory, feminist political science has followed a similar trajectory, developing from the inclusion of information about women to the subject matter of politics to an intellectual concern to establish a gendered discipline. Feminist political science first sought to eliminate the 'bad science' of exaggerating and misconstruing differences between women and men, assuming a male political universe and using data without their original accompanying qualifications. They were troubled by the distortions present in the mainstream of behavioral political science arising, they believed, from an epistemology that separated facts and value and privileged so-called value-free factual accounts. The effect of concern only with putative facts was to allow behavioral political science to factor women out of consideration as political subjects. An overemphasis on measurable political behavior prevented recognition of '...not only the... exclusion of women from what is traditionally political, but also the inclusion of politics in what women have traditionally done' (Nelson 1989, quoted in Carroll and Zerilli 1993). Families, communities, and voluntary groups all went largely unresearched as constitutive of political life. Inevitably, feminist political scientists sought to expand definitions of politics, lending strength to other challenges to overconcern with electoral behavior and elite participation, at the same time as examining differences in men's and women's political behavior. Discussion of the significance of the public/private split and the confinement of women to the private realm were an important challenge to narrow constructions of political power. So too were studies of the crucial part played by the state in regulating sexuality (see Carroll and Zerilli 1993, Lovenduski 1998).

Once it was necessary (and possible) to study the effects of women on previously male-dominated political institutions and organizations, the need for an adequately theorized notion of gender was accepted. As increasing numbers of women appeared in legislatures, councils, and other representative bodies as well as in executive organizations and public

administrations of various kinds, a considerable debate developed about the causes, stability, and results of gender differences in political activity. An especially powerful line of argument for both feminist activists and their opponents was the contention that women should not seek equal presence to men in powerful positions because they would be 'incorporated' and changed by it, they would become unable to represent women because they had been turned by male created institutions into political men. The counterclaim was that once a significant number and diversity of women occupy decision making positions in previously 'masculine' institutions they would be transformed, their gendering altered.

This argument, so important in feminist debate, has echoes in the conventions of the study of political processes (Randall 1991). Once arrangements, practices, and rules are institutionalized, actors learn them, invest in them, and make decisions accordingly. Particular institutions become associated with particular behavior and outputs that may be classified as masculine or feminine (Duerst-Lahti and Kelly 1995). The practices that insulate the positions of typical members and are notoriously difficult to change are gendered.

4. Feminism and Gender Studies

Feminist intellectual concern about women and power generated research in the arts, humanities, and social sciences. In the fields of history, sociology, social psychology, literature, and cultural studies, a substantial body of work about different aspects of gender and power was produced. In addition a significant field of 'new' men's studies examined masculinity as a category both of privilege and oppression (Carver 1996). Analysis of masculinities has also received considerable attention from feminist theorists who also claimed that differences among women were at least as important as differences between women and men, that race, class, age, ethnicity, sexuality, and physical ability were, like sex, constitutive of political identity and therefore of political behavior and power.

5. Feminism, Gender Studies, and Political Science

When feminist political scientists contested narrow definitions of politics, it was because they wanted to expand their research to aspects of 'private' life and to examine the role of government and the state in constructing and enforcing certain kinds of relationships between women and men. As it became more widely understood that both women and men have gender, debates across disciplinary boundaries became a common feature of feminist research on politics.

Once perceived, the effects of gender cannot be left to one side or held constant for the sake of other experiments. As it has come to be defined, gender is embedded in individuals, relationships, institutions, and organizations. The gender debates are part of a continuing process of definition refinement and theory production that enhances understanding of political institutions and processes. A 'gendered' political science looks 'not just at particular theories and assumptions as they have been applied to women, but as they are constructed and apply in any case' (Sapiro 1991). Institutional research traditions in political science require the close description of contexts, norms, processes, and cultures to which a gendered vocabulary may be applied. It follows that the gendered interests present in political arrangements should be acknowledged and analyzed.

In summary, feminist scholars have shown that it is difficult to think of an aspect of politics that is unaffected by gender. However, gender is not a neat and tidy concept to be slotted into unchanging research strategies. Using the concept of gender involves rejection of the definitions of politics implied by the conventions of the public/private split, re-examination of the productions of conventional political science for gender dimensions and, on that foundation, the construction of a political science that is gender informed.

See also: Feminist Movements; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Gender and Feminist Studies in Sociology; Rules in the Legal Process; Women's Suffrage

Bibliography

- Carroll S J, Zerilli M G 1993 Feminist challenges to political science. In: Finifter A W (ed.) *Political Science the State of the Discipline, II*. American Political Science Association, Washington, DC
- Carver T 1996 *Gender is not a Synonym for Women*. L. Reiner, Boulder, CO
- Dahlerup D 1988 From a small to a large minority: Women in Scandinavian politics. *Scandinavian Political Studies* 11: 275–98
- Delmar R 1986 What is feminism? In: Mitchell J, Oakley A (eds.) *What is Feminism?* Blackwell, Oxford, UK, pp. 8–33
- Dietz M 1985 Citizenship with a feminist face: The problem of maternal thinking. *Political Theory* 13(1)
- Duerst-Lahti G, Kelly R M (eds.) 1995 *Gender Power, Leadership and Governance*. University of Michigan Press, Ann Arbor, MI
- Lister R 1997 *Citizenship: Feminist Perspectives*. Macmillan, London
- Lovenduski J 1998 Gendering research in political science. *Annual Review of Political Science* 1: 333–56

- Mouffe C 1995 Democracy, power and the political. In: Benhabib S (ed.) *Democracy and Difference: Contesting the Boundaries of the Political*. Princeton University Press, Princeton, NJ, pp. 245–57
- Pateman C 1988 *The Sexual Contract*. Polity Press, Cambridge, UK
- Phillips A 1995 *The Politics of Presence*. Clarendon Press, Oxford, UK
- Randall V 1991 Feminism and political analysis. *Political Studies* 39: 513–32
- Sapiro V 1991 Gender politics, gendered politics: The state of the field. In: Crotty W (ed.) *Political Science: Looking to the Future*. Northwestern University Press, Evanston, IL, pp. 165–87
- Squires J 1999 *Gender in Political Theory*. Polity Press, Cambridge, UK
- Young I M 1990 *Justice and the Politics of Difference*. Princeton University Press, Princeton, NJ

J. Lovenduski

Feminist Psychotherapy

1. What is Feminist Therapy?

Feminism is defined in Webster's New Word Dictionary (1978) as (a) the principle that women should have political, economic, and social rights equal to those of men, and (b) the movement to win such rights for women. While this is a relatively benign definition, the range of reactions and beliefs attached to the word feminism is vast and often emotionally laden.

These diverse views of feminism have resulted in conflicting definitions in the field of psychotherapy and have led to questions of whether one should more aptly refer to *feminisms* and psychotherapies rather than a unified theory of feminist therapy. In fact in many ways 'feminist therapy' is a moniker for an approach or value system that can be applied to all mental distress. Neither monolithic nor static feminist perspectives can be applied to all therapeutic orientations (e.g., *Interpersonal Psychotherapy*; *Feminist Theory: Psychoanalytic*).

1.1 Core Principles

Although they may vary in their application, at the core of a feminist approach are the following assumptions (Worell and Remer 1992).

- (a) Women's problems cannot be solved in isolation from gender politics which often result in women's lower social status and oppression in most societies.
- (b) Equal status and empowerment are vital not only for women but for all oppressed groups.
- (c) Values enter all human enterprises: neither science nor clinical cases can be value free.

(d) Women's experience and knowledge should be appreciated and studied.

(e) Few individual women can achieve parity alone, individual and collective action is necessary to achieve the social and political change that underpins many stated 'mental' problems.

Brown (1994) captures it well when she says 'what makes practice feminist is not who the clients are or the specific techniques used, but how the therapist thinks about what she does' (p. 2).

2. The Birth of Feminist Therapy

The women's movement in the late 1960s prompted a reconsideration of roles for women and men in western society. Consciousness raising groups provided a forum for women to make the link between their personal experience and the political context in which these experiences were constructed. In spite of their turning away from the accepted expertise of the 'establishment' one of the outcomes of consciousness rising groups was the development of feminist therapists. Through setting up 'women's' groups and 'feminist' supervision, services run by women for women began to develop.

As emergent ideologies challenged traditional views, the field of psychology was also pushed in new directions. This produced revised theories of psychological development and a demand for new explanations to prevent and remediate human problems. In 1975 a new discipline, The Psychology of Women, was established providing a foundation or (at least one academic home) for an applied science dedicated to counseling women.

While the consideration of gender in the prevalence, etiology, diagnosis, and treatment of human problems had long been neglected (*Age, Race, and Gender in Organizations*), there is now increasing awareness that the labeling of illness, the election to seek care, the dialogue between the professional and the 'helper,' and the 'cure' offered are all transactions that are influenced and bound by social organizations, as well as socially constructed themselves.

2.1 Data Supporting the Disproportionate Number of 'Mentally Ill Women'

Studies of clinical and community samples report that a high proportion of individuals with signs of depression, anxiety, panic, and eating disorders are women. Community health and mental health utilization rates are higher for women and a disproportionate share of psychoactive drugs are prescribed to women—many with deleterious or unknown side effects. In addition, it has been suggested that continuous exposure to disempowered positions may be

associated with women's depression (Feree et al. 1999). In addition physical and sexual abuse has been recorded in significant numbers of women seeking mental health care (*Childhood Sexual Abuse and Risk for Adult Psychopathology*).

2.2 Dissatisfaction with Traditional Treatment

Worell and Remer (1992) in a classic text delineate numerous sources of discontent with prevailing treatment options.

(a) Dissatisfaction with traditional theories that depicted male traits as the norm and women as deficient by comparison.

(b) Omission of women from the knowledge base of psychology.

(c) Failure to recognize that psychopathology may reflect power imbalances rather than intrapersonal difficulty.

(d) Uncovering sex bias in psychotherapy.

(e) Disregard by many health professionals for women's self reported experience.

(f) The practice of attributing blame and responsibility to women for their experience of sexual and physical violence.

(g) 'Mother blaming' such that women's interdependence and involvement is pathologized and men's responsibility for their abuses of power remains unchallenged.

(h) Increasing medicalization of women's psychological problems, including misuse of diagnoses and prescriptive drugs.

3. From The Western Individual to The Global Issue: Changing Focus Over Time

The initial focus of feminist therapy was to name the biases indicated above and to create a validated discipline to address them. Consciousness raising groups increased personal insights and improved relationships between women. Although these groups helped women see how the personal could be political, the wider political impact of these groups was limited and the focus was on the individual (Seu and Heenan 1998). The movement was primarily contained to western societies with the experience of the 'other' women of the east infrequently noted.

The next step was to describe and implement clinical practice. Alongside grass roots movements, scholars began to focus on issues specific to women's needs. An important contribution was Gilligan's (1982) *In a Different Voice* in which she argued that women's relationship-oriented concern for care and connectedness is devalued by the dominant androcentric culture's emphasis on logic and reasoning. She called for men and women to embrace a wider range of emotional qualities.

The 1980s and 1990s also saw a proliferation of writing, primarily in the USA, on ways to teach 'empowering behaviors' such as assertiveness training and psychoeducation.

In addition, psychoanalytic scholars in America and Europe explored ways of incorporating feminist object relations theory and its emphasis on connectedness into its framework, while Lacan (1977) provided direction for many professionals working on cultural theory.

Although the field became more adept at questioning, the answers were not forthcoming and indeed the inherent tension between political theory and therapeutic practice has not been resolved. There has, however, been greater attention paid to the global condition of women.

As a result, the third stage within feminist therapy will be the naming of the biases within the profession (Seu and Heenan 1998). Influenced by postmodern deconstructionist thought, there have been continuing challenges to the notion of 'feminist therapy' as a singular idiom. Eurocentric and heterosexual biases in teaching and treatment have been criticised and the esteemed position of gender bias as the 'ultimate' social oppression has been questioned.

Feminist analyses have become increasingly cross-cultural and cross-disciplinary as a greater effort has been made to understand not only the complexity of the social issues 'masquerading' as individual psychopathology but the near universal struggle to achieve and maintain power in a complex, modernizing, global environment.

Studies exploring the social construction of illness—borrowing from sociology, anthropology, and history as well as medicine and psychology have focused as much on gender issues as on power issues. By widening the gender lens to survey hierarchal imbalances more generally, feminist theories have found a way to answer male and female needs in an array of areas from self-definition, personal agency, and respectful care delivery to the negotiation of improved working environments and rejection of limiting fashion norms.

The tension between the individual and the societal approach along with connotations attached to 'feminisms' has led to several interesting questions for the future. For example (a) should a 'true' feminist approach to care even involve treating the individual? or (b) do we even need the term feminist at all if in fact the goal is parity for all humans?

4. Research Questions and Concerns

To amply answer the questions generated by a feminist perspective one must first gain support from societal (governmental, academic, and personal) agencies that manage funding. That questions of female mental health have been omitted in the past reflect that this is

not all easy. However, all who study the mental health of women need not be feminists or utilize feminist research methods, in fact they could be men!

Striegel-Moore (1994) has delineated research guidelines for feminist scholarship in this field: affirmation of a positive view of women (challenges the deficit model of women vs. men), adoption of a 'contextual' approach, utilization of a broad spectrum of research methods (applies qualitative and quantitative methods and establishes a collaborative relationship with research participants), and consideration of the implications of research findings for social change (studies diverse women in natural settings).

The challenge for the future will be to apply these methods across disciplines and across countries to help answer questions such as 'how the outside (environmental influences) get inside (and become individual pathology)?,' and when problems are mostly female specific, Why women? Why now? (at this point in history). It may be that by looking across specific diagnoses to syndromes of female distress that new approaches to treatment and prevention will be found (see *Syndromal Diagnosis versus Dimensional Assessment, Clinical Psychology of*). It may also be possible to better understand societal dilemmas by decoding these personal struggles.

5. Future Directions: The World as Living Laboratory

Changing societies and gender roles may provide *in vivo* opportunities to explore the impact of modifying political and social structures on human adjustment. Cross-disciplinary input will be the key. In addition, access to technology may enhance efforts to distribute knowledge and care to a broader audience (see *Technological Change in a Multicultural Context, Applied Psychology of*). The impact of globalization and technology on women's freedoms and their applications in the therapy environment will be critical areas for study in the twenty-first century.

See also: Feminist Theory: Psychoanalytic; Gender and Feminist Studies in Psychology; Women and Psychiatry

Bibliography

- Brown L S 1994 *Subversive Dialogues: Theory in Feminist Therapy*. Basic Books, New York
- Feree M M, Lorber J, Hess B B 1999 *Revisioning Gender*. Sage Publications, London
- Gilligan C 1982 *In a Different Voice*. Harvard University Press, Cambridge, MA
- Hare-Mustin R T, Marecek J (eds.) 1990 *Making a Difference: Psychology and the Construction of Gender*. Yale University Press, New Haven, CT
- Lacan J 1977 *Ecrits: A Selection*. Tavistock, London

Seu I B, Heenan M C 1998 *Feminism and Psychotherapy: Reflections on Contemporary Theories and Practices*. Sage Publications, London

Striegel-Moore R 1994 A feminist agenda for psychological research on eating disorders. In: Fallon P, Katzman M A, Wooley S (eds.) *Feminist Perspectives on Eating Disorders*. Guilford Press, New York

Worell J, Remer P 1992 *Feminist Perspectives in Therapy—An Empowerment Model for Women*. Wiley, New York

M. A. Katzman

Feminist Theology

Feminist theology presumes what many world religions have for centuries precluded—namely, women's full humanity, the immediacy of women's relation to the divine, and the cultural import of women's reflections on the sacred. Committed to supporting cultural transformation through the politics of spiritual practice (Spretnak 1994), feminist theology has attempted to dismantle dominative power and hierarchy in all their guises—from the private, familial and sexual to the institutional, ecclesiastical and political.

In the interest of transforming power relations, feminist theology works both to deconstruct the androcentric religious imagination, its philosophical underpinnings and its socio-political incarnations, and to construct or to create livable, religious imaginings that shelter the cultural becoming of women and that generate the formation of open, just and egalitarian communities. The term feminist theology, broadly defined, may include women's re/invention of goddess and wiccan traditions, critical reconstruction of the world religions, i.e., Judaism, Christianity, Islam, and Buddhism, as well as women's philosophical, non-theistic, or spiritualist departures from them.

1. Philosophical Background

Philosophically speaking, feminist theology developed in response to the 'deracination' (Daly 1984) of women's lives within the scope of patriarchal, monotheistic religions. Deracination names the effect on women of those religions whose engagement with 'the transcendent' subliminally assumes an abjection of matter—from the Latin 'mater' meaning mother (Kristeva 1982). Inasmuch as Platonic philosophy—which radically split spirit from matter, while aligning 'man' with enspirited mindfulness and 'woman' with (fallen or corrupted) nature—has served as infrastructure for the western monotheistic religions, women have been engendered theologically as devoid of spirit. As a sexual and reproductive body, woman

was subjugated to male authority, e.g., wife to husband, daughter to father, penitent to male priest, etc. Additionally inlaid with Aristotle's hierarchical 'chain of being,' western religions granted men greater ontological status and the privilege, therefore, of ruling over women, animals and earth—all of which, given their assumed passive and receptive contours, awaited man's invigorating agential influence. Built upon the story of man's 'second birth' into the transcendent realms, a birth without resort to woman, these religions idealized male selfhood and generated essentially homosocial cultures (Irigaray 1985). These religions libidinally installed an androcentric socio-political and economic matrix which feminists named 'patriarchy,' because all 'goods' (in terms of values and political rights as well as economics) were circulated through the generations of the fathers and sons. Consequently, women within the purview of these religions have been, as philosopher Daly put it, 'physically, mentally, and emotionally separated from our Original, elemental Race...' (1984). That is, women have been religiously precluded from exercising historical agency on behalf of ourselves as women and on behalf of the culture at large.

Androcentric religions have in various ways instilled in women a sense of self-loathing, existential shame and fear of autonomy. In Christianity, e.g., women have been blamed for all frustrations with the flesh, rationalized as the consequence of Eve's eating of the Tree of Knowledge of Good and Evil (Genesis 3). According to the normative reading of this story, one woman's exercise of free will ontologically fractured and debased the world, bringing death, disease, suffering, and transience. Since the fleshy body has been viewed as the inheritance of the mother; and since woman has been viewed as a derivative of man (Genesis 2, 1 Corinthians 11:8), Christianity disavowed the flesh and imagined salvation as a resurrection of 'the (ideal or imaginal) body' freed from the 'bondage' of development and decay, death, disease, and sexual desire (Romans 8). Not only do women thus serve as social scapegoats for Christianity, women have been the physical buffers between the projects generated by 'man's self-infinetizing spirit' (Ruether 1983) and the material impingements of finitude. That is, by socially domesticating women, men withhold from the purview of culture the needs and delimitations of physical bodies, including 'bodies' of land, water, etc.

2. Theoretical Contours

Inasmuch as the women's movement initially understood itself to be a 'spiritual' movement (Christ and Plaskow 1979, Spretnak 1994), feminist theology could be said to have been inherent to feminism. Women's exclusion from societal power structures was, feminists surmised, fundamentally linked to

women's exclusion from socio-religious symbol systems. As Daly observed wryly, where 'God is male, the male is God' (1985b)—a sentiment seconded by Spretnak, who observed that '(t)he long-standing justification in Western culture for regarding the male as the norm... has been the implicit acknowledgement that men are the same sex as God' (1994). This insight into the culturally constitutive nature of the religious imagination, a critique supported by Feuerbach's definition of religion as projection (*The Essence of Christianity* 1957), led feminist theologians to attempt to interrupt the cultural absolute of male socio-political power by displacing male religious images and by interrupting their sacred storylines. (Consequent to this insight and owing to feminists' own interest in 'changing the divine subject' so as to authorize women's power thereby, women's theological enterprise has sometimes been named 'feminist thealogy'—'thea-' being the feminine of the Greek masculine 'theo-', meaning 'God.' See Goldenberg (1979).) Given the incarnational relation between religious symbol and cultural power relations, among feminist theologians' first tasks was the critique of the masculine images by which religions have referred to the sacred as God, Father, Lord, King, Son, etc., and to humanity as 'sons' or 'men of God.' Not only do such images exclude women as religious and political subjects, these images—feminist theologians were convinced—led to the abuse and violation of women and children. 'Women's experience,' feminist theology's epistemological locus, referred to women's shared experiences of living under the purview of this exclusive male engendering of power. Specifically, such images religiously legitimated the practice of dominative power in relation to all that the system conversely objectified. Hence, women experienced this 'dominion of God' as 'domination'—specifically as battering, rape, poverty, the dislocation and domestication of our lives, etc.

Citing instead texts such as Genesis 1:27 ('So God created humankind in (God's) image,...male and female...'), Galatians 3:28 (In Christ, 'there is no longer male and female...'), and the Pentecostal outpouring of the Spirit upon 'all flesh' (Joel 2, Acts 2), feminists asserted women's unmediated relationship to the sacred, claiming thereby women's theological authority, the validity of women's spiritual experience, and the theological subject matter of women's daily lives. Authorized by Spirit, feminist theologians began a critique of male clerical and ecclesiastical leadership, of the male authorship and interpretation of the canon, of the functional closure of revelation vis-a-vis the sacred canon. Feminist theology not only supported, but was identified actively with the Women-Church movement and the movement for women's ordination—accomplished in Reformed Judaism as well as within many Protestant (e.g., Presbyterian, Lutheran, American Baptist, United Methodist) and Anglican-Episcopal Christian

circles. Yet the problem for women in the religions of 'the sacred book' remained: sacred texts have been written by and for men and have circulated for centuries in closed circuits of male interpreters. Where females did appear in these texts, feminists soon surmised that men had 'used (these) women to think with' (Karen King). Further, the religious sacraments of these communities have often worked to delegitimize even that which was prescribed as 'women's work,' e.g., in the church birthing becomes baptism, feeding becomes eucharist. Faced with the question of whether sacred texts and existing religious institutions had the symbolic resources to promote women's well-being, some women, beginning in the early 1970s and onward, have left and invested themselves in the recovery of goddess and wiccan spiritualities, traditions which have supported human embodiment (Eller 1993). Among those women who (also) stayed related to the existing monotheistic religions, a 'hermeneutic of suspicion' was employed to generate a critical distance from canonical texts. By invoking 'suspicion,' Biblical scholar Schüssler-Fiorenza discouraged women from expecting to discover truth 'in' a text conceived by men and encouraged feminists instead to deduce the 'truth' of scripture by a text's ability to raze systems of domination and exploitation (1984). Another feminist strategy for creating an authority base has been the recovery of the hidden histories of women, both in canonical texts, e.g., Wire's rhetorical positioning of Paul's otherwise anonymous discourses as *The Corinthian Women Prophets* (1990), and in Christian communities throughout the centuries, e.g., Ruether and McLaughlin's *Women of Spirit: Female Leadership in the Jewish and Christian Traditions* (1979). Nevertheless, Tribble's *Texts of Terror* (1984) reminded feminists that, given the patterns of 'inferiority, subordination and abuse of the female in ancient Israel and the early church' (3), e.g., Hagar, Tamar, and the daughter of Jephthah, women's tragic stories do not always yield to a positivist reconstructive resolution. Along with the recuperative labor of feminist Biblical and historical theologians, women have attempted to re-imagine the sacred, a project that has outraged some denominational leaders and led to heresy charges against some women. First, recovering feminine images of the divine from scripture, e.g., Sophia, Wisdom, Spirit, feminists have then applied themselves to rethinking ritual and theology based upon these images. Among academic theologians, McFague's (1987) *Models of God* developed the metaphors of God as Mother, Lover, and Friend so as to promote a change of consciousness—from a consciousness founded upon the theological notion of transcendence as separation from and ascendance over women and nature, towards a consciousness of belonging to the 'evolutionary ecosystem of our cosmos' (9). In *She Who Is* Johnson (1992), proposed a Christian trinitarian model of Spirit–Sophia, Jesus–Sophia and Mother–Sophia—a

constructive project aimed at thinking women into the economy of divine love, while also thinking divine sociality in terms of the lives of women.

As feminist theology developed, feminists, observing that traditional theology reflected only men's spiritual experience, undertook a reconstruction of the central doctrines of their respective communities. Whereas Christian theology had, following Augustine, psychically repressed the corporeal body, senses, emotions, and relations, feminists made the body—women's minds, wills, desires, and relationships—the central site of revelatory encounter and salvific experience. Whereas God had been conceived as absolutely transcendent, feminists have insisted on divine immanence—hence, the transition from invoking the divine as 'God' to 'Spirit.' Similarly, if according to the androcentric imagination creation was said to have transpired by divine fiat or to have come about through artisanal handicraft, i.e., God as 'Creator' or 'Maker,' feminists have spoken of the divine 'birthing' the world, of the world as 'God's Body' (McFague 1993, Jantzen 1984) and of humans as cocreators of cosmic process.

Finally, whereas for Christians Jesus' crucifixion has been seen historically as the occasion for sacrificial grace entering in to right a recalcitrant and wayward world, feminists have questioned positioning a human injustice as an ontological necessity and have gravitated toward two other views. First, feminists, rather than memorializing a unique individual from the historical past, reconstructed christology as the prophetic practice of justice and egalitarian relations within the on-going Jesus-community. Second, some feminists have shifted the christological focus away from crucifixion and towards incarnation. The use of the incarnational metaphor, pervasive in contemporary feminist philosophical theory, resonates with the histories of Christian women who have asserted their subjective agency in and through this analog, e.g., Proulx Boneta and Guglielma of Milan (Ruether and McLaughlin 1979).

3. Context and History of Development

While prefigured by the late nineteenth century feminists (e.g., Stanton's *The Woman's Bible* 1895), 'feminist theology' refers to the theological critique and reconstruction of religious symbols and systems over the past four decades of the women's movement (1960s–1990s). Initial publications by Schüssler-Fiorenza and Daly took place in Europe (1968), where they had worked with other early European feminists, including Catherina Halkes. By the early 1970s, Schüssler-Fiorenza and Daly re/turned to the United States where they joined the scene that included Rosemary Ruether, Letty Russell, and Judith Plaskow. Women's theological writing swelled during the 1970s and 1980s with publication's such as Daly's

Beyond God the Father (1985a) and *Pure Lust* (1984), Ruether's *Sexism and God-Talk* (1983) and Schüssler-Fiorenza's *In Memory of Her* (1983).

As feminist theory became conscious of the contextual situatedness of knowledge, feminist theology was refracted increasingly through the differences of all who had been forced together under the universalization of history's anonymous 'other,' 'woman.' Inasmuch as 'women's experience' was invoked as the epicenter of feminist theology in the 1970s, and inasmuch as this category was—women soon learned—dependent upon contextual, ethnic, class, and race factors, feminist theology in the 1980s became pluralized. Written now as feminist theologies, the term recognized the distinctive liberative agendas of women of color.

African-American women, protesting that feminist calls for 'sisterhood' glossed over black-white racism and the double oppression of race and gender, gleaned the name 'womanist' from Alice Walker's *The Color Purple*. If feminists read with suspicion, womanists—Delores Williams suggested—read with an eye toward survival; she found the epitome of the black woman survivor in Hagar (1993). Womanists' theological reflection often incorporates black women's literature, especially that of Toni Morrison, Alice Walker, and Paule Marshall, womanists having found in this literary canon models more relevant to the goal of improving black women's quality of life.

As differences continued to refract, Hispanic women like Ada Maria Isasi-Diaz wrote *Mujerista Theology* (1996). The indigenous voice of Paula Gunn Allen in *The Sacred Hoop* (1986) interrupted any too facile linkage of North America with the Christian continent. Jewish women resisted the supercessionist imperialism of Christian women (Plaskow 1990). From the beginning, lesbian voices have checked Christianity and Judaism's too facile heterosexist perspectives (e.g., Carter Heyward 1989 *Touching Our Strength*).

As with women of color in North America, women's theology in the densely populated and economically poorer nations has had to address survival needs and has often been undertaken without standing against men, who may consider the gender question secondary to the community's goal of liberation from neo-colonialism. Significant voices from feminist theology in global perspective include Mercy Amba Oduyoye from Ghana, Kwok Pui-Lan of Hong Kong, Chung Hyun-Kyung from South Korea, Mary John Mananzan of the Philippines, Ivone Gebarra from Brazil and Elsa Tamez from Costa Rica. Women from the nations evangelized by the western religions have come more recently to celebrate the hybridity of their religious subjectivity. The recovery of indigenous traditions—some of which, like Korean shamanism, claim stronger spiritual leadership roles for women—have become significant factors in creating out of the received religious tradition a vital and practical re-

ligious syncretism (Chung, Hyun Kyung 1990 *Struggle to be the Sun Again*).

In summary, the pluralization of 'woman' lead to distinct feminist theologies. Questions remain as to whether 'feminist' can serve as a canopy or whether such a move asks women of color to submit to the universalism of white, western experience. Though European theologians, e.g., Dorothee Soelle and Luise Schottroff, have—given greater impetus by their proximity to socialist analysis—included more economic reflections, feminist theology in general has not done as well at critiquing class as race issues. If feminist theology initially was configured by uniting women around victimization, increasingly feminist theologies have attempted to build constructively around women's shared goals. Parallel to this, a significant shift in the energies of feminist theologies can be felt—a move away from the ensconcing of opposition, especially 'resistance,' a position still within the logic of dualism; a move towards a 'nonapocalyptic' feminism, employing a 'methodological ambivalence' to sift through the 'power lines' so as to transform cultures (Catherine Keller 1996 *Apocalypse Now and Then*).

4. Probable Directions

In her 1975 book *New Woman New Earth*, Ruether noted that '(s)exism and ecological destructiveness are related in the symbolic patterns of patriarchal consciousness...' The disclosure of this symbolic connection between women's oppression and the domination of the earth led to the ecofeminist theology of the 1990s. If the collections of ecofeminist essays are indicative (e.g., Adams), ecological concern may convene an even greater diversity of feminist theologians—including women activists from Buddhism, Hinduism, and indigenous religions worldwide. Since women's bodies have served as buffer from life's material impingements for those who believe in the extraterrestrial transcendence of androcentric religions; and, inasmuch as ecosystemic collapses from industrial, agricultural, and multinational industries appear to be increasing, women—who, according to UN statistics, perform two-thirds of the human work hours, receive 10 percent of the pay, and own less than 1 percent of the world's habitable land—will be increasingly adversely impacted (*Progress for Half the World's People* in issues of the 1980s by the UN Association of the US 1980). Insofar as feminist theologies remain orthopraxic, i.e., committed to good practice, rather than correct belief ('orthodoxy'), a burgeoning of ecofeminist theologies might be expected.

Further, a new interest in the sciences could revitalize a certain stream of feminist theology. In the 1980s, the articulation of a relational ontology, undergirded with the process metaphysics of Alfred North

Whitehead, lead to the process-oriented theologies of Rita Nakashima Brock, Catherine Keller, Marjorie Suchocki, and Sheila Devaney, among others. More recently, feminist philosophers of science, e.g., Carolyn Merchant, Donna Haraway, and Vandana Shiva, have been critiquing Christianity's influence upon western scientific paradigms. Given this alliance of interest, feminist theologians might be expected to further the deconstruction of Western Christian mythemes in hopes of interrupting the planetary destructiveness of current and envisioned biotechnologies. At the same time, feminist theologians could be expected—given the fertile overlap between the new physics and theology—to offer new cosmological articulations.

In the 1990s, feminist theology took on a decidedly philosophical hue. To be sure, women like Daphne Hampson (England) and Daly, convinced that Christianity was and would remain inherently patriarchal, dated themselves 'post-Christian.' Consequent to that move, Daly insisted that she wrote philosophy, not theology. However, for those who continue to work within the sacred texts and traditions of Christianity and Judaism, postmodern discourse—heralding the end of totalizing worldviews and the disappearance of the subject—has caused feminist theology—like feminist theory, upon which it is dependent—to rethink its critical terms. While both a totalizing worldview, i.e., 'patriarchy,' and the evocation of female subjectivity had been used to achieve feminism's early liberative goals, this new philosophically reflective stage should not be seen necessarily as a departure from feminist theology's commitment to political transformation. Increasingly, postcolonial theory—especially in the field of Biblical studies—and deconstruction are being used to revitalize or reshape liberative theologies. Sharon Welch modeled the viability of employing post-structuralism on behalf of emancipatory transformation in her 1985 *Communities of Resistance and Solidarity* as did Rebecca Chopp in her 1989 *The Power to Speak*. More recently, the French psychoanalytic and poststructuralist insights of Kristeva and Irigaray have been used as hermeneutical tools to address topics such as eucharist, belief, the incarnation, spirit, etc. (Kim et al. 1993). Nevertheless, whether or not this philosophical tone might—amidst the transformation of other critical terms, e.g., the divine, transcendence, belief, etc.—represent a reconsideration of what counts as theology remains an unarticulated and therefore unanswered question for the field.

See also: Feminist Epistemology; Feminist Ethics; Feminist Theory; Feminist Theory and Women of Color; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Radical Lesbian; Religion and Gender; Religion: Evolution and Development

Bibliography

- Adams C A (ed.) 1993 *Ecofeminism and the Sacred*. Continuum, New York
- Christ C P, Plaskow J (eds.) 1979 *Womanspirit Rising: A Feminist Reader in Religion*. Harper & Row, San Francisco
- Daly M 1984 *Pure Lust: Elemental Feminist Philosophy*. Beacon Press, Boston
- Daly M 1985a *Beyond God the Father: Toward a Philosophy of Women's Liberation*. Beacon Press, Boston
- Daly M 1985b *The Church and the Second Sex*. Beacon Press, Boston
- Eller C 1993 *Living in the Lap of the Goddess: The Feminist Spirituality Movement in America*. Crossroad, New York
- Goldenberg N R 1979 *Changing of the Gods: Feminism and the End of Traditional Religions*. Beacon Press, Boston
- Irigaray L 1985 *Speculum of the Other Woman*. Cornell University Press, Ithaca, NY
- Irigaray L 1993 *Sexes and Genealogies*. Columbia University Press, New York
- Jantzen G 1984 *God's World, God's Body*. Westminster, Philadelphia, PA
- Kim C W M et al. 1993 *Transfigurations: Theology and the French Feminists*. Fortress, Minneapolis, MN
- Kristeva J 1982 *Powers of Horror: An Essay on Abjection*. Columbia University Press, New York
- McFague S 1993 *The Body of God: An Ecological Theology*. Fortress Press, Minneapolis, MN
- Plaskow J 1990 *Standing Again at Sinai: Judaism From a Feminist Perspective*. Harper & Row, San Francisco
- Ruether R R 1983 *Sexism and God-Talk: Toward a Feminist Theology*. Beacon Press, Boston
- Ruether R, McLaughlin E 1979 *Women of Spirit: Female Leadership in the Jewish and Christian Traditions*. Simon and Schuster, New York
- Russell L M (ed.) 1988 *Inheriting Our Mother's Gardens: Feminist Theology in Third World Perspective*, 1st edn. Westminster Press, Philadelphia
- Schüssler-Fiorenza El 1983 *In Memory of Her: A Feminist Theological Reconstruction of Christian Origins*. Crossroad, New York
- Schüssler-Fiorenza El 1984 *Bread Not Stone: The Challenge of Feminist Biblical Interpretation*. Beacon Press, Boston
- Spretnak C (ed.) 1994 *The Politics of Women's Spirituality: Essays by Founding Mothers of the Movement*, 2nd edn. Anchor/Doubleday, New York
- Stanton E C 1895 *The Woman's Bible*. European Publishing Co., New York
- Trible P 1989 *Texts of Terror*. Fortress, Philadelphia, PA
- Williams D S 1993 *Sisters in the Wilderness: The Challenge of Womanist God-Talk*. Orbis Books, Maryknoll, NY
- Wire A C 1990 *The Corinthian Women Prophets*. Fortress, Minneapolis, MN

S. V. Betcher

Feminist Theory

Feminist theory encompasses a range of diverse ideas, all of which originate with the following beliefs: (a) society is patriarchal, structured by and favoring men;

(b) traditional ways of thinking support the subordination of women and the neglect or trivialization of issues particularly affecting women; and (c) this patriarchal order should be overthrown and replaced with a system that stresses equality for both sexes. Feminist theory impacts all institutions—medical, legal, academic, and social, for example—and can be used to illumine all issues affecting humans. The diversity of thought within feminist theory lies in the fact that women across the world differ from each other in many ways—including race, ethnicity, nationality, class, sexual orientation, or educational background, for example—and that these differences result in varying views of feminism and gender equality as expressed in liberal feminist, Marxist-socialist, radical ‘libertarian’ and radical ‘cultural,’ post-modern, and global feminist theories. Yet for all the diversity within feminist theory, there remains the belief that despite women’s many differences, women everywhere share some basic ‘sameness.’

In understanding feminist theory or feminist philosophy, it is necessary to understand theory and why humans develop theories. Philosophy helps us understand what it means to be rational beings, to have knowledge, and to express virtues such as benevolence or justice. Humans develop theories to explain elements of reality, how we interact with one another, and who we are as persons.

However, the history of philosophy and those humans who have developed the theories has centered around male humans. Western philosophy is dominated by the writings and thinking of men and, therefore, it often reflects stereotypes or biases against women, or simply omits any point of view that might be particular to women and their lives. For example, basic concepts of political theory—autonomy, freedom, individualism, power, competition, justice—have been designed to fit the experiences of males in male-dominated cultures (Kourany 1998). Particularly reflecting the daily lives of men who do not birth and raise children or provide the major, mundane caretaking duties for a family, these ideas represented for many centuries the heights to which human nature should aspire. Conversely, traits associated with women for many centuries—emotionality, passivity, cooperation, nurturing, compassion—were thought of as inferior and weak. Clearly, here are philosophical ideas that favor the experiences of men over women, relegating women’s experiences to the inferior realm.

Feminist theory, usually described as beginning in the late eighteenth and early nineteenth centuries along with women’s movements and suffrage movements, reconceives these historical theories and includes the experiences or general characteristics of women in addition to men. Women are, after all, the sex with the capacities of gestation, lactation, and menstruation. There are specific differences in the bodily functions of women and men. Traditionally, women have been associated with these bodily functions, associated

with the ‘lesser’ of many hierarchical oppositions: mind/body, sacred/profane, reason/emotion, active/passive, spiritual/material. This duality—particularly where it privileges so-called ‘masculine’ virtues and values over ‘feminine’ ones—is rejected in feminist theory.

Feminist theory, as a ‘new’ way of thinking, is sometimes difficult to understand because it does not represent one uniform view. The basic premise of feminist theory is that women have been excluded from traditional philosophy; moreover, not only have women been excluded, but also people of ‘different’ economic classes, races, ethnicities, or nationalities. Feminist theory recognizes many flaws in traditional (Western, Judeo-Christian, European, male) philosophy and offers a new way of thinking about issues affecting humans and our world. Additionally, feminist theory operates with the understanding that our current society is patriarchal, structured by and favoring men, and that this patriarchal order should be overthrown and replaced with a system that emphasizes equal rights and just and fair distribution of resources for both sexes. Because this element of feminist theory particularly impacts political and moral theory and the basic functions of our society, there are, as mentioned above, different ‘political’ interpretations or understandings of feminist theory, ranging from so-called liberal perspectives to so-called global perspectives. So great is the diversity of feminist theory that some critics fail to see that no matter how they label their particular way of thinking, virtually all feminist theorists agree that in order to be ‘feminist’ a theory must (a) proceed on the assumption that women and men do not share precisely the same situation in life; (b) offer action guides ‘that will tend to subvert rather than reinforce the present systematic subordination of women’; (c) provide strategies for dealing with issues that arise in private or domestic life; and (d) ‘take the moral experience of all women seriously, though not, of course, uncritically’ (Jagger 1992).

1. Different Feminist Theories

Liberal feminism receives its classic formulation in Mary Wollstonecraft’s *A Vindication of the Rights of Woman* and in John Stuart Mill’s *The Subjection of Women*, two nineteenth-century works. Liberal feminists find the roots of women’s oppression in the cultural constraints that hinder women from competing in the public world, including the worlds of politics, medicine, law, and finance. Because traditional society holds the false belief that women are, by nature, less mentally and physically capable than men, society often excludes women from the arenas mentioned above. Insisting this is discriminatory, liberal feminists demand that society give women the same educational and occupational opportunities that men

have. Liberal feminism is clearly reflected in the work of the National Organization of Women, current laws against sexual harassment or discrimination in the workplace, and past lawsuits against same-sex schools that receive funding from the state, including female tax-payers.

In contrast to liberal feminism, radical feminism holds that the patriarchal system that oppresses women is so flawed it cannot be reformed, but must be completely eliminated. In overthrowing patriarchy's institutions and assumptions, radical feminism singles out for attack social and cultural institutions such as the family and church, as well as traditional ideologies about the sexual and reproductive rights, responsibilities, needs, and preferences of men and women.

Initially, radical feminists—sometimes referred to as 'radical libertarian feminists' (Tong 1998)—aimed to explore what they saw as the pleasures of sex: consensual sex between men and women, lesbian sex, sex with both men and women, autoeroticism, sado-masochistic sex, and even intergenerational sex. They sought to free women from the beliefs that 'good' sex could be experienced only in a 'love relationship,' and that sex for sex's sake was somehow 'bad' or promiscuous. In addition, radical libertarian feminists wished to help women avoid the burdens of human reproduction, going so far as to recommend that natural reproduction be replaced by technological reproduction. They agreed with thinkers such as Shulamith Firestone that no matter how much educational, political, and economic equality women achieve, nothing fundamental will change for women so long as their reproductive roles remain the same. Natural reproduction, said Firestone, is not in women's best interests. Pregnancy is 'barbaric,' and natural childbirth is 'at best necessary and tolerable' and at worst 'like shitting a pumpkin' (Firestone 1970) and had best be replaced by *in vitro* fertilization and *ex utero* gestation. What is more, insisted Firestone, natural reproduction is not in men's or children's interests either, for as soon as men and women are truly free to engage in polymorphous, perverse sex, it will no longer be necessary for men to display only masculine identities and behaviors and for women to display only feminine ones. Freed from their gender roles at the level of biology (i.e., reproduction), said Firestone, women would no longer have to be passive, receptive, and vulnerable, sending out 'signals' to men to dominate, possess, and penetrate them in order to keep the wheels of human procreation spinning. Instead, men and women would be encouraged to become either equally masculine and feminine (mono-androgynous) or as differently masculine and/or feminine as they wished (polyandrogynous). As a result, not only would men and women become androgynous persons, all of culture would become androgynous. Furthermore, in this newly-evolved androgynous culture, the categories of the technological and the aesthetic, together with the categories

of the masculine and the feminine, would disappear through what Firestone termed 'a mutual cancellation—a matter–antimatter explosion, ending with a poof!' (Firestone 1970).

Firestone's 'poof' proved too much not only for the general public and most liberal feminists, but also for some radical feminists, particularly those who were beginning to wonder whether women would really gain true liberty by engaging in permissive sex, refusing to bear children, and becoming androgynous persons, that is, persons who are free to develop both 'female' qualities—nurturance, compassion, tenderness, sensitivity, cooperativeness—and 'male' qualities—aggressiveness, leadership, initiative, competitiveness. This concerned group of radical feminists—sometimes called 'radical cultural feminists' (Tong 1998) or 'essentialists' (Alcoff 1988)—began to caution that sex, usually understood as heterosexual sex, is more dangerous than pleasurable for most women. They urged women to extricate themselves from the institution of so-called compulsory heterosexuality (Rich 1994), which they viewed as 'characterized by an ideology of sexual objectification (men as subjects/masters; women as objects/slaves) that supports male sexual violence against women' (Ferguson 1984). Radical cultural feminists insisted that, as it has been experienced so far, heterosexuality is men's sexuality. It is about men seeking to control women's sexuality: representatively in pornography, and actually through the use of prostitutes or the selective harassment, rape, and physical abuse of women in their power. Only if women can free themselves from sex as men want it, can women discover what sex as women want it might be.

In addition to stressing the dangers of heterosexual relations and the pleasures of lesbian relations, radical cultural feminists emphasized that artificial reproduction would more likely disempower than empower women. They urged women to see artificial insemination by donor, *in vitro* fertilization, and plans for an artificial womb not as new procreative options for women but as means for men to exercise complete control over women's procreative powers—their ability to determine whether the human species will continue or not (Rich 1976, Ruddick 1984).

Finally, radical cultural feminists rejected the idea of androgyny as a desirable goal for feminists, replacing it with proposals to affirm women's essential 'femaleness' (Echols 1983). Far from believing the liberated woman must exhibit both masculine and feminine traits and behaviors (traits or behaviors culturally or stereotypically associated with a particular gender), radical cultural feminists expressed the view that it is better to be female/feminine than it is to be male/masculine. Women should not try to be like men, they said. On the contrary, they should try to be more like women, emphasizing the values and virtues culturally associated with women including 'interdependence, community, connection, sharing,

emotion, body, trust, absence of hierarchy, nature, immanence, process, joy, peace, and life' (Jaggar 1992)), and de-emphasizing the values and virtues culturally associated with men including 'independence, autonomy, intellect, will, wariness, hierarchy, domination, culture, transcendence, product, asceticism, war, and death' (Jaggar 1992). Women's liberation, according to radical cultural feminists, exists in women's conviction that it is better to be women than men.

Somewhat unconvinced by the liberal and radical feminist agendas for women's liberation, Marxist-socialist feminists claim it is impossible for anyone, especially women, to achieve true freedom in a class-based society where the wealth produced by the powerless many ends up in the hands of the powerful few. With Engels, Marxist-socialist feminists insist women's oppression originated in the introduction of private property, an institution that obliterated whatever equality of community humans had previously enjoyed. Private ownership of the means of production by relatively few persons, originally all male, inaugurated a class system whose contemporary manifestations are corporate capitalism and imperialism. Reflection on this state of affairs suggests that capitalism itself, not just the larger social rules that privilege men over women, is a primary cause of women's oppression. If all women—rather than the 'exceptional' ones alone—are ever to be liberated, the two-headed beast of capitalist patriarchy or patriarchal capitalism (take your pick) must be killed. Agreeing with this thought, Marxist-socialist feminist Mitchell (1971) stressed that women's condition is determined by the structures of production (as Marxist-socialists believe), reproduction and sexuality (as radical feminists believe), and the socialization of children (as liberal feminists believe). Women's status and functions in all of these structures must change if women are to be men's equals.

Common to liberal, radical, and Marxist-socialist feminist thought is a desire to view women as somehow the same, even if this sameness falls short of the extreme 'essentialism' of radical-cultural feminism. Women's sisterhood and solidarity are the dominant notes of feminist theory as discussed so far. The problem with such a view of women is, however, that women are not only different from men, but also from each other depending on their class, race, ethnicity, nationality, age, religion, marital status, sexual orientation, and so forth. This point about women's difference, about the tendency of some feminist theorists to confuse 'Women' with a particular kind of woman—perhaps white, Western, middle-class—is the core conviction of multicultural feminism. In the 1990s, the buzzword 'cultural diversity' affected social, legal, educational, and religious institutions in the US, and multicultural feminism emerged as part of this diversity. Multicultural feminists stress that although it is hard to be a woman, it is harder to be a woman of

color than a white woman; a poor woman than a rich woman; a lesbian woman than a heterosexual woman; an old or average-looking woman than a young or beautiful woman. Repeatedly, multicultural feminists explain how the idea of 'sameness'—the seemingly benevolent notion that 'down deep' we are all the same—could counter-intuitively be used as an instrument of oppression rather than liberation. In this connection, multicultural feminists reflect in particular upon the works of feminist theorists such as Spelman (1988) who pleaded with feminists not to make the mistake historian Kenneth Stampp made by asserting 'that innately Negroes are, after all, only white men with black skins, nothing more, nothing less.' Why, asked Spelman, is it that black men are really white men 'down deep' and not that white men are really black men 'down deep'? Could it be that white people still think 'white' is definitely the best way to be—the 'gold standard' for all people? Fearing the presence of some well-intentioned 'Kenneth Stampps' within the circle of feminist theorists—most of whom are white, relatively-privileged, and heterosexual—Spelman warned that 'If, like Stampp, I believe that the woman in every woman is a woman just like me, and if I also assume that there is no difference between being white and being a woman, then seeing another woman "as a woman" will involve seeing her as fundamentally like the woman I am. In other words, the womanness underneath the Black woman's skin is a white woman's, and down inside the Latino woman is an Anglo woman waiting to burst through a cultural shroud' (Spelman 1988). If feminist theorists really value equality, insisted Spelman, then they must take care not to use themselves as the measure of 'womanhood.'

In addition to multicultural feminists, so-called postmodern feminists emphasize women's differences. For example, Julia Kristeva, a postmodern feminist, notes that feminism has neglected multicultural diversity. Oliver, author of an article entitled *Julia Kristeva's Feminist Revolutions*, explains:

Feminists in the United States are struggling with this very issue. The feminist movement has had to realize that it is a white middle class movement that has worked to exclude women whose interests and needs are somehow different. Paradoxically as soon as feminism defines 'woman' it excludes all sorts of women (Oliver 1993).

Inspired by a variety of antifoundationalists, especially the deconstructionist Jacques Derrida and the psychoanalyst Jacques Lacan, postmodern feminists resolve the old philosophical debates about the nature of reality—is it 'one' or 'many'?—decidedly in the direction of 'many-ness.' Along the way postmodern feminists such as Hélène Cixous challenge Western dualistic thinking itself (reason/emotion, mind/body, male/female) as the primary cause of human oppression, seeking to liberate human beings from a rigid conceptual grid that privileges some of society's

members over others (Cixous and Clement 1986). Insisting that not being one of society's favored members (white, rich, heterosexual, male) is an advantage, postmodern feminists stress the advantages of being rejected, unwanted, shunned, frozen out, disadvantaged, unprivileged, abandoned, dislocated, and otherwise marginalized. They claim that precisely because women have not been socially favored, women can lead the revolution to let human differences flourish.

Reconsidering their emphasis on the idea of differences as the 'essential' description of the human condition, some multicultural and postmodern feminists now wonder whether they might have gone too far in their rejection of the idea of sameness, thereby endangering feminism's ability to speak on behalf of the best interests of women in general. After all, if women have nothing more in common than two X chromosomes, it is conceivable that women share few, if any, interests in common. If women do not share common interests, then it makes little, if any, sense to insist, for example, that pornography degrades women or that female genital mutilation harms women or that confining women to the domestic sphere limits their human development. If women are essentially different, if they are more different from each other than similar to each other, then 'issues' affecting women in general do not exist.

Increasingly convinced that, like the idea of sameness, the idea of difference can become destructive, many multicultural and postmodern feminists have cast aside or at least modified their perspectives in order to join hands with so-called global feminists and ecofeminists. Although global feminists are well aware that women in the US are very different from women in Kenya or Thailand, for example, they are nevertheless inclined to believe that the women in these countries share enough common interests to become each other's political allies and moral supports. In a recent article, feminist theorist Okin (1998) reminds 'First World' feminist theorists that feminist activists, particularly those in the 'Third World,' are finding that women across the world do have a lot in common. Okin stresses that at several international meetings, women attending from countries throughout the world acknowledged 'that women (everywhere) are greatly affected by laws and customs having to do with sexuality, marriages, divorce, child custody, and family life as a whole;' that they 'are much more likely to be rendered sexually vulnerable than men and boys;' and that their 'work tends to be valued considerably less highly than...men's work' (Okin 1998). First World women, insists Okin, should work with Third World women to achieve for all women the kind of freedom and well-being some women already have. To do so is not an exercise in cultural imperialism, but a simple response to a call for assistance.

Okin's case for cooperation between First World women and Third World women is a cogent one. But

feminists who are sincere about eliminating all forms of human oppression, beginning with gender oppression, will need to do more than talk about the need to help all women achieve all that is rightfully theirs as women. Privileged feminists must, as many global feminists have suggested, be prepared to give up some of their material luxuries so that women far more oppressed than they can attain all of their necessities (Mies 1993a). Stressing that there is only so much of any one material good to be distributed among the world's population, global feminists claim that feminists must take the lead in living more simply so that life on earth can continue through the next millennium and more. Bluntly put, global feminists maintain that if feminists are really serious about ending all oppression, beginning with gender oppression, they must stop being oppressors themselves.

Of course, it is not easy to stop being an oppressor, particularly if one intends to actively carry out the theories of ecofeminists. According to ecofeminists, human beings are connected not only to each other, but also to the nonhuman world: animal and even vegetative. Unfortunately, we do not always acknowledge our responsibilities to each other, let alone to the nonhuman world. As a result, we deplete the world's natural resources with our machines, pollute the environment with our toxic fumes, and stockpile weapons of total destruction. In so doing, we delude ourselves that we are controlling nature successfully to benefit ourselves. In point of fact, says ecofeminist King (1995), nature is already rebelling, and each day the human self is impoverished as yet another forest is 'detreed' and yet another animal species is murdered to extinction. The only way not to destroy ourselves, insist ecofeminists, is to strengthen our relationships to each other and the nonhuman world by refusing to engage in acts of violence, particularly warfare; refusing to eat animal flesh; and refusing to lead luxurious lifestyles. The question remains, of course, whether all feminists are required to lead lives of self-restraint so that not only gender equity but human equity is achieved; or whether, instead, ecofeminist are feminist 'saints' and 'heroes' that all feminists need not follow.

2. Feminist Theory and Other Disciplines

Many of the feminist theories described above center on political or social sciences and systems. Although most work has been accomplished in the social sciences, and much of the feminist movement in the past 200 years has emphasized these areas, feminist theory impacts every discipline and particularly impacts the increasingly important areas of science, medicine, technology, religion, ethics, education, as well as philosophy.

Since it is impossible to describe in a few pages the ways in which feminist theory has transformed the

entirety of human knowledge, I focus here only on the way it has affected two of philosophy's many subfields: epistemology and ontology. Feminist theory challenges the traditional ontological assumption that the more separate the 'self' is from others, the more autonomous, unique, successful, and superior that self is. Traditionally, philosophers have portrayed 'autonomous man' as a biological male—an independent self geared toward maximizing his self-interest effectively and efficiently. Such a self is always on guard against the ominous 'other' who many interfere with his life projects. Even when the 'other' convinces him to establish cooperative relations in a community, autonomous man proceeds cautiously, always seeing every individual he encounters as a potential threat and competitor for resources. Feminist theory, however, offers a different interpretation of how people come to understand their 'selves.' According to feminist ontologists, including Whitbeck (1989), people come to understand their selves through others, not against them. We are historical creatures, shaped by our relationships with other people, our parents, siblings, friends, and colleagues. Although we differ from one another, we relate to each other reasonably well because we also have much in common. This different ontology would result in a different view of the world—one in which we work together for a common good instead of competing against each other in order to secure the best only for ourselves.

Similarly, feminist theory impacts the field of epistemology. As described by Code (1998), 'the issue is less of doing philosophy "in a feminist voice", or even "in a different voice"—as some feminists propose—than of discerning whose voices have been audible, and whose muffled, in the articulation of prevailing theories ...' As with the impact of feminist theory on other disciplines such as history, literature, or art, we find that the story has been told from the male point of view, using a male's voice, and using the experiences of males. At least in the Western world, philosophers and other theoreticians (mostly men until relatively recently) have claimed that the only way to achieve true knowledge is by distancing one's self from the object of one's inquiry and subjecting it to the powers of dispassionate reason. This style of knowing is in estimation of many feminists 'masculine' or 'male.' Because men tend to conceive of themselves as separate, autonomous selves, they favor ways of thinking that permit them to keep their study subjects at arm's length. Keller (1983, 1985) notes that in science, for example, objectivity has been defined in terms of the separation of the subject (the scientist) from the object of study. She believes it is more than coincidence that the men developing science conceived their methodology in terms of what was emphasized as 'masculine'—separative autonomy. Emotional connections with one's subject matter were seen as contaminating knowledge—as subjective, biased, 'feminine.' Keller insists, however, that emotional

connections sometimes yield useful insights. Indeed, some of our deepest insights come from the ability to empathize and to connect with those whose behavior we study; and because women typically understand themselves as intimately and intricately connected to others, it stands to reason that women might be able to see, or first see dimensions of reality to which many men are initially oblivious (Keller 1983, 1985).

3. *First, Second, and Third Waves of Feminism*

Feminist theory, in sum, is an evolving set of ideas. Feminist theory—or 'academic feminism' (Nussbaum 1999) in America—historically has been 'allied to the practical struggle to achieve justice and equality for women. Feminist theory has been understood by theorists as not just fancy words on paper; theory is connected to proposals for social change' (Nussbaum 1999). As a discipline or philosophical approach to other disciplines, feminist theory may not be separated from the very real struggles of the feminist movement.

At this point, the beginning of the twenty-first century, the feminist movement has accomplished great equality in many areas, including politics, the workplace, and the legal system. However, there is much more work needed for true equality in all spheres of life. Feminist theory—understood here as the scholarship or thinking behind such practical feminist actions such as the reform of rape, domestic violence, or sexual harassment laws, for example—changes according to the position of women in society and the world. Just as the feminist movement is often described as having three waves—we are in the midst of the third wave now—feminist theory has evolved similarly. The first wave of US feminism is rooted in eighteenth- and nineteenth-century liberal thought as articulated by Mary Wollstonecraft, John Stuart Mill, and Harriet Taylor Mill, in particular. As described above, liberal feminism emphasized the equal rational and moral abilities of men and women, and the fact that women must be able to leave the domestic realm and enter the public realm if they are to use these equal abilities. One of the major issues of first-wave feminism in the US was securing for women the right to vote, that is, the right to participate as equal members in our political society.

Second-wave feminism centered on the fact that suffrage would never make women equal to men unless women had the same educational, occupational, and professional opportunities that men had. In addition to participating in the public realm, women must be freed from obligations in the domestic world, including sexual and reproductive obligations. Major issues in second-wave feminism included birth control and abortion, and attacks on the images portraying women as sex-objects or baby machines.

Third-wave feminism is very similar to the global and ecofeminist theories described above. In par-

ticular, third-wave feminism emphasizes that the feminist movement in the US does not represent feminism all over the world. Third-wave feminism is aware of women's differences, including race, class, and national origin; but it is also aware that women across the world have much in common, and that women can work together as equals to achieve the particular forms of freedom and well-being that mean the most to them. Like first- and second-wave feminists, third-wave feminists hunger for justice. However, their vision of justice is exceptionally large and demanding for it is based on the assumption that oppression, including gender oppression, will continue to define the human condition unless feminists and other like-minded theorists and activists take it upon themselves to convince humankind as a whole that in order for each person to have enough, no person can 'have it all.' According to ecofeminist Mies, this thought was best expressed by Kamla Bhasin, an Indian feminist. Bhasin stated that:

sustainable development ... is not compatible with the existing profit-and-growth oriented development paradigm. And this means that the standard of living of the North's (First World) affluent societies cannot be generalized. This was already clear to Mahatma Gandhi sixty years ago, who, when asked by a British journalist whether he would like India to have the same standard of living as Britain, replied: 'To have its standard of living a tiny country like Britain had to exploit half the globe. How many globes will India need to exploit to have the same standard of living?' From an ecological and feminist perspective, moreover, even if there were more globes to be exploited, it is not even desirable that this development paradigm and standard of living was generalized, because it has failed to fulfill its promises of happiness, freedom, dignity and peace, even for those who have profited from it' (see quotation in Mies 1993b).

Emerging between the lines of Bhasin's words is, it would seem, the profile of the ideal third-wave feminist, a feminist for a truly new millennium.

4. Conclusion

Feminist theory includes many ideas, and may be frustrating for someone seeking a one-sentence answer to the question 'what makes a view feminist?' However, this characteristic of feminism is not new or particular to contemporary, third-wave, global feminism or ecofeminism. In 1913, a woman wrote: 'I myself have never been able to find out precisely what feminism is ... I only know that people call me a feminist whenever I express sentiments that differentiate me from a doormat, or a prostitute' (Gibbs 1992). In essence, feminist theory is a set of ideas originating with the belief that women are not subordinate to men or only valuable in relationship to men (servant, caretaker, mother, or prostitute), and that the disciplines, systems, and structures in place in our world today may be changed for the better if infused with a feminist point of view. But it is more than this.

Feminist theory sets an agenda for action, the aim of which is justice and equality for women everywhere and, of course, also for the men and children to whom they are inextricably linked.

See also: Beauvoir, Simone de (1908–86); Critical Race Theory; Family Theory: Feminist–Economist Critique; Feminist Epistemology; Feminist Ethics; Feminist Legal Theory; Feminist Movements; Feminist Political Ecology; Feminist Theology; Feminist Theory and Women of Color; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gay/Lesbian Movements; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in Geography; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Gender, Feminism, and Sexuality in Archaeological Studies; Multicultural Feminism: Cultural Concerns; Political Thought, History of; Queer Theory; Social Movements and Gender; Theory: Sociological

Bibliography

- Alcoff L 1988 Cultural feminism versus poststructuralism: The identity crisis in feminist theory. *Signs: Journal of Women in Culture and Society* 13(3): 406–22
- Cixous H, Clement C 1986 *Sorties*. In: Wing B (trans.) *The Newly Born Woman*. University of Minnesota Press, Minneapolis, MN
- Code L 1998 Voice and voicelessness: A modest proposal? In: Kourany J (ed.) *Philosophy in a Feminist Voice: Critiques and Reconstructions*. Princeton University Press, Princeton, NJ
- Echols A 1983 The new feminism of Yin and Yang. In: Snitow A, Stansell C, Thompson S (eds.) *Powers of Desire: The Politics of Sexuality*. Monthly Review Press, New York
- Engels F 1972 *The Origin of the Family, Private Property and the State*. Pathfinder Press, New York
- Ferguson A 1984 Sex wars: The debate between radical and liberation feminists. *Signs: Journal of Women in Culture and Society* 10(1): 106–35
- Firestone S 1970 *The Dialectic of Sex*. Morrow, New York
- Gibbs N 1992 The war against feminism. *Time* March 9: 50–4
- Harding S, Hintikka M (eds.) 1983 *Discovering Reality: Feminist Perspectives on Epistemology, Metaphysics, Methodology, and Philosophy of Science*. Reidel, Dordrecht, Germany
- Jaggar A 1992 Feminist ethics. In: Becker L, Becker C (eds.) *Encyclopedia of Ethics*. Garland, New York
- Keller E F 1983 *A Feeling for the Organism: The Life and Work of Barbara McClintock*. Freeman, San Francisco, CA
- Keller E F 1985 *Reflections on Gender and Science*. Yale University Press, New Haven, CT
- King Y 1995 Healing the wounds: Feminism, ecology, and nature/culture dualism. In: Tuana N, Tong R (eds.) *Feminism and Philosophy*. Westview Press, Boulder, CO

- Kourany J 1998 Introduction: Philosophy in a feminist voice? In: Kourany J (ed.) *Philosophy in a Feminist Voice: Critiques and Reconstructions*. Princeton University Press, Princeton, NJ
- Mies M 1993a The myths of catching-up development. In: Mies M, Shiva V (eds.) *Ecofeminism*. Zed, London
- Mies M 1993b The need for a new vision: The subsistence perspective. In: Mies M, Shiva V (eds.) *Ecofeminism*. Zed, London
- Mitchell J 1971 *Women's Estate*. Pantheon, New York
- Nussbaum M C 1999 The professor of parody. *The New Republic* February 22: 37–46
- Okin S M 1998 Feminism, women's human rights, and cultural differences. *Hypatia* 13(2): 32–52
- Oliver K 1993 Julia Kristeva's Feminist Revolutions. *Hypatia* 8(3): 94–114
- Rich A 1976 *Of Woman Born*. Norton, New York
- Rich A 1994 Compulsory heterosexuality and lesbian existence. In: Jaggar A (ed.) *Living with Contradictions: Controversies in Feminist Social Ethics*. Westview Press, Boulder, CO
- Ruddick S 1984 Maternal thinking. In: Trebilcock J (ed.) *Mothering: Essays in Feminist Theory*. Rowman and Allanheld, Totowa, NJ
- Spelman E V 1988 *Inessential Woman: Problems of Exclusion in Feminist Thought*. Beacon, Boston
- Tong R 1993 *Feminine and Feminist Ethics*. Wadsworth, Belmont, CA
- Tong R 1998 *Feminist Thought: A More Comprehensive Introduction*, 2nd edn. Westview Press, Boulder, CO
- Whitbeck C 1989 A different reality: Feminist ontology. In: Garry A, Pearsall M (eds.) *Women, Knowledge, and Reality: Explorations in Feminist Philosophy*. Unwin Hyman, Boston

R. Tong

Feminist Theory and Women of Color

Feminist theory has been at the forefront of new directions in political, social, and cultural theory. These developments are inherently indebted to the internal critique within feminism made by 'women of color' who have been pivotal in raising questions of 'difference' around such social axes as class, racism, ethnicity, sexuality, and the problematic of global inequities. The point is that the experience of 'being a woman' may vary greatly according to class background, color, sexual orientation, religion, and so on. The critique consists of debates that emerged through political contestation both within and outside the women's movements, drawing attention to the centrality of slavery, colonialism, imperialism, and postcoloniality in understanding contemporary gender relations and global predicaments.

1. Emergence of 'Women of Color' as a Political Subject

The terms 'women of color' and 'white women' throw into relief the political nature of discourses and practices in and through which these terms emerged as

political subjects and became conceptual components of social, political, and cultural theory. They show how seemingly neutral words such as 'color' may assume specific meanings in different contexts so that, as in this instance, the color of 'whiteness' is placed into question. The political subject of 'women of color' decenters 'whiteness' as a modality of power constituted in varying forms in different times and places through historical processes of slavery, colonialism, and imperialism. It challenges the ways in which white European-descent woman came to be seen as superior to other categories of women. Although black, brown, and other categories of 'nonwhite women' have been engaged in feminist practices for centuries, the concept of 'women of color' first emerged in post-World War II Northern America and Western Europe, especially Britain. Hence, the primary focus here is upon the anglophone debate in these locations.

Political tensions in Western feminism surrounding the interrelationship between 'race' and other factors such as class and gender in the USA date back to the antislavery campaigns. During the decade of the 1830s, for example, American women became increasingly active in the abolitionist movement where they learnt to champion their own right to engage in political work and where their experience of relative marginalisation compelled them to form separate women's antislavery societies. The first female antislavery society was formed in 1832 by black women in Salem, Massachusetts, followed by similar societies established by white women in other locations. Paradoxically, when the motion for women's suffrage was first introduced amid immense controversy at the Seneca Falls Anti-Slavery Convention of 1848, black women were conspicuous by their absence. This omission was incredible, not least because black women already had brought into the arena of public debate issues such as women's education which the Convention was only just beginning to address. In May of 1866, when women decided to establish an Equal Rights Association incorporating struggles for Black emancipation and woman's suffrage into a single campaign, a number of eminent speakers, including the white feminist Elizabeth Cady Stanton, argued against it. In class terms too, this was predominantly a middle-class women's movement. While a few individual black women, as for instance Sojourner Truth, were able to participate in specific events, nonetheless, the movement overall did not take on board the contradictory relationship between racism and class, or the question of sexuality in any significant way. Nor was the plight of Native American peoples, or non-European immigrants an identifiable feature of these debates and activities (Davis 1981).

As a form of self-conscious protest, feminism in Britain arose in the seventeenth century. Here, as in the USA, the early women's rights movement and later the Suffragists failed to give sufficient priority to the needs of working-class women or the issue of 'race.'

This is not to deny that there were some women as, for example, Annie Besant who was active on the anti-colonial front as well as on gender issues, and Sylvia Pankhurst who placed considerable emphasis on the conditions of the working class. Nevertheless, the effects of racism and class inequality did not become a major feminist concern at this stage. Such amnesia about issues of 'race' cannot be attributed to the lack of presence of Irish, Jewish, African, and South Asian descent people—the primary target of racisms of the period—because they were far from absent in Britain from the sixteenth to the early twentieth century. Moreover, the first usage of the term 'race' in Britain also dates back to the sixteenth century and the subsequent history of the discourse of 'race' is interlinked with slavery, colonialism, and the Holocaust of Jews, gays, gypsies, and others. Racism therefore can be said to be one of the key factors in the formation of Western societies.

The last five decades of the twentieth century witnessed the emergence of a variety of social movements: anticolonial movements for independence, Civil Rights and the Black Power movement, the Peace movement, student protest, and the Workers' movements. The Women's Liberation Movement was formed during the late 1960s and the 1970s as a part of this intellectual and political ferment. Women of color were active in all these groups. Yet, a significant number of early publications by 'Second Wave' white feminists seemed to display a certain disregard of racism as an *internal* feature of Western patriarchal relations. This neglect drew critical scrutiny from women of color. One of the first critiques was launched by the Combahee River Collective, a black, lesbian feminist organization from Boston, USA. In 1977, they produced a document that demonstrates the complexity of theorizing women's subordination when analyzing experiences based on simultaneous intersection of diverse forms of injustice: 'We know that there is such a thing as racial-sexual oppression which is neither solely racial nor solely sexual, e.g., the history of rape of Black women by white men as a weapon of political repression (Smith 1983, p. 275). The Collective saw itself 'actively committed to struggling against racial, sexual, heterosexual, and class oppression' and advocates 'the development of integrated analysis and practice based upon the fact that the major systems of oppression are interlocking (Smith 1983, p. 272). Pointing to the global dimensions of gender, the text speaks of the impact of 'political-economic systems of capitalism and imperialism' and emphasizes the question of institutional racism as well as 'racism in the white women's movement.' It repudiates those political positions which accept that the end justifies the means: '... as feminists we do not want to mess over people in the name of politics' (Smith 1983, p. 282). In taking a stance against biological determinism, this feminist discourse articulates a certain non-essentialism even as its notion of identity

politics would seem to exemplify what Spivak-Chakravorty (1999) was later to call 'strategic essentialism.' The idea of 'simultaneously interlocking oppressions' is one of the most productive insights of this period.

While the composition of 'women of color' in terms of ethnicity may vary from place to place, the generalized category itself remains wide-ranging and global. According to the anthology *Home Girls*, for example, 'the terms Third World Women and women of color... designate Native American, Asian American, Latina and Afro-American women in the USA and the indigenous peoples of Third World countries wherever they may live' (Smith 1983, p. liv). Taken together, women of color comprise the great majority of women of the world and a very high proportion live in conditions of material disadvantage. Hence, the centrality of a concern with social inequality in women of color feminisms.

Writings by women of color have a history of being treated as 'personal testaments' rather than 'theory.' The question of what counts as theory and whose intellectual work gets treated as being theoretical has been the subject of much debate (cf. Collins 1990). A related point concerns the *theoretical* subject of 'women of color.' In her analysis of the book *The Bridge Called My Back*—a seminal American text of the early 1980s—Norma Alarcon argues that the theoretical subject of *Bridge* is a figure of multiplicity, representing consciousness as a 'site of multiple voicings' seen 'not as necessarily originating with the subject but as discourses that traverse consciousness and which the subject must struggle with constantly.' It is associated with subjectivity that is deeply marked by 'psychic and material violence' and it demands a thorough 'reconfiguration of feminist theory' (Alarcon in Anzaldúa 1990, pp. 359–65). In other words, the theoretical subject of 'women of color,' in its very constitution, de-centers the Eurocentric discourse of the rationalist, masculinist subject of modernist philosophy. It challenges essentialist readings of skin tone or physical appearance as inherent 'difference' and disrupts any notion of 'woman' as a unitary category.

2. *Black British Feminism*

A similar yet different process occurred in 1970s Britain when 'women of color' came to be figured as 'black.' This was a consequence of coalition politics among African, Caribbean, and South Asian descent groups who borrowed the Black Power vocabulary but re-signified 'black' to embrace all 'nonwhite' people. The concept of 'black' was designed to substitute the colonial term 'colored people' and interrogate 'chromatism'—the hierarchy based on differences of skin tone—that could potentially threaten unity between these groups. Black British feminism was forged

through the work of local women's organizations active around such issues as wages and conditions of work, immigration law, fascist violence, reproductive rights, and domestic violence. By 1978, local groups had combined to form a national body called the Organisation of Women of Asian and African Descent (OWAAD). This network held annual conferences, published a newsletter, and served as an active conduit for information, intellectual conversations, and political mobilization. The ensuing dialog entailed sustained analysis of racism, class, and gender with much debate as to the best means of confronting their outcomes while remaining alive to cultural specificities:

Our group organises on the basis of Afro-Asian unity, and although that principle is maintained, we don't deal with it by avoiding the problems this might present, but by having on-going discussions... Obviously, we have to take into account our cultural differences, and that has affected the way we are able to organise (OWAAD cited in Mirza 1997, p. 43).

Like its 'women of color' counterpart in the USA, emerging black feminism in Britain came to prefigure later theories of 'difference.' The internal conflicts within OWAAD, especially around homophobia, proved salutary so that, even as British 'black feminism' assumed a distinctive political identity separate from 'white feminism,' engaging the latter in heated debate, it did not become oblivious to the contradictions of its internal heterogeneity. The careful reflection and analysis that went into the deliberations resulted in a very significant corpus of scholarship.

Hazel Carby's essay *White woman listen* is symptomatic of conflicts and tensions that characterized feminist debate during the 1980s. Carby develops a trenchant critique of key feminist concepts—'patriarchy,' 'the family,' and 'reproduction'—drawing out their interconnection with racism and colonial history. Taking issue with the Euro-centric approach of certain white feminist discourses, she stresses that racism is not simply about exclusion but equally about the manner of inclusion itself. As Chandra Talpade Mohanty has argued, Western feminisms during this period tended to reproduce the very categories through which the 'West' constructs and represents itself as superior to others. In 1984, a group of British feminists of color collectively edited a special issue of the journal *Feminist Review* (Issue 17). This text, together with the edited collection *Charting the Journey* (Grewal et al. 1988), represents intellectual intervention through work that was the outcome of collective thinking. Some of the main points addressed in these publications may be summarized as follows:

(a) As a consequence of capitalist colonial relations, labor migrants from former colonies are constituted as a *racialized class fraction* in the metropolitan society.

(b) The institutionalization of racism and gender inequality within global sociopolitical and cultural relations means that 'women of color' may face some

similar injustices all over the world, but they will also share experiences with the metropolitan white working class.

(c) Women of color in Britain are subjected to 'double patriarchy': a generalized, all pervasive, racialized form of British patriarchy; and, specific configurations of patriarchal norms current among particular groups of people of color.

(d) Lesbians of color have to contend with: heterosexism in British society; homophobia among people of color; and, eurocentricity and racism among white lesbians and gay men. These aspects must be addressed as part of feminist theoretical and political practice.

3. Postmodernism and the Theory of Difference

Feminist theories of the 1970s and 1980 were informed by conceptual repertoires drawn largely from 'modernist' theoretical and philosophical traditions of European Enlightenment such as liberalism and Marxism. The 'postmodernist' critique of these perspectives, including their claims to universal applicability, had precursors within anticolonial, antiracist, and feminist critical practice. Postmodern theoretical approaches found sporadic expression in Anglophone feminist works from the late 1970s. But, during the 1990s they became quite a significant influence, in particular their poststructuralist variant. Some of the main points emerging from the encounter between women of color feminism and postmodern/post-structuralist theories are that:

(a) It is no longer tenable to conceptualize 'white feminism' and 'feminism of color' as if they are mutually exclusive entities, each carrying some unchanging transhistorical 'essence.'

(b) Contrary to analysis where process is reified and understood as personified in the bodies of individuals, these two distinctive yet overlapping sets of feminisms are now viewed as representing historically contingent relationships, contesting fields of discourses, or 'reading strategies.'

(c) The concept of 'agency' has been reconfigured through poststructuralist appropriations of psychoanalysis in order to take account of psychic and emotional life. The concept of subjectivity now covers both conscious agency and unconscious psychic dynamics, and it represents a site of multiplicity where effects of racism and ethnicity intersect with other structural and experiential dimensions.

(d) Reassessment of the notion of experiential authenticity has highlighted the limitations of 'identity politics.' Experience is now understood primarily as a 'signifying practice' referring to the process of making sense of the world symbolically and narratively.

(e) Poststructuralist insistence that meaning is relational; that identity and subjectivity are not 'products' but on-going processes; that power can be

both coercive and productive; that subordination can occur through modes of inclusion no less than exclusion means that poststructural paradigm has much in common with theoretical interventions made by women of color over the years.

Overall, critical but productive conversations with theories of poststructuralism have resulted in new theoretical frameworks for refashioning the analysis of 'difference' (cf. Hooks 1990, Grewal and Kaplan 1994).

4. Postcoloniality, Transnationality, and Diasporic Space

One outstanding feature of the work of feminists of color during the 1990s is a concern with the potential of combining strengths of modern theory with post-modern insights. This interest in interdisciplinary 'creolized theory' would seem to stem from the following preoccupations:

- (a) an interest in analyzing the effects of racism without recourse to the discourse of 'race';
- (b) the need for theorizing ethnicity and cultural difference outside the imperatives of racism;
- (c) the importance of understanding political identities as contingent forms of resistance;
- (d) the necessity of distinguishing between political identity as conscious agency and identity conceptualised in terms of subjectivity;
- (e) the importance of analyzing locality and globality as relational categories;
- (f) sustaining the previous emphasis on colonial histories and 'material conditions' at the same time as analyzing cultural issues; and
- (g) the need to attend to the problematic of the category 'mixed race.'

This work has taken several forms. Some developments, especially in the field of literary studies, are known as 'postcolonial' theory. When used in the sense of a reading strategy, postcolonial theory emphasizes the point that both the 'metropolis' and the 'colony' were altered deeply by the colonial process and that these articulating histories have a mutually constitutive role in the present. Postcolonial feminist studies foreground 'Othering' processes underlying colonial and postcolonial discourses of gender. Frequently, such work uses poststructuralist frameworks, especially Derridean deconstruction and Foucauldian discourse analysis. Some scholars have attempted to combine poststructuralist approaches with neo-Marxist or psychoanalytic theories. Other feminists of color have transformed 'border theory' (Anzaldua 1990, Mani 1998, Spivak 1999)

A related development is associated with valorization of the term diaspora. The concept of diaspora increasingly is used in analyzing the mobility of peoples, commodities, capital, and cultures in the context of globalization and transnationalism. The

concept is designed to analyze configurations of power—both productive and coercive—in 'local' and 'global' encounters in specific spaces and historical moments. Brah (1996) addresses the concept of diaspora alongside that of Anzaldua's theorization of 'border' and the widely debated feminist concept of 'politics of home.' The intersection of these three terms is understood through the concept of 'diaspora space' which covers the entanglements of genealogies of dispersal with those of 'staying put' so that the 'native' becomes as much a 'diasporian' as the diasporic subject becomes the native. The term 'homing desire' is used to think through the question of home and belonging; and, both power and time are viewed as multidimensional processes. Importantly, diaspora space embraces the intersection of 'difference' in its variable forms, placing emphasis upon emotional and psychic dynamics as much as socioeconomic, political, and cultural differences. Difference is thus conceptualized as social relation, experience, subjectivity, and identity. Home and belonging is also a theme of emerging literature on 'mixed-race' identities which interrogates the concept of 'race' as an essentialist discourse with racist effects (Tizzard and Phoenix 1993, Zack 1993, Ifekwunigwe 1999). Accordingly, the idea that you are mixed-race if you have black and white parents is jettisoned. Instead the analytical focus is upon varying and variable subjectivities, identities, and the specific meanings attached to 'differences.'

Theoretical perspectives developed by women of color are central features of innovations in feminist theory. They herald a new phase in feminist alliances and coalitions. They promise new ways of formulating democratic futures (Alexander and Mohanty-Talpade 1997).

See also: Critical Race Theory; Cultural Assimilation; Cultural Critique: Anthropological; Cultural Relativism, Anthropology of; Evolutionism, Including Social Darwinism; Feminist Epistemology; Feminist Ethics; Feminist Theology; Feminist Theory; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Multicultural Feminism: Cultural Concerns; Race and Gender Intersections; Theory: Sociological

Bibliography

- Alexander J, Mohanty-Talpade C (eds.) 1997 *Feminist Genealogies, Colonial legacies, Democratic Futures*. Routledge, London and New York

- Anzaldúa G (ed.) 1990 *Making Face, Making Soul*. Aunt Lute Foundation Books, San Francisco
- Bhavnani K K (ed.) 2000 *A Reader*. Oxford University Press, Oxford, UK
- Brah A 1996 *Cartographies of Diaspora, Contesting Identities*. Routledge, London and New York
- Collins P H 1990 *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Unwin Hyman, Boston
- Davis A Y 1981 *Women, Race and Class*. Women's Press, London
- Essed P 1991 *Understanding Everyday Racism: An Interdisciplinary Theory*. Sage, Newbury Park, CA
- Grewal I, Kaplan C (eds.) 1994 *Scattered Hegemonies: Postmodernity and Transnational Feminist Practices*. University of Minnesota Press, Minneapolis, MN
- Grewal S et al. (eds.) 1988 *Charting the Journey: Writings by Black and Third World Women*. Sheba, London
- Hooks B 1990 *Yearning: Race, Gender and Cultural Politics*. South End Press, Boston
- Ifekwunigwe J O 1999 *Scattered Belongings: Cultural Paradoxes of 'Race', Nation, and Gender*. Routledge, London and New York
- Jarrett-Macauley D (ed.) 1996 *Reconstructing Womanhood, Reconstructing Feminism: Writings on Black Women*. Routledge, London
- Lewis G 2000 *Race, Gender, Social Welfare: Encounters in a Postcolonial Society*. Polity, Cambridge, MA
- Mani L 1998 *Contentious Traditions: The Debate on Sati in Colonial India*. University of California Press, Berkeley
- Mirza H S (ed.) 1997 *Black British Feminism: A Reader*. Routledge, London and New York
- Smith B (ed.) 1983 *Home Girls: A Black Feminist Anthology*. Kitchen Table Press, New York
- Spivak G Chakravorty 1999 *A Critique of Postcolonial Reason: Towards History of the Vanishing Present*. Harvard University Press, Cambridge, MA
- Sudbury J 1998 *Other Kinds of Dreams: Black Women's Organizations and the Politics of Transformation*. Routledge, New York and London
- Tizzard B, Pheonix A 1993 *Black, White or Mixed Race?: Race and Racism in the Lives of Young People of Mixed Parentage*. Routledge, London and New York
- Trinh T Min-Ha 1989 *Woman, Native, Other: Writing Postcoloniality and Feminism*. Indiana University Press, Bloomington and Minneapolis
- Zack N 1993 *Race and Mixed Race*. Temple University Press, Philadelphia.

A. Brah

Feminist Theory: Ecofeminist and Cultural Feminist

Just as there are a variety of feminisms and feminist theories, there are a variety of ecofeminisms and ecofeminist theories. What distinguishes them as 'ecofeminist' is a twofold commitment to analyzing and undoing historical connections between the unjustified dominations of women (and other human

subordinated groups) and nonhuman nature. This article explores the varieties of ecofeminist theories. It reveals similarities and differences among ecofeminist theories, and suggests directions for future ecofeminist theory and research.

1. The Terms 'Ecofeminist' and 'Cultural Feminist'

The term '*ecofeminisme*' was coined in 1974 by French feminist Francoise d'Eaubonne (1974) to call attention to women's potential to bring about an ecological revolution. Ecofeminism has emerged around the globe as a powerful political movement and theory. Despite important differences among ecofeminist theories, they share a common concern for making visible historical associations of 'women' with 'nature,' and, at least in Western societies, the attendant devaluation of both as inferior to men and male-gender identified 'culture' (Merchant 1980).

2. Three Types of Ecofeminist Positions

Historically, the first strand of ecofeminism was *cultural feminism* (sometimes referred to as 'radical' or 'spiritual' feminism). Developed in the early 1970s, cultural feminism reclaims women-nature connections as liberating and empowering expressions of women's capabilities to care for nature. Some cultural ecofeminists argue that women's reproductive capacities provide a biological tie with nature, making women 'closer to nature' than men. Other cultural ecofeminists claim that 'women's closeness to nature' is embedded in deep social and psychological structures, making women's ways of knowing and moral reasoning better suited to solving environmental problems. Some argue for a resurrection of pre-patriarchal religions and spiritual practices (e.g., Goddess worship or pagan rituals) which honored women's bodies and procreative powers. But all cultural ecofeminists agree that women have a special and useful relationship to the physical world (bodies, nature) which can help end the unjustified dominations of both women and nature.

Cultural ecofeminism has been criticized by *social ecofeminists* for making essentialist, universalist, and ahistorical claims about both women and nature. Social ecofeminists claim that there is no essential (biological, natural, innate) nature of 'women' (*contra* cultural ecofeminism's essentialism), no homogenous 'women's experience' or 'women's way of knowing' (*contra* cultural ecofeminism's universalism), and no ahistorical concept of 'women' (*contra* cultural ecofeminism's ahistoricism). Rather, women's identities are socially constructed, historically fashioned, and materially reinforced through the interplay of a diversity of race/ethnicity, class, sexual orientation, age,

ability, marital status, and geographic factors ('social constructivism').

A middle ground is suggested by *materialist* (or *socialist*) *ecofeminism*. Socialist/materialist ecofeminism claims that women–nature connections are both socially constructed *and* biologically predisposed: women's biology (bodies, reproductive capacities) has played a key role historically in the oppression of women and the identification of 'women' with 'nature,' but women's biology is *not* destiny. Rather, it is the social, material, and political relationships between women and nature which are important.

What all three strands of ecofeminism add to feminism is the insistence on the relevance of ecology and environmentalism to feminism. While contemporary feminism has made progress in making visible the interconnections among various forms of human oppression (e.g., sexism, racism, classism, heterosexism, ageism, ableism, ethnocentrism), all ecofeminists insist that feminist analyses be extended to show how the unjustified domination of nonhuman nature (or 'naturism') is connected to sexism and other human 'isms of domination.'

3. The Intellectual Context of Ecofeminist Theory and Practice

Ecofeminism has always been a grassroots political movement motivated by pressing pragmatic concerns: women's and environmental health; deforestation, desertification, and soil erosion; the treatment of animals; science, development and technology; peace, and antinuclear activism. The Chipko Movement in India, Mothers of East Los Angeles, and the Women's Pentagon Actions are examples of the sort of political organizing actions characteristic of ecofeminism (Warren 1997b).

Ecofeminist theory grew out of such grassroots activism. It provides perspectives about gender from a combined environmental *and* feminist perspective, by showing ways human relationships to both 'nature' and 'culture' are gendered, raced, and classed, and how the dominations of women and nature are embedded in patriarchal (and other) structures and practices. Ecofeminist theory offers ecological notions of the self which challenge dominant Western models of rationality, knowledge, and ethics, and which establish connections between biodiversity and cultural diversity.

Ecofeminist scholarship is interdisciplinary. In philosophy, it focuses on conceptual issues by challenging patriarchal world views characterized by value-hierarchical and dualistic thinking, coupled with a 'logic of domination' which 'justifies' relationships of dominance and subordination (Warren 1996, p. 21). Ecofeminist philosophy challenges the 'culture/nature' dualism, as well as related 'reason/emotion,' 'mind/body,' 'mental/physical,' 'objective/subjective,' and

'human/nonhuman nature' dualisms. That is because these value dualisms historically have functioned to sanction as superior the male-identified realm of culture (mind, reason, objectivity) and as inferior the female-identified realm of nature (the physical, the body, emotion, subjectivity). Ecofeminist theorists argue that such hierarchical value dualisms are either false or conceptually incoherent. They incorrectly presuppose a clear separation of culture from nature (mind from body, reason from emotion, human from nonhuman), an ecologically unsound view of humans as discontinuous with nature, and an untenable ethical view which justifies the control, conquest, and mastery of the nonhuman environment in order to serve human needs.

Ecofeminists have offered different historical accounts of the connections between patriarchy, androcentric (male-centered) thinking, and the destruction of nonhuman nature. For example, environmental historian Carolyn Merchant argues in *The Death of Nature* that the change in gender-laden metaphors about nature in the history of science contributed to 'the death of nature.' Merchant argues that the classical Greek metaphor of nature as 'organic,' kindly, benevolent, nurturing female metaphor was replaced in the sixteenth and seventeenth centuries by the metaphor of nature as 'mechanistic,' disorderly, chaotic, instrumental, passive, and inert. While both metaphors feminize 'nature,' only the mechanistic metaphor permitted the removal of ethical constraints against the 'rape of Mother nature,' thereby permitting the 'mining of her womb' for the sole benefit of man.

By contrast, Riane Eisler argues that patriarchal patterns of domination began about 4500 BC E with the invasion of Indo-European societies by nomadic tribes from Eurasia. Ecofeminist philosopher Val Plumwood locates the historical roots of naturism in patriarchal dualisms which originated in classical Greek philosophy and the rationalist tradition (whereby rationality is taken to be both the hallmark of humanness and what elevates humans over nonhuman animals and nature).

Ecofeminist theorists often focus on ways sexist-naturist language maintain, sanction, and perpetuates harmful value hierarchical dualisms. Women are described in animal (nature) terms as pets, dogs, cats, pigs, cows, sows, foxes, chicks, serpents, bitches, beavers, mares, old mares, old bats, old hens, mother hens, pussies, pussycats, cheetahs, bird-brains, hare-brains, elephants, and whales. Animalizing or naturalizing women *in a patriarchal culture* where animals are seen as inferior to humans (men) thereby reinforces and authorizes women's inferior status. Similarly, language which feminizes nature *in a patriarchal culture* where women are seen as subordinate or inferior reinforces and authorizes the domination of nature. 'Mother Nature' is raped, mastered, conquered, mined, controlled, poked, prodded, and pried into; her secrets are penetrated and her womb is put

into service by the 'man of science.' Virgin timber is felled, cut down; fertile (not virile) soil is tilled and land that lies fallow is barren (not impotent) and useless, like a woman who cannot conceive a child. The exploitation of nature and animals is justified by language which is feminizing them; the exploitation of women is justified by language which is naturalizing women (Warren 1997b).

Ecofeminist theory has spawned a new area called 'ecofeminist literary criticism' (Murphy 1995). Focusing on contemporary writing by women (particularly women of color), ecofeminist literary criticism uses ecofeminist insights to develop a body of literature which honors biotic and human (especially gender) differences, while challenging mainstream notions of humans, culture, nature, and wilderness in the traditional Western literary canon (including traditional nature writing by men). Paula Gunn Allen, Margaret Atwood, Rachel Bagby, Annie Dillard, Sally Gearhart, Susan Griffin, Linda Hogan, Ursula Le Guin, Marge Piercy, Leslie Marmon Silko, Alice Walker, and Terry Tempest Williams are just a few of the authors who explore these connections through literature.

One of the first ecofeminist literary texts was cultural ecofeminist Susan Griffin's *Woman and Nature* (1978). In the prologue to her book Griffin writes,

He says that woman speaks with nature ... He says he is not part of this world, that he was set on this world as a stranger. He sets himself apart from woman and nature.

And so it is Goldilocks who goes to the home of the three bears, Little Red Riding Hood who converses with the wolf, Dorothy who befriends a lion, Snow White who talks to the birds, Cinderella with mice as her allies, the Mermaid who is half fish, Thumbelina courted by a mole. (And when we hear in the Navajo chant of the mountain that a grown man sits and smokes with bears and follows directions given to him by squirrels, we are surprised. We had thought only little girls spoke with animals.)

We are the bird's eggs. Bird's eggs, flowers, butterflies, rabbits, cows, sheep; we are caterpillars; we are leaves of ivy and sprigs of wallflower. We are women. We rise from the wave. We are gazelle and doe, elephant and whale, lilies and roses and peach, we are air, we are flame, we are oyster and pearl, we are girls. We are women and nature. And he says he cannot hear us speak.

But we hear (Griffin 1978, p. 1).

Griffin's writing has been controversial because of its essentialist and universalist underpinnings. Still, it is one of the earliest and most powerful pieces of ecofeminist literature to reveal the deeply-seated symbolic and philosophical associations between women and nature in the Western intellectual tradition.

An ecofeminist concern about language, images, stories, rituals, and symbol-systems about 'women' and 'nature' is at the heart of cultural ecofeminism's spiritual claims. Cultural ecofeminists defend earth-based, feminist spiritual symbols (such as Gaia and the Goddess) as a necessary part of ecofeminism. Charlene Spretnak writes:

Ecofeminism grew out of radical, or cultural, feminism (rather than from liberal feminist or socialist feminism) ... In the mid-1970s many radical/cultural feminists experienced the exhilarating discovery, through historic and archaeological sources, of a religion that honored the female and seemed to have as its "good book" nature itself ... We would not have been interested in "Yahweh in a skirt," a distant, detached, domineering godhead who happened to be female. What was cosmologically wholesome and healing was the discovery of the Divine as immanent in and around us (Spretnak 1990, p. 5).

Similarly, cultural ecofeminist Starhawk defends a nonpatriarchal, earth-based spirituality as crucial to ecofeminist politics:

Earth-based spirituality is rooted in three basic concepts that I call immanence, interconnection, and community. The first—immanence—names our primary understanding that the Earth is alive, part of a living cosmos. What that means is that spirit, sacred, Goddess, God—whatever you want to call it—is not found outside the world somewhere—it's in the world: it *is* the world, and it is us ... When we understand that the Earth itself embodies spirit and that the cosmos is alive, then we also understand that everything is interconnected ... The kind of spirituality and the kinds of politics we're called upon to practice are rooted in community ... The goal is the creation of a community that becomes a place in which we can be empowered and in which we can be connected to the Earth and take action together to heal the Earth (Starhawk 1990, pp. 73–4).

On the other hand, ecofeminist theologian Elizabeth Dodson Gray was among the first ecofeminists to examine the roles that sexist-naturist imagery plays in the patriarchal heritage of the Judeo-Christian tradition *without* entirely rejecting those traditions. In her book, *Green Paradise Lost*, Gray claims that a destructive hierarchy of beings is at the heart of biblical accounts of creation:

In this biblical view of the nature of things woman comes after and also below man. Woman was created (according to this chronologically earliest account of the creation of the world in Gen. 2) out of man's body (rather than from a woman's body as happens naturally) ... Then come children, so derivative that they are not even in the Creation story ... Then come animals, who do not have the unique human spirit at all ... Thus animals are below. Further down are plants, which do not even move about. Below them is the ground of nature itself—the hills and mountains, streams and valleys—which is the bottom of everything just as the heavens, the moon and the stars are close to God at the top of everything (Gray 1979, p. 5).

In this hierarchical order of being, the higher up one goes, the closer one is to all that is spiritual and superior:

What is clearly articulated here is a hierarchical order *of being* in which the lower orders—whether female or child or animal or plant—can be treated, mistreated, violated, sold, sacrificed or killed at the convenience of the higher states of spiritual beings found in males and in God ... Nature ... is

obviously a prize candidate for the most ruthless “mastering” of all (Gray 1979, pp. 5–6).

Much of ecofeminist theory involves ethical considerations. Minimally, ecofeminist ethics attempts to develop theories and practices concerning humans, nonhuman animals, and the natural environment which are not male-biased and which provide a guide to action in the prefeminist present. Typically, ecofeminist ethics reconceives humans as both ecological selves and moral agents, and reframes human–nonhuman nature relationships in ways which centralize values of care, appropriate reciprocity, kinship, integrity, and well-being—values typically underplayed in traditional Western ethics.

Ecofeminism also challenges the conception and practice of mainstream Western science and epistemology. As philosopher Lori Gruen states,

Following the arguments made by feminist philosophers of science, Marxists, cultural critics, and others, ecofeminists believe that facts are theory-laden, theories are value-laden, and values are molded by historical and philosophical ideologies, social norms, and individual processes of categorization (Gruen 1994, p. 124).

According to Gruen, ecofeminist theory always grows out of and examines the social contexts in which moral and epistemic claims are generated. Recognition of the interdependence of science and society, facts and values, reason and emotion ‘is the first step towards any legitimate knowledge’ (Gruen 1994, p. 134).

Biologist Donna Haraway concurs. Her arguments for ‘situated knowledges’ and ‘cyborg feminism’ challenge scientific assumptions of impartiality, objectivity, fact, and truth by revealing their embeddedness in structures of power and dominance. Haraway locates the promise of ecofeminism in disrupting the nature–culture dualism by seeing nature as both active, unpredictable, unstable, ungendered subject with agency (not simply some ‘resource to be mapped and appropriated’) and also as constructed artifact (Haraway 1991, p. 199).

Ecofeminism’s deep roots in political activism have led to the emergence of ecofeminist political theory based on coalition-building, direct action, and new conceptions of democracy. Political scientist Noël Sturgeon argues for a conception of ecofeminism as an ‘oppositional political discourse and set of practices imbedded in particular historical, material, and political contexts’ (Sturgeon 1997, p. 3). Sturgeon characterizes ecofeminism as a social movement through which change is produced by numerous kinds of ‘action,’

including that of the deployment of symbolic resources, shifts in identity construction, and the production of both popular and scholarly knowledge—as well as direct action, civil disobedience, strikes, boycotts, demonstrations, lobbying, and other more traditionally recognized forms of political action (Sturgeon 1997, p. 4).

Sturgeon argues that, as a social movement, ecofeminism is and must be involved in theorizing about concepts (e.g., of nature, women, race).

The power of ecological feminist theory lies in its ability to provide insights into the interconnections among the dominations of women and other human ‘others’ and the domination of nonhuman nature. The future of ecofeminist theory and research is to continue to show why, where, and how to accomplish the dismantling of interlocking human systems of unjustified domination.

See also: Feminist Ethics; Feminist Theology; Feminist Theory; Feminist Theory and Women of Color; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gender and Environment; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology

Bibliography

- d’Eaubonne F 1974 *Le Feminisme ou La Mort*. Pierre Horay, Paris
- Eisler R 1987 *The Chalice and the Blade*, 1st edn. Harper and Row, Cambridge, MA
- Gray E D 1979 *Green Paradise Lost*. Roundtable Press, Wellesley, MA
- Griffin S 1978 *Woman and Nature: The Roaring Inside Her*, 1st edn. Harper and Row, New York
- Gruen L 1994 Toward an ecofeminist moral epistemology. In: Warren K J (ed.) *Ecological Feminism*. Routledge, London, pp. 120–38
- Haraway D J 1991 *Simians, Cyborgs and Women: The Reinvention of Nature*. Routledge, New York
- Merchant C 1980 *The Death of Nature: Women, Ecology, and the Scientific Revolution*, 1st edn. Harper and Row, San Francisco
- Murphy P D 1995 *Literature, Nature, and Other: Ecofeminist Critiques*. State University of New York Press, Albany, NY
- Plumwood V 1993 *Feminism and The Mastery of Nature*. Routledge, London
- Spretnak C 1990 Ecofeminism: our roots and flowering. In: Diamond I, Orenstein G F (eds.) *Reweaving the World: The Emergence of Ecofeminism*. Sierra Club Books, San Francisco, pp. 3–14
- Starhawk 1990 Power, authority, and mystery. In: Diamond I, Orenstein G F (eds.) *Reweaving the World: The Emergence of Ecofeminism*. Sierra Club, San Francisco, pp. 73–86
- Sturgeon N 1997 *Ecofeminist Natures: Race, Gender, Feminist Theory and Political Action*. Routledge, New York
- Warren K J (ed.) 1997a *Ecofeminism: Women, Culture, Nature*. Indiana University Press, Bloomington, IN
- Warren K J (ed.) 1997b Taking empirical data seriously. *Ecofeminism: Women, Culture, Nature*. Indiana University Press, Bloomington, IN, pp. 3–20

Warren K J 1996 The power and the promise of ecological feminism. In: Warren K J (ed.) *Ecofeminist Philosophy*. Rowman and Littlefield, Lanham, MD, pp. 19–41

K. J. Warren

Feminist Theory: Liberal

Liberal feminist theory is liberal political theory and feminist political theory—meaning that it advances the theoretical tradition of liberalism while it contributes to a political theory based on the foundation that men and women are of equal human worth. It is also a critical practice of assessing the laws and social practices sustained by liberal theory which might inhibit the realization of men's and women's equality.

Classical liberalism describes a limited government intended to secure the individual freedom and the political equality of (some) men (Locke 1998). Rational, free, and equal men contract in the public sphere to organize limited government that will protect their autonomy. Contemporary welfare liberals respect the rational autonomy and freedom of individuals but interpret equality more broadly and accept a role for the state in fostering political equality.

Because liberal feminists respond to theoretical traditions, laws, and social practices that have changed over time, liberal feminism is not defined quickly or narrowly. Some liberal feminists, particularly early liberal feminists, work within the framework of classical liberalism and seek to extend equality to women based on the assertion that women like men are rational individuals. More recently others challenge the public/private dichotomy that is sustained by the theory of the state as a contractual arrangement among autonomous rational individuals (Pateman 1988). Thus, some liberal feminists work within liberal theory to change its implications for women while others stretch the bounds of liberal theory to adapt it to feminist demands (Phillips 1993). In addition, due to their concern about inclusion, many liberal feminists engage in criticism of some aspects of liberal feminist thinking itself. Thus, the boundaries between liberal feminist inquiry and the broader feminist project of making theory, law, and social practice and the criticism thereof more inclusive social processes are not tidily delineated.

1. The Historical Tradition of Liberal Feminism

Liberal feminists of the eighteenth, nineteenth, and early twentieth centuries asserted women's ability and right to participate in public life at least through the suffrage, examined the institution of marriage and the legal framework of divorce and property laws, and

challenged the practices of denying women access to the same quality education as men and to the professions. Liberal feminists sought to end the treatment of women as legal dependents on their husbands or fathers. Through examination of laws and practices, liberal feminists including Mary Astell (1666–1731), Mary Wollstonecraft (1759–99), Harriet Taylor (1807–58), John Stuart Mill (1806–73), Elizabeth Cady Stanton (1815–1902), and Virginia Woolf (1882–1941) drew on the liberal tradition's value of equality and individual freedom to argue that, just as social status at birth was no longer a legitimate basis on which to discriminate among men as liberals argued, so also sex at birth was no longer a legitimate basis on which to discriminate against women.

Among liberal feminists, justifications varied. To defend their claim, some feminists argued that men and women were of equal human worth; some argued that some women were superior to some men and thus all women should be given opportunity to demonstrate their worth; and some argued that women's virtues were superior to men's and thus warranted influence on public economic and political activity. Mill drew on each of these when it was strategically prudent. However, in general, the liberal argument took the form of asserting that individual development required the ability to educate and expand one's faculties such that to deny women access to education and the professions was to deny the development of their full human potential. Further, liberal feminists argued that society as a whole would benefit from the public contributions of women.

While liberal feminists asserted that society in general would benefit from educated women's contributions, generally these early liberal feminists did not pay adequate attention to the circumstances of working class women. The male model of life experience to which they wanted women to gain access was that of a white professional class man. However, more activist liberal feminists such as Maria Stewart (1803–79) and Susan B. Anthony (1820–1906) did pay attention to the circumstances of Black women and working class women.

In the latter twentieth century the arguments for women's equal access to education and the professions and for treatment as autonomous individuals (rather than individuals assumed to be dependent on a father or husband) by the law were heard from feminists again. However, although some feminists focussed their attention on the plight of the white, middle class, heterosexual, married Western woman (Friedan 1974), in general the theoretical and practical attention of liberal feminists switched from equality to inclusion. While demanding that women get the same opportunities as advantaged men, feminists also interrogated the very desirability of a western male model of life experience where professional and public activities were seen as more important and fulfilling than community and family activities. Instead, these

feminists sought to change the gendered practices of men and women at home and in public such that domestic caregiving was given social, legal, and economic recognition commensurate with its social value (Okin 1989; Steinem 1983) (see *Gender and Feminist Studies*).

Further, women of color, lesbians, and women from around the world encouraged western feminists to think beyond the bounds of white, middle class, Western experience (Lorde 1979, Mohanty 1984). Consequently, during the twentieth century, the political agenda of liberal feminism moved beyond a focus on married women's property rights, divorce laws, and the suffrage to include labor laws, social issues including child-care, welfare, health-care and education, access to safe legal abortion, international development, the distribution of economic resources, and human rights (NOW 1967, Sen and Grown 1987, Bunch 1990).

Even as the liberal feminist agenda expanded, theorists and activists wondered if liberal feminism was theoretically adequate for challenging norms of colonialism, global inequality, racism, patronage politics, and other forms of hierarchy that complicate analysis of gender inequality. For example, at the dawn of the twenty-first century, women's human rights activism is a global movement of national and international feminists acting in locally specific ways and sharing their experience globally. This movement ostensibly promotes the liberal values of freedom and equality, but the scope of freedom and equality extends well beyond classical liberal bounds to include global provision of basic needs, education, and healthcare and access to development resources (Tinker 1990, Kabeer and Subrahmanian 1996). Thus, the social interests of contemporary liberal feminists extend far beyond those of early liberal feminists. Most early twenty-first century liberal feminists seek to challenge fundamental values, practices, and norms within and across societies (Charlesworth 1994, Afkhami 1995, Okin 1999).

2. Feminism within Liberal Theory

In the tradition of liberal feminist theory, theorists have not only promoted liberal values of freedom and equality for women but also re-examined those same values in light of women's experiences. While working within the liberal tradition, liberal feminists work to change liberal theory by examining its accepted arguments and underlying assumptions. Thus, liberal feminists advance feminism and liberalism. For example, Carole Pateman (1988) shows that liberal theory's use of contract doctrine assumes freedom and equality that do not exist among the contracting parties. In a private contract between the sexes, women lose their ability to contract as autonomous individuals with others in public. Thus, women are not

equal parties to the original political contract on which classical liberal social contract theorists' base liberal democracy. For women the 'sexual contract' is not an exercise of freedom. Rather than wanting to rewrite the sexual contract and the political contract as some feminist theorists have argued, and as some political theorists have done by sanitizing their texts of sex-specific language, Pateman criticizes the use of classical liberal theorists' fundamental tool, consent theory or social contract theory. Pateman undermines the oppressive roots of contract theory and reveals its use in the validation and regulation of slavery, employment contracts, marriage and divorce, prostitution, and surrogacy. She reveals that the liberal premise of natural individual freedom and equality is a fallacy and as such is an inappropriate basis for political theory.

MacKinnon (1993) argues that liberal notions of equality are based on an historical tradition dating back to Aristotle according to whom treating people equally means treating likes alike, and unlikes differently. MacKinnon argues that contemporary liberal theory sustains this perspective. Consequently, although equality is an important foundation of liberal theory, as defined it is not an adequate basis for challenging exploitative inequality where the inequality is the result of a difference that is perceived to be natural or ordained by God. MacKinnon proposes an alternative conception of equality as the absence of hierarchy. Such a definition enables the 'unlikes' to demand equality.

In addition to challenging the underpinnings of liberalism, liberal feminists also challenge the scope of liberal inquiry. For example, Okin (1989) challenges the liberal assumption that justice can be assumed to exist within families and between family members and instead argues that the social institutions of marriage and childrearing practices create a vulnerability of women in marriage that is unjust and leads to harm to women and children. Okin proposes either changing the traditional gendered division of private and public work and related gender differentiation of men's and women's roles in families and in public or compensating those women and their children who are made vulnerable by the existing gendered system. Specifically, until society does not perpetuate gender difference in economic means and opportunities, wives are entitled to and should receive half of their husbands' incomes. Such an arrangement should continue even in divorce for as long as the gendered arrangement existed during the marriage. With these and related proposals, Okin challenges society to take responsibility for the vulnerability of women and children that is created by the institution of gendered division of work in marriage and its impact on other social institutions.

Because they challenge the root norms of liberalism—freedom and equality—these feminists might be classified as 'radical'; yet their contributions

to liberal theory's cornerstones of freedom, equality, and justice make them liberal feminists as well. By re-examining the meanings of freedom and equality and the institutions that liberals rely on to promote them, liberal feminists advance liberal theory while advancing the feminist agenda of promoting theory, law, and social practices that respect men and women as having the equal worth of all human beings (see *Feminist Theory: Radical Lesbian*).

3. Feminist Critiques of Liberal Feminism

While engaged in a critical project of its own, liberal feminist theory has also been the target of criticism. However, liberal feminists generally respond to criticism by expanding their collective agenda and self-understanding such that liberal feminism continues to evolve.

Marxist feminists have criticized capitalism as causing the sexual division of labor in the family and have criticized liberal feminists' focus on equal opportunity (see *Feminist Theory: Marxist and Socialist*). Liberal feminists likewise are concerned about the sexual division of labor in the family (see *Family and Gender*) but they seek to change social norms regarding male and female roles in family maintenance—increasing women's role in providing the economic resources for the family and promoting equal opportunity for women in economics and politics while increasing men's role in providing for the care of the family (Okin 1989). In addition liberal feminists working in development have challenged the global distribution of resources and economic power without objecting to a capitalist organization of economics (Sen and Grown 1987).

Radical and lesbian feminists accuse liberal feminists of accepting patriarchy and the heterosexual model of family such that any challenge to gender norms within the family is not radical enough. In other words, according to their radical critics, liberal feminism does not get at the root causes of gender inequality (see *Feminist Theory: Radical Lesbian*). Liberal feminists respond by proposing family related policies that incorporate the experiences and needs of single and lesbian women and do not assume that families are headed by heterosexual married couples.

Although generally responding to its critics in ways which expand the scope of its own agenda, liberal feminist theory is less responsive to postmodern feminist theory. Postmodern feminists interrogate categories and challenge analyses that rely on norms of 'woman' and 'gender' as being exclusionary and limiting even in their attempt to be critical and analytical (see *Feminist Theory: Postmodern*). Liberal feminists respond by being self-conscious about their own ability to exclude through their assumptions (as liberal theorists exclude through their assumptions), but they are unwilling to abandon the use of categories of analysis as a tool for their critical endeavor

(Benhabib 1995). Moreover, they accuse postmodern feminists of abandoning a genuine concern for the excluded by undermining the tools necessary to analyze their exclusion (Okin 1995, Nussbaum 1992).

In sum, because it responds to theoretical traditions, laws, and social practices that have changed over time, liberal feminism is not narrowly definable. Moreover, due to their concern about inclusion, many liberal feminists not only engage in a critique of liberalism but also incorporate into their analyses criticism that has been directed toward certain aspects of liberal feminist thinking. Thus, liberal feminism is part of the broader feminist project of making theory, law, and social practice and the criticism of these a more inclusive social process.

See also: Evolutionism, Including Social Darwinism; Feminist Epistemology; Feminist Ethics; Feminist Theology; Feminist Theory; Feminist Theory and Women of Color; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Liberalism; Liberalism: Historical Aspects; Liberalism: Impact on Social Science; Mobility: Social; Women's Suffrage

Bibliography

- Afkhami M (ed.) 1995 *Faith and Freedom: Women's Human Rights in the Muslim World*. Syracuse University Press, Syracuse, NY
- Anthony S B 1972 Woman wants bread, not the ballot! In: Schneir M (ed.) *Feminism: The Essential Historical Writings*, 1st edn. Random House, New York, pp. 137–42
- Astell M 1970 *Some Reflections upon Marriage*. Source Book Press, New York
- Basu A (ed.) 1995 *The Challenge of Local Feminisms: Women's Movements in Global Perspective*. Westview Press, Boulder, CO
- Benhabib S 1995 Feminism and postmodernism: An uneasy alliance. In: Benhabib S, Butler J, Fraser N, Cornell D (eds.) *Feminist Contentions: A Philosophical Exchange*. Routledge, New York, pp. 17–34
- Bunch C 1990 Women's rights as human rights: Toward a revision of human rights. *Human Rights Quarterly* 12(4): 486–98
- Charlesworth H 1994 What are 'Women's International Human Rights?' In: Cook R J (ed.) *Human Rights of Women: National and International Perspectives*. University of Pennsylvania Press, Philadelphia, pp. 58–84
- Flax J 1995 Race/gender and the ethics of difference: A reply to Okin's 'gender inequality and cultural differences.' *Political Theory* 23(3): 500–10
- Friedan B 1974 *The Feminine Mystique*, 10th anniversary edn. Norton, New York

- Kabeer N, Subrahmanian R 1996 Institutions, relations and outcomes: A framework and case studies for gender-aware Planning. IDS discussion paper, Institute for Development Studies, Brighton, UK
- Locke J 1988 *Two Treatises of Government*. Cambridge University Press, Cambridge, UK
- Lorde A 1979 The master's tools will never dismantle the master's house. In: Moraga C, Anzaldua G (eds.) *This Bridge Called My Back: Writings by Radical Women of Color*. Kitchen Table, Women of Color Press, New York, pp. 98–101
- Lorde A 1981 An open letter to Mary Daly. In: Moraga C, Anzaldua G (eds.) *This Bridge Called My Back: Writings by Radical Women of Color*. Kitchen Table, Women of Color Press, New York, pp. 94–7
- MacKinnon C 1993 Crimes of war, crimes of peace. In: Shute S, Hurley S (eds.) *On Human Rights: The Oxford Amnesty Lectures 1993*. Basic Books, New York, pp. 83–109
- Mill JS 1980 *The Subjection of Women*. Attar, Arlington Heights, IL
- Mohanty C T 1984 Under Western eyes: Feminist scholarship and colonial discourses. In: Talpade Mohanty C, Russo A, Torres L (eds.) *Third World Women and the Politics of Feminism*. University of Indiana Press, Bloomington, IN, pp. 51–80
- National Organization for Women 1967 NOW Bill of Rights. In: Jaggar A M, Rothenberg P S (eds.) *Feminist Frameworks: Alternative Theoretical Accounts of the Relations between Women and Men*. McGraw-Hill, New York, p. 159
- Nussbaum M 1992 Human functioning and social justice: In defense of Aristotelian essentialism. *Political Theory* 20(2): 202–46
- Okin S M 1989 *Gender, Justice and the Family*. Basic Books, New York
- Okin S M 1994 Gender inequality and cultural differences. *Political Theory* 22(1): 5–24
- Okin S M 1995 Race gender and the ethics of difference—Response. *Political Theory* 23(3): 511–16
- Okin S M 1999 *Is Multiculturalism Bad for Women?* Princeton University Press, Princeton, NJ
- Pateman C 1988 *The Sexual Contract*. Stanford University Press, Stanford, CA
- Phillips A 1993 *Democracy and Difference*. The Pennsylvania State University Press, University Park, PA
- Sen G, Grown C 1987 *Development, Crisis, and Alternative Visions: Third World Women's Perspectives*. Monthly Review Press, New York
- Stanton E C 1972 Seneca Falls Declaration. In: Schneir M (ed.) *Feminism: The Essential Historical Writings*, 1st edn. Random House, New York, pp. 76–82
- Steinem G 1983 *Outrageous Acts and Everyday Rebellions*. New American Library, New York
- Stewart M 1987 *America's First Black Woman Political Writer*. Indiana University Press, Bloomington, IN
- Taylor H 1970 Enfranchisement of women. In: Rossi A S (ed.) *The Subjection of Women in Essays on Sex Equality*. University of Chicago Press, Chicago, pp. 91–121
- Tinker I 1990 A context for the field and for the book. In: Tinker I (ed.) *Persistent Inequalities: Women and World Development*. Oxford University Press, New York, pp. 3–13
- Wollstonecraft M 1795 [1792] *A Vindication of the Rights of Women: An Authoritative Text, Backgrounds, Criticism*. Poston C (ed.). Norton, New York

B. A. Ackerly

Feminist Theory: Marxist and Socialist

1. Definition of the Field

Socialist feminist and Marxist feminist theory has been a major form of both feminist theory and of socialist theory. Key elements include: the analysis of gender in its intersection with class; a focus on the material, economic, and political dimensions of social relations; a prioritization of the analysis of inequality rather than difference; a focus on social transformation rather than either the recognition of difference or of equal opportunities; and location within the Enlightenment tradition with its commitment to reason, rather than postmodernist scepticism about the possibility of objective explanation.

It was conventional in the 1980s to distinguish three major types of feminist theory in which socialist and feminist theory was one alongside liberal feminism (with a focus on individuals and equal opportunities) and radical feminism (with a focus on men's oppression of women). Since then the major division in feminist theory has become that between modernist theories such as socialist feminism and postmodernist theories (described below). There is a fuzzy rather than clear-cut boundary around socialist and Marxist feminist theory. For instance, there is an issue of how much autonomy can be granted to gender relations from class relations before the theory is better classified in materialist feminism or radical feminism rather than socialist feminism. Or again, it is unclear how much significance can be accorded to culture before a different categorization becomes appropriate.

There are various sources of variation within socialist feminism including among others: the extent of the centrality of class relations as against the centrality of men's direct exploitation of women and the relative importance of economic and cultural dimensions.

Marxist feminist and socialist feminist theory overlap significantly, though the extent of this varies in different times and places. Sometimes the terms are used interchangeably, while in other contexts there are specific demarcations. When there are demarcations, this most often involves Marxist feminism being more focused on class analysis and socialist feminism on the specificity and autonomy of gender dynamics. Further, materialist feminist theory overlaps with socialist feminism, in that it has an emphasis on the material exploitation of women by men, though the extent of the priority given to gender is such that it is sometimes distinguishable from socialist feminism.

2. Historical Development

Socialist feminist theory first developed in Western capitalist societies in the eighteenth century, initially in the context of radical communitarian forms of social-

ism which embraced women's liberation as a core part of their practice, such as that of Owenism. Marxist feminist thought developed a little later, in the nineteenth century, in the context of Marxist socialist practice and theory and of both class struggle and first-wave feminism.

While Marx himself had little to say about feminism, Engels (1941) did, writing of the changing forms of the family using a historical materialist method. He drew attention to reproduction as well as to production, and the way in which the different social organization of the form of reproduction in the family affected many other dimensions of society. The merit and precise meaning of the distinction between reproduction and production has been the subject of extensive later analysis. Gilman (1896), a materialist feminist and a major theorist of first-wave feminism, focused the analysis of women's labor onto the way that men were the beneficiaries of the exploitation of women. This theoretical focus is echoed in the work of contemporary theorists such as Delphy (1984), who likewise see the expropriation of women's labor in the household as key to the explanation of women's subordination. Materialist feminism shares with socialist and Marxist feminism a concern with the centrality of labor, but pulls away from it as a result of the assumption of the independence of gender dynamics from those of class. In the revitalization of feminism in the West from the 1960s, there was a flowering of many hues of socialist feminist and Marxist feminist thought, both within and without the academy. This led to a major rethinking of the significance of gender for social theory and the social and behavioural sciences more broadly. Specific attention was paid to the analyses of women's labor both at home and in the labor market. During the 1970s and 1980s there was a significant extension in the range of substantive concerns which were subjected to a socialist feminist analysis, including the state, sexuality, culture, and male violence.

3. Domestic Labor Debate

The domestic labor debate of the 1970s set out to theorize the relationship between housework and capitalism. Housework was understood broadly in the context of domestic labor. This was intended to include childcare as well as housework, and indeed other forms of care work and servicing that women did at home without a wage. The protagonists debated the possible benefits of domestic labor for capitalism, and for husbands and fathers. Capitalists were seen to benefit from the nurturance and care of male workers who were more productive as a consequence, and also from the rearing of the next generation of well-socialized workers. The debate explored the extent to which it was possible to utilize traditional Marxist terminology and apply it to gender inequality, and the extent to which a new set of concepts was needed. In

particular, they asked whether domestic labor was productive, unproductive, or nonproductive, of value, and surplus value. This was important to Marxist theory of capitalism, since labor that was productive of value and surplus value was considered to be more central to the workings of capitalism than work that was unproductive or nonproductive. These feminist writers thus challenged successfully the traditional Marxist assumption that women's domestic labor was marginal to capitalism, as part of their argument that Marxist theory needed a better appreciation of the significance of gender relations, and of women as a potential political force.

The debate was a highly sophisticated exploration of these theoretical issues, which eventually left far behind its starting point in the everyday oppression of women—too far behind for some critics of the entire debate. The whole field was criticized for economism and its use of arcane technical language. With the fall of the kind of socialism practiced in Eastern Europe Marxist conceptualization has become regarded as less significant. Today, Marxist concepts are utilized rarely in the understanding of housework. Nevertheless, the question of the nature of the relationship of domestic labor to the market economy remains an important one, and one pursued by socialist feminists using a less overtly Marxist terminology. Today the focus is on domestic labor as care-work (Gardiner 1997), and on the variations in the extent to which the state provides for the collective and social performance of this labor, rather than its being left as a private matter for individual households (Folbre 1994).

4. Women's Waged Labor

The inequality between women and men in the labor market was theorized as resulting from capitalism, with major debates as to the mechanisms by which the connection was to be understood. The inequalities include: lower wages; lesser job security; fewer fringe benefits such as employer-based pension schemes; occupational segregation. The nature of the causes of the greater vulnerability to exploitation of women's labor than men's was the subject of much debate. This greater vulnerability was seen sometimes as stemming from women's position in the family that ill-prepared them with the skills needed for the labor market. Further, divisions and lack of solidarity between workers as a result of gender and racialized divisions were seen to facilitate such exploitative strategies by employers (Cockburn 1983).

A variant of socialist feminism, sometimes referred to separately as dual-systems theory, theorized the subordinate position of women in the labor market and family as a result of the interlocking of the system of capitalism with a system of patriarchy (Hartmann 1976). Here the subordination of women in the system of patriarchy interacted with that of women's subordination within capitalism, and was, in certain

periods and locations, fuelled by the organization of men on their own behalf, such as in the male-only or male-dominated trade unions of the late nineteenth century. Dual-systems theory played an important role in the development of analyses of occupational segregation, since it was hard to see that this form of labor market rigidity was in a simple sense a result of employers' interests, or indeed women's interests, but rather that to be understood it was necessary to include a notion of male workers' interests. The relationship between patriarchy and capitalism was sometimes seen as harmonious, in that the subordination of women in one system supported their subordination in the other, and sometimes in tension, as when the demand for labor by expanding capitalist economies pulled women out of patriarchal households (Walby 1986).

Recent socialist feminist analysis on women's labor has included analysis of the strategy for delivering comparable worth, that is, equal pay for work of equal value using a legally-based intervention. This focuses on the extent to which patriarchal work systems are resilient to feminist campaigns rooted in equality politics. While most of the analysis of comparable worth has taken place in the US, with largely pessimistic conclusions as to the tenacity, depth, and flexibility of patriarchal practices in the workplace, the strongest legislation is actually within the EU, where analysis is still in its early stages.

Economics as a discipline has integrated a feminist perspective to a lesser extent than some other disciplines, for instance, sociology. Yet, there has been a sustained critique of and engagement with mainstream economics, in which socialist feminist theoretical leadership has played a significant role (England 1992, Cook et al. 2000).

5. *Class Divisions*

In socialist and Marxist feminist theory class divisions are important in a number of ways. First, the class divisions between capital and labor structures women's lives, both their domestic labor and their participation in waged labor. Second, the class divisions between women mean that it is not possible to generalize about women as a whole, since women in different class locations will have different interests. For instance, a woman who works as a cleaner in a private household will not have the same economic interests as the woman who employs her.

6. *Culture*

Most socialist and Marxist feminist theory has conceptualized culture as a secondary dimension of women's lives. While patriarchal ideologies that support women's oppression are noted, these are not usually seen as the most important part of the explanation.

However, since the late 1980s the perceived importance of culture has grown within much Western social science, although the extent of this change is uneven and is more acute in the UK than some other countries. This 'cultural turn' has had an impact upon socialist and Marxist feminist theory in two ways. First, the turn to culture was incorporated within a self-defined Marxist feminism, as part of a new and more sophisticated interpretation of Marxism. For instance, there were attempts to address questions of sexuality and to integrate Marxism with Freud and with Foucault. Second, there was a turn away from socialist and Marxist feminism, and a critique of this school of thought as economistic. The overall impact was to displace the concern with women's paid and unpaid labor from its central location in theoretical questions within feminist theory, even though empirical studies continued on women's labor.

7. *'Race,' Ethnicity, and Gender*

Since the 1980s there have been criticisms of feminist (and other) academic writing for insufficient attention to issues of 'race' and ethnicity. This was not only a criticism of the scarcity of studies focused on issues of race and ethnicity. More importantly, it was a criticism of the lack of theoretical centrality given to race and ethnicity in feminist theory. The strongest criticism was that socialist and Marxist feminist theory was inherently unable to deal with the complexity needed to theoretically engage with 'race' and gender simultaneously.

However, attempts at analyzing the relationship of gender and ethnicity by the critics have themselves been subject to problems, not least of which have been tendencies to, first, empirical depth at the expense of explanatory power; and second, a concern with difference at the expense of inequality. These issues are re-articulated in the debate on postmodernist developments in feminist theory.

The integration of issues of analyses of ethnicity, gender, and class remains an important theoretical debate today, and re-emerge in different forms in the topics discussed below, especially those of: postmodernism/modernism; recognition/transformation; agency/structure; development and globalization.

7.1 *Postmodernism/Modernism*

Socialist and Marxist feminist theory is modernist rather than postmodernist and draws strongly on an Enlightenment heritage. This involves several elements of which two will be noted here. First, there is an assumption that explanation, while fallible, is both possible and desirable, and that merely describing the world, albeit in complex and interesting ways, is too modest an ambition for social science. The search for an explanation of the world is necessary in order to understand how to change the world. Without an

explanation of the way the world works, attempts at change, such as a feminist political project, are directionless. Second, there is a search for simplicity in the explanation of large and complex phenomena, in terms of powerful underlying motors of change, and of underlying large-scale social structures. This is in contrast with the postmodernist fascination with complexity that has a tendency to merely descriptive accounts of processes.

7.2 Recognition/Transformation

Socialist and Marxist feminist theory seeks to analyze processes of transformation rather than simply recognizing existing social relations or limiting demands for social justice to that of equal opportunities (Fraser 1995). There has been a vigorous debate in feminist (and other social) theory as to the merits of the politics of recognition and the politics of equality, as to which is more effective in achieving social justice. Equality politics has been criticized because, in its focus on the unequal distribution of material privileges, it neglects the significance of cultural differences that are alleged to be more important. Recognition politics focuses on cultural dimensions of social life and on the positive valuation of difference in the political arena. Two examples of recognition politics are the demand for respect for religious customs and for quotas for the representation of subordinate social groups in political bodies, such as parliaments. An example of equality politics is a drive for equal pay through equal opportunities laws. The socialist feminist critique of this debate holds that the strategy of recognition politics can merely embed existing inequalities even as it recognizes the cultural specificity of subordinate groups, while equal opportunity politics, by seeking redistribution within an existing hierarchical system, does not fundamentally alter that system. By contrast, the socialist feminist argument is that a focus on transformation is needed since it is the overarching system of social inequality which needs to be changed, not just which groups, such as women or racialized minorities, are disadvantaged. This entails a less hierarchical division of labor and less unequal distribution of income. It is argued that the goal of transformation demands reasoned argument to build a new social project, rather than embracing existing differences and the projects of existing groups, or seeking the redistribution of existing privileges. This debate has echoes in many branches of contemporary social and political theory, as it grapples with integrating theoretical concern with difference into traditional social theory.

7.3 Agency/Structure

Socialist and Marxist feminist analysis uses concepts of social structure. This is in contrast to some forms of feminist theory that focus primarily on agency and

performativity. This conceptualization of the social world is needed in order to capture the deep rootedness of social processes, their persistence over time, and to enable an analysis which connects large-scale processes of different kinds into an overarching system. This does not deny the relevance of agency, despite many criticisms of structural analysis to this effect, since collective political agency is regularly theorized by socialist feminists as a key route to social change. But rather it involves rooting agency in structure, so as to understand that while people choose, they do not do so under conditions of their own making.

7.4 Globalization and Development

While within some branches of Western feminist theory there was a move away from structural accounts of gender inequality during the 1980s and 1990s, this change has not been universal. In particular, within Development Studies the socialist feminist tradition has remained strong (Moghadam 1996). This has been partly because of the agreed significance of the issue of economic poverty and of the role of international capital in the shaping of the process of development.

Since the late 1990s the interest in the issue of globalization is leading to a resurgence in the types of analysis which traditionally have been at the core of socialist feminist analysis, such as the transformation of work and collective political agency. This is because of the desire to understand the interconnectedness of large-scale global processes, and the role of international capital in the changes associated with globalization.

8. The Mainstreaming of Socialist Feminist Theory?

It would be conventional to suggest that socialist and Marxist feminist theory has lost most of its vibrancy during the 1980s and 1990s and that the ascendancy in feminist theory has passed to theories which are more cultural and literary in substantive and theoretical focus. Indeed since the fall of the Wall between East and West Europe Marxist theory has become unfashionable, being regarded as disproven by historic events. Certainly hypotheses about the revolutionary potential of the proletariat need to be rejected.

However, this should not lead to the underestimation of the impact of socialist feminist theory and the continuing significance of its core theoretical concerns. Several of the initial positions of socialist feminist theory which were once regarded as unusual and original have become accepted by traditions which might once have been hostile opponents of such theory. These include, for example, arguments as the importance of labor to understanding gender inequality, and of the centrality of gender to the understanding of the labor market, which are now mainstream assumptions. Today, it is rare for main-

stream social science to deny the relevance of a gender perspective to their subject matter, unlike in the 1970s. Once gender issues were confined to family studies, but today the relevance of gender to a wider set of issues, such as the labor market, is much more readily accepted. To a significant extent then, socialist feminist theory has become mainstreamed and in so doing has lost some of its distinctive identity. In parallel the broadening of the substantive and theoretical concerns within socialist feminist theory has meant the loss of the specificity of some dimensions of a distinctive socialist feminist project (Walby 1997).

However, the move to postmodernism in much feminist theory in the 1980s and 1990s, left socialist feminism as an important carrier of the commitment to modernist forms of theorizing, with a focus on explanation, a search for deep structures and their interconnections, and an interest in the conditions under which gender relations might be transformed. With the development of analysis of globalization the core theoretical concerns of social feminism are again central to current debates, even if they are no longer labeled as socialist feminist.

See also: Feminist Epistemology; Feminist Ethics; Feminist Political Ecology; Feminist Theology; Feminist Theory; Feminist Theory and Women of Color; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Marxism in Contemporary Sociology; Marxist Geography; Marxist Social Thought, History of; Theory: Sociological

Bibliography

- Cockburn C 1983 *Brothers: Male Dominance and Technological Change*. Pluto Press, London
- Cook J, Roberts J, Waylen G (eds.) 2000 *Towards a Gendered Political Economy*. Macmillan, London
- Delphy C 1984 *Close to Home: A Materialist Analysis of Women's Oppression*. Hutchinson, London
- Engels F 1941 *The Origin of the Family, Private Property and the State*. Lawrence and Wishart, London
- England P 1992 *Comparable Worth: Theories and Evidence*. Aldine de Gruyter, New York
- Folbre N 1994 *Who Pays for the Kids? Gender, the Structures of Consent*. Routledge, London
- Fraser N 1995 From redistribution to recognition? Dilemmas of justice in a 'post-socialist' age. *New Left Review* 212 (July/August): 68–93
- Gardiner J 1997 *Women, Care and Economics*. Macmillan, Basingstoke, UK

- Gilman C P (1896/1966) *Women and Economics: A Study of the Economic Relation between Men and Women as a Factor in Social Evolution*. Harper & Row, New York
- Hartmann H 1976 Capitalism, patriarchy and job segregation by sex. *Signs* 13: 137–69
- Moghadam V (ed.) 1996 *Trajectories of Patriarchy and Development*. Clarendon Press, Oxford, UK
- Walby S 1986 *Patriarchy at Work: Patriarchal and Capitalist Relations in Employment*. Polity Press, Cambridge, UK
- Walby S 1997 *Gender Transformations*. Routledge, London

S. Walby

Feminist Theory: Postmodern

The work of postmodern theorists has been a major influence in feminist theory since the 1980s. Building on the work of theorists such as Michel Foucault, Jacques Derrida, and Jean-Francois Lyotard, postmodern feminist theory emphasizes the constituting function of language and discourse in the construction of sexuality and the structure of gender relations. Breaking from modernist feminist theories such as liberal feminism and Marxist/socialist feminism, postmodern feminists reject the founding assumptions of modernist/Enlightenment thought: the attempt to find universal truth, to comprehend the totality of social reality and to define the essential nature of 'woman.' Through the postmodern method of deconstruction they attempt to displace these certainties. They emphasize the discursive nature of social reality and analyze the mechanisms by which gender identity is constituted discursively.

1. Postmodern Theories

Theories termed postmodern have their roots in Nietzsche's challenge to the fundamental values of the Western tradition, his 'transvaluation of values.' French structuralism and psychoanalysis also played a major role in the development of postmodernism. Levi-Strauss's structuralism focused on the way in which the overall structures of cultures create meaning for social actors. Jacques Lacan's psychoanalytic theory asserted that language constitutes the psyche, specifically the definitions of masculinity and femininity; he defined the possibility of meaning in language as deriving from the phallus. These and related theories produced a generation of thinkers, particularly in France, who looked to language and discourse as the constituent elements of social reality. They turned away from the quest for universal truth, what they called 'metanarratives,' and turned instead to the specific definition of 'truth' established by discourses within particular societies.

Lyotard's influential *The Postmodern Condition* (1984) defined the parameters of the postmodern

project. Lyotard argued that 'knowledge' in the contemporary world was necessarily plural and diverse, that the metanarratives that defined modernism no longer hold sway. Identifying postmodernity as discontinuous with modernism, Lyotard argued for new definitions of truth and knowledge appropriate to the postmodern condition. Michel Foucault and Jacques Derrida further challenged the modernist definition of a single, absolute Truth. Their distinctive contribution to postmodern thought was the theory of the death of 'man' or the subject. Foucault argued that the subject of modernism, the autonomous agent who constitutes meaning, is neither universal nor absolute, but, rather, the product of specific discourses at a particular juncture in Western thought. Derrida's work attacked the opposition between the rational, autonomous subject of the Western tradition and the object of knowledge that this subject constitutes. His critique called into question a particular aspect of this subject: the author who creates the meaning of the text. Both Foucault and Derrida argued that this subject, the centerpiece of modernism and humanism, is doomed to extinction.

One of the principal aspects of postmodern theories is an attack on the binaries—dichotomous, hierarchical oppositions—that are at the center of modernist thought. Postmoderns argue that modernist thought is structured by a series of binaries: rational/irrational, nature/culture, subject/object, reason/emotion, absolute/relative, male/female. A major aspect of the postmodern project is the attempt to displace these binaries. Postmoderns argue, first, that they are the result of historically situated discourses and, second, that they distort and limit understanding. The focus of postmodern method is the attempt to 'deconstruct' these binaries, to reveal the unstated presuppositions informing them and the strictures they impose on thought.

2. The Convergence of Feminism and Postmodernism

There is widespread agreement among intellectuals that Western thought is witnessing a major shift at the end of the twentieth century: the turn away from the absolutes of modernist thought and its central figure, the autonomous subject, and the embrace of situational, perspectival knowledge and a relational, constituted subject. Both postmodernism and feminism are at the forefront of this intellectual shift. Both engage in a radical critique of modernist thought, attacking the roots of its understanding of knowledge and the subject. They agree that the binarisms that inform modernist thought are necessarily gendered, rooted in the central opposition between male and female and the subordination of the feminine entailed by this opposition. Postmoderns attack this binarism

on the grounds that it distorts knowledge and presupposes a false opposition between the essentially feminine and the essentially masculine. At the heart of the modern feminist movement is an attack on the male/female binarism on the grounds that it presupposes the inferiority of women. The postmodern and feminist convergence on this central issue was the initial basis of the attraction of postmodern theories for feminists. The postmodern project of deconstructing 'man' converged in many ways with the feminist project of challenging patriarchy.

Many feminists first encountered postmodern themes in the work of what are now called the 'French Feminists.' Luce Irigaray, Julia Kristeva, and Helene Cixous theorized about the construction of the feminine within language, particularly the construction of female sexuality. Building on the work of Jacques Lacan who defined the position of women in language as a 'lack,' they argued that the oppression of women in Western societies is rooted in language and what they identified as its 'phallographic' assumptions. They emphasized that language creates subjects, particularly the subject 'woman,' and that this creation both defines and distorts female sexuality. The French Feminists accepted Lacan's thesis that the basis for meaning in language is the phallus, but, instead of accepting his conclusion that women are, therefore, erased by language, they attempted to fashion a 'feminine writing' that would give women a voice.

The work of the French Feminists, although not strictly postmodern, presaged many of the themes of postmodern feminism. The French Feminists identified language as the source of meaning, subjectivity, and sexuality. They questioned the binarisms that structure Western thought and the inferior status for women that they entail. Their attempt to define a 'feminine writing,' however, appears to contradict the postmodern impulse to reject essentialism. This aspect of the French Feminists' thought continues to be controversial for postmodern feminism.

3. Postmodern Feminist Theories

The clearest statement of the major issues surrounding postmodern feminism is found in an influential article by Fraser and Nicholson (1990). They argued that feminism and postmodernism are the most important political-cultural currents of the 1980s and that the two approaches can both profit from a closer alliance. Their central thesis is that the strengths of each movement could and should correct the weaknesses of the other. The strength of postmodernism, its rejection of metanarratives, can counter feminism's tendency to construct such narratives. Conversely, feminism's orientation toward social criticism can rectify postmodernism's avoidance of social critique. Calling for an integration of the two approaches, Fraser and Nicholson argued that a postmodern feminism would

be an approach that avoids the metanarratives of modernism but could nevertheless engage in large historical narratives and analyses of societal macro-structures.

The late 1980s and 1990s witnessed an explosion of discussions of postmodernism and feminism. Two books outlined the issues facing a postmodern feminism (Hekman 1990, Weedon 1987). Butler's (1990) book defined the distinctive approach to the subject that would characterize postmodern feminism. Relying on the postmodern critique of the modernist subject as well as psychoanalytic theory, Butler argued that there is no essence to gender identity but, rather, that gender is constituted through the performance of gender roles. Butler's thesis was that gender identity is a regulative ideal that grounds the norm of heterosexuality. She asserted that it is precisely the belief in gender identity, a belief that has been central to feminist theory and political practice since the advent of the feminist movement, that prevents us from identifying the political processes that produce that coherent identity. She thus concluded that gender identity, far from being the means of women's liberation, was the ground of women's oppression.

Butler's book served as a kind of lightning rod for both defenders and critics of postmodern feminism. A series of books and articles in the 1990s further established the parameters of the approach. A notable element of this formulation was the connection to Marxism/socialism. Several prominent postmodern feminists argued for a redefinition of socialist feminism along postmodern lines. Haraway (1990), in her influential article, argued that women need a new identity for the postmodern age that can replace the constructed revolutionary identity of socialist feminism. Like Butler, Haraway attempted to demonstrate how the postmodern critique of identity could be used to formulate a feminist politics appropriate to the postmodern era. In addition to extending the discussion of gender identity, her work also brought postmodern themes into another area of feminist concern: the critique of Western science. Another argument for the connection between postmodern feminism and Marxism/socialism came from Gayatri Spivak. Proclaiming herself a Marxist feminist deconstructionist with an overtly political agenda, Spivak argued for the political usefulness of postmodern feminism. Arguing against critics of the approach who defined it as exclusively a white, male theory, she also demonstrated how deconstructive methods could be used to critique issues involving race and colonialism.

Psychoanalytic theories became important in the emerging definition of postmodern feminism as well. Flax (1990) argued for the convergence of postmodernism, feminism, and psychoanalysis. Identifying feminist theory as a type of postmodern philosophy, Flax argued that, together with psychoanalysis, postmodernism and feminism contribute

to the fundamental questioning of the certainties of the Western tradition. Together, she asserts, these three movements force a redefinition of those certainties and encourage questioning of the need for imposing order on social reality and defining its universal structure.

Another influential postmodern feminist theorist, Drucilla Cornell, applied postmodern theories to feminist legal studies. Although legal theory would seem to be an inappropriate subject for postmodern deconstruction, Cornell (1991) argued that legal theory needs a new choreography of sexual difference that can only be supplied by postmodern deconstruction. In a series of books on deconstruction and the law, Cornell developed the ethical dimensions of deconstruction and argued for its usefulness in liberal legal theory.

Postmodern themes have appeared in nearly every aspect of feminist thought. The strongest proponents of a postmodernism feminism are found in English literary criticism where the influence of Derrida and Foucault is pervasive. But postmodern themes also appear in women's history, political theory, philosophy, cultural studies and feminist critiques of science.

4. Critiques of Postmodern Feminist Theories

Since it was first introduced in feminist theory, postmodernism has had a mixed reception in the feminist community. The well-known collection, *Feminism/Postmodernism* (Nicholson 1990), is representative of this reception: the articles are almost evenly divided between critics and defenders of a post-modern feminism. The feminist critiques of postmodern feminism that have emerged in the 1980s and 1990s have focused on the fundamental aspects of the postmodern approach, questioning its appropriateness for feminist theory and practice.

The most pervasive critique of a postmodern feminism is that it precludes the possibility of a feminist politics. Di Stefano, for example, argued that feminism is 'firmly, if ambivalently, located in the modernist ethos' (1990). The basic concepts of the contemporary women's liberation movement, 'woman' and 'liberation,' are, she asserts, thoroughly modernist concepts. These concepts are deconstructed and rejected by postmodernism, thus destroying the foundation of feminist theory and politics. Di Stefano also questions the possibility of founding feminist politics on an approach that ignores gender altogether.

Ebert (1996) elaborates on this critique in one of the most comprehensive attacks on postmodern feminism. Ebert decries what she defines as the dominance of 'ludic feminism.' She argues that a feminism founded on postmodern assumptions about linguistic play, difference, and the priority of discourse will result in the substitution of a politics of representation for

radical social transformation. Ebert laments the fact that many socialist feminists have embraced postmodernism and thus abandoned the revolutionary resistance politics that defines Marxism. Ebert and other critics of postmodern feminism claim that by reducing everything to discourse and rejecting the notion of the truth of social reality, postmodernism precludes the possibility of a feminist politics. Far from fostering a politics of resistance, these critics claim, postmodernism entails either nihilism or an ultimately conservative defense of the *status quo*.

Defenders of postmodern feminism argue that politics in general and feminist politics in particular do not require the metanarratives of modernist politics. They also point to the work of postmodernists such as Foucault who clearly place their work in a tradition of political resistance. But, although Ebert and other critics of a postmodern feminism concede that some postmodern thinkers are interested in resistance, they assert that without a view of the truth of social reality, feminists cannot proclaim the truth of women's oppression, a necessary element of any feminist politics.

This critique is related closely to another pervasive criticism of postmodern feminism: its failure to produce systematic knowledge of social reality. Although Fraser and Nicholson argued that systemic social critiques are not precluded by postmodernism this argument has not been persuasive for many feminists. On the contrary, many feminists have argued that postmodernism's approach to knowledge produces nothing but meaningless textual analysis of the infinite differences of social reality (Hartsock 1990). Bordo (1990) argued that this aspect of postmodernism leads to 'gender scepticism' and the infinite play of differences without distinction. These critics conclude that postmodernism's resistance to comprehensive theories precludes a coherent understanding of social reality without which social change is impossible.

Another major focus of critiques of a postmodern feminism is the concept of the subject. Butler and other postmodern feminists' deconstruction of the modernist subject is central to their theoretical project. Yet their critics have claimed that this deconstruction leaves feminism with nothing to replace this subject. The modernist subject, they claim, had agency and autonomy, the postmodern subject has neither. Worse, postmodern theories of the subject produce what its critics call a 'social dupe,' a subject which is wholly constituted by discourse and thus incapable of agentic action. This not only produces passivity but fails to explain the possibility of resistance, including feminist resistance politics. The defenders of postmodern feminism claim that this critique is rooted in modernist assumptions that falsely link agency to the modernist subject. But the question of agency and the 'social dupe' remain central to critiques of postmodern feminism.

Postmodernism's emphasis on texts and discourse is

at the root of another critique of postmodern feminism: the charge that deconstructing materiality denies the reality of women's pain. Critics such as Bordo (1993) argue that a central concern of feminism must be an account of the very real, material pain inflicted on women's bodies by the structures of patriarchy and that postmodern feminism's rejection of materiality precludes this. The postmodern feminist rejoinder formulated by Butler (1993) is that focusing on discourses does not ignore the materiality of pain, but, rather, offers the possibility of transforming it. Butler claims that materiality must be the object of feminist inquiry, not its ground.

The origins of postmodern theories have also been the object of feminist critiques. Postmodern theorists, critics charge, are elite white men who ignore gender issues and write jargon-filled discourse that excludes the uninitiated. Feminists writing on race and ethnicity have been particularly vehement critics of this aspect of postmodern feminism. They claim that what Barbara Christian called the 'race for theory' marginalizes women of color. Specifically, these critics assert that the unique perspectives and theorizing formulated by women of color are ignored and negated by postmodern feminist theory.

One consequence of these critiques of postmodern feminism has been a return to the universals of modernism on the part of some feminists. Okin has argued that the anti-essentialism and hyperconcern with differences among women that characterizes contemporary feminism is a dangerous trend. She claims that the result of these emphases is a paralyzing degree of cultural relativism and the impossibility of making any generalizations about women's condition. A similar concern motivates the recent work of Martha Nussbaum. Nussbaum asserts that without a concept of universal human functioning feminists will be unable to address the wrongs done to women throughout the world. Like Okin, she is concerned about the extremes of what she sees as cultural relativism fostered by postmodern feminism and other approaches that emphasize differences among women.

In summary, the jury is still out on postmodern feminist theories. They have had a pervasive influence in contemporary feminism and have generated vehement defenders and critics. On the positive side postmodern theories have aligned feminism with the intellectual shift that characterizes contemporary Western thought and brought it into the era of differences. On the negative side postmodernism has raised questions about the foundations of any feminist politics and the status of the truth claims feminists advance.

See also: Cultural Critique; Anthropological; Cultural Relativism, Anthropology of; Deconstruction: Cultural Concerns; Feminist Epistemology; Feminist Ethics; Feminist Political Ecology; Feminist Theology; Feminist Theory; Feminist Theory and Women

of Color; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Political Thought, History of; Postmodernism in Geography; Postmodernism in Sociology; Postmodernism: Methodology; Postmodernism: Philosophical Aspects; Theory: Sociological

Bibliography

- Bordo S 1990 Feminism, postmodernism, and gender scepticism. In: Nicholson L (ed.) *Feminism/Postmodernism*. Routledge, New York, pp. 133–56
- Bordo S 1993 *Unbearable Weight: Feminism, Western Culture, and the Body*. University of California Press, Berkeley, CA
- Butler J 1990 *Gender Trouble*. Routledge, New York
- Butler J 1993 *Bodies that Matter*. Routledge, New York
- Cornell D 1991 *Beyond Accommodation: Ethical Feminism, Deconstruction and the Law*. Routledge, New York
- Di Stefano C 1990 Dilemmas of difference: Feminism, modernity, and postmodernism. In: Nicholson L (ed.) *Feminism/Postmodernism*. Routledge, New York, pp. 63–82
- Ebert T 1996 *Ludic Feminism and After: Postmodernism, Desire, and Labor in Late Capitalism*. University of Michigan Press, Ann Arbor, MI
- Flax J 1990 *Thinking Fragments: Psychoanalysis, Feminism, and Postmodernism in the Contemporary West*. University of California Press, Berkeley, CA
- Fraser N, Nicholson L 1990 Social criticism without philosophy: an encounter between feminism and postmodernism. In: Nicholson L (ed.) *Feminism/Postmodernism*. Routledge, New York, pp. 19–38
- Haraway D 1990 A manifesto for cyborgs: Science, technology, and socialist feminism in the 1980s. In: Nicholson L (ed.) *Feminism/Postmodernism*. Routledge, New York, pp. 190–223
- Hartsock N 1990 Foucault on power: A theory for women? In: Nicholson L (ed.) *Feminism/Postmodernism*. Routledge, New York, pp. 157–75
- Hekman S 1990 *Gender and Knowledge: Elements of a Postmodern Feminism*. Polity Press, Cambridge, UK
- Lyotard J-F 1984 *The Postmodern Condition: A Report on Knowledge* (trans. Bennington G, Massumi B). University of Minneapolis Press, Minneapolis, MN
- Nicholson L (ed.) 1990 *Feminism/Postmodernism*. Routledge, New York
- Weedon C 1987 *Feminist Practice and Poststructuralist Theory*. Blackwell, New York

S. Hekman

Feminist Theory: Psychoanalytic

Psychoanalytic theory, originated by Sigmund Freud and elaborated by numerous competing camps of

disciples, has been a major influence on feminist theory throughout the twentieth century. Feminists were among the earliest proponents of psychoanalysis because it acknowledged sexual desires in all people including women, who at the time were considered asexual. But over the course of the century, the theory fell out of favor among feminists, many of whom maintained that it contributed to women's oppression instead of their liberation. In the 1970s interest in psychoanalysis among feminists enjoyed a renaissance, and by the end of the century it formed part of the canon of feminist theory in the humanities. However, the theory still faces a great deal of resistance from feminists in the social sciences.

The most important aspect of psychoanalysis for feminists is its theory of sexual development. In the Freudian account, all infants are polymorph perverse, meaning that humans are not innately predisposed to heterosexual genital activity, but also capable of homosexuality, fetishism, perversions, and all other sexual variations. Freud argued that boys and girls begin life with similar psychological dispositions, but they follow different paths to adult heterosexuality owing to their anatomical differences and the expectations placed upon them by their families and society in general. Boys' development diverges from girls' around age three, during the oedipus complex, when they must relinquish their erotic longing for their mother in the face of castration threats from their father. To maintain their physical integrity, boys will typically project their aggressive fantasies onto their fathers, identify with their fathers, and then internalize the aggression imputed to the father in the form of a punitive superego. Freud wrote relatively little about girls' development, which is surprising since he treated many women patients, some of whom were the subjects of his published case studies. He did note that as beings without penises, the threat of castration was useless against girls' incestuous desires. But he maintained that the inherent superiority of the penis as a satisfying sexual organ inspired such envy among women that they are (unconsciously) compelled to seek a substitute penis of their own, typically by subordinating themselves to men in the hopes of obtaining a male offspring.

Freud's description of the childhood stages of sexual development recognized that individuals could and often did stray from the typical patterns of resolution, due to idiosyncracies of their physiology and the peculiar configurations of their family lives. When individuals did not successfully progress through the stages of development, mental illness could result. Freud maintained that unresolved early childhood traumas associated with these stages of sexual development become lodged in the unconscious and reappear in neurotic symptoms. These symptoms can be relieved through psychoanalytic therapy, a technique that involves free association, recovery of repressed traumatic memories and affects, and trans-

ference, all in the context of a uniquely structured relationship between analyst and patient. Thus, his theory provided both an account of typical sexual development, and a method for curing adult neurosis produced by deviations from those typical patterns.

This theory of sexual development appealed to many feminists who were contemporaries of Freud. Although several of Freud's women proteges, including Karen Horney, rejected his notion of penis envy in women, she and other feminist theorists applauded his emphasis on the psychological similarities between men and women. At a time when most of his notable contemporaries emphasized essential differences in men's and women's natures, Freud observed that individuals contain both masculine and feminine characteristics, and that women are as passionate as men, ideas that resonated with feminists Emma Goldman and Margaret Sanger. Feminists also took from Freud an indictment of women's proscribed roles in society. In Freud's view, marriage and family arrangements demanded an enormous and inordinate amount of repression and sacrifice on the part of women which made them hostile to civilization and often generated mental disorders. But where Freud saw few options or alternatives to these conventional practices, some feminist theorists, including Elsie Clews Parsons, saw this as a critique of Western marriage and an inspiration to experiment with alternative forms of sexual relationships.

Feminism was dormant in mid-twentieth century, but psychoanalysis at the time was experiencing a period of growth and popularity. When feminism re-emerged in the 1960s and 1970s, psychoanalysis was targeted as an enemy of women's liberation. Instead of being associated with sexual modernism, which it had represented in the 1920s, psychoanalysis had become alloyed with social conservatism. Feminist activists and theorists, including Betty Friedan, charged that psychoanalysts typically blamed strong and assertive mothers for causing mental illness in their families and a whole host of other social problems. The concept of 'penis envy' was again criticized for defining women as innately inferior to men. Therapy was seen as a tool of patriarchy for enforcing submission to sex roles and brainwashing women into accepting compulsory heterosexuality by diagnosing lesbians as sick and deranged; Mary Daly noted that 'therapist' could be read 'the/rapist,' symbolizing the attitude of psychoanalysis to women's oppression and suffering (Buhle 1998, Gardiner 1992).

In this climate of feminist antagonism to psychoanalysis, two books appeared that eventually inspired a reassessment of Freudian theory by feminists: Juliet Mitchell's *Psychoanalysis and Feminism*, published in Great Britain in 1974, and Nancy Chodorow's *Reproduction of Mothering*, published in the US in 1978. Both writers were active in the women's liberation movement, and both identified themselves as socialist feminists. They turned to psychoanalysis

because it offered a compelling explanation of the origin and the intractability of gender domination. Psychoanalysis was unique, they argued, in its ability to explain how patriarchal attitudes and practices are inscribed and internalized in both men and women. They read Freud as showing how, through the normal course of sexual development, both men and women learn to unconsciously desire and reproduce male domination and devaluation of women. In other words, they understood the theory as a description of the inner workings of a patriarchal society—not a proscription for one. Freud had in fact elided this distinction in his work, but Mitchell and Chodorow insisted that there were critical moments in psychoanalytic theory that could be emphasized to enhance its usefulness for feminists intent on changing society to benefit women.

For example, Chodorow demonstrated that the psychoanalytic theory of sexual development could account for why women typically conform to society's expectation that they bear primary responsibility for rearing children. She agreed with radical feminists who argued that this primary mothering is ordered by a sexist society that provides few options for emotionally fulfilling productive activity for women. But she used psychoanalytic theory to uncover the unconscious dynamics that caused women to willingly embrace this family role. She argued that when women do all the mothering, this sets up asymmetries in boys' and girls' sexual development that leads to gender differences in adult men's and women's ability and unconscious desire to render adequate care to infants. Girls develop their primary identity through connectedness to their mothers, whereas boys develop a sense of self as separate from their primary caretaker and all the nurturing qualities she typically embodies. Boys learn that to be 'masculine' requires that they deny and denigrate feminine qualities in themselves and in others. Thus, Chodorow demonstrated that psychoanalysis provides a compelling account of the origin of gender differentiation, helps to explain the desires and proclivities of women to nurture infants, and offers an explanation of men's general devaluation of women, without succumbing to either biological or social determinism, pitfalls that feminist critics of Freud seldom avoided. Psychoanalysis also suggests a strategy for changing these patterns—co-parenting. Chodorow argued that if men were equally involved in childcare, women would be freed to perform other creative and productive work. Also, children would learn to associate nurturing activities with both men and women, so boys would no longer denigrate women (and qualities currently associated with femininity) in the course of developing a masculine sense of self.

Chodorow's approach draws heavily from the 'objects relations' branch of psychoanalysis, which emphasizes the importance of social relationships and interactions with other people (unfortunately termed objects in the development of the self). This tradition

is the most popular one among the sociologists who advocate psychoanalysis, including Chodorow, Jessica Benjamin, and Miriam Johnson. Miriam Johnson, for example, analyzes the function of the peer group in male psychological development, focusing on its importance in directing men's aggression against women. Political theorist Jane Flax also uses objects-relations theory to challenge notions of universal reason and justice, which she regards as undergirding male domination and other forms of political inequality. The goal of these and other feminist works in the objects relations tradition is to explore the gendered psychological underpinnings and consequences of various forms of social organization with the ultimate aim of criticizing those forms and advocating social and political change.

The other major branch of psychoanalysis that feminists have appropriated is associated with the work of French psychoanalyst Jacques Lacan. This tradition is strongest in the humanities. Juliet Mitchell, who began her career as a literary critic, endorsed a Lacanian approach in her appropriation of psychoanalysis for feminism, arguing that language is an important mechanism through which socialization occurs and unconscious conflicts are expressed. Because of its emphasis on fantasy and imagination, this perspective lends itself more readily than objects-relations theory to the analysis of literature and culture. Other feminist theorists who are associated closely with the Lacanian tradition include Julia Kristeva, Jane Gallop, and Elizabeth Grosz. While all criticize Lacan as sexist, they share his view that acquisition of the culture's dominant symbolic language is the means through which male subjectivity is privileged and male domination is enforced. They also use Lacan's methods of textual analysis to uncover feminine resistance to this symbolic order through hidden expressions of women's unique sexual desire. These theorists reject the notion of a stable self or gender identity; masculinity and femininity are considered the products of discourse, not the psychological attributes of individuals (e.g., Grosz 1990).

The popularity of this perspective is reflected in the fact that familiarity with this branch of feminist psychoanalytic theory is now expected in most graduate gender programs in English, literature, and cultural studies. Psychoanalysis has been less successful at integrating the canon of gender scholarship in the social sciences, however. The perspective is virtually ignored, if not vilified in most psychology departments, and sociologists have generally been wary about embracing the theory.

Feminist social scientists have been reluctant to endorse psychoanalytic theory for a variety of reasons, some of which reflect resistances common among social scientists in general. First, the methodology of psychoanalysis—the in-depth, case study of a neurotic individual typically compiled through prolonged one-

on-one discussions with a therapist—is considered incompatible with accepted research methodologies in the social sciences. Psychoanalysis violates principles of random selection, and its emphasis on the abnormal or pathological to explain the normal patterns of development, goes against the social science logic of focusing on central tendency.

A related problem concerns the different levels of analysis in psychoanalysis and the social sciences. It is very difficult to detect patterns of group behavior based on personal, idiosyncratic information about an individual's unconscious motivations, or to infer anything about individuals' experiences based on broad-based, impersonal social forces. Feminists are not alone among social scientists in detecting this incompatibility, but it has special consequences for theorizing about gender. Since feminism begins with the premise that women as a group are disadvantaged compared to men, generalizations about gender differences are central to the project of feminist theory. But psychoanalysis focuses on individual uniqueness; the assumption is that no two people share identical unconscious needs, desires, or motives. Consequently, gender identity itself can and does have different meanings for different individuals, a point that has been emphasized by Chodorow in her later work, aptly titled *Femininities, Masculinities, Sexualities* (1994). Not coincidentally, Chodorow's movement away from large-scale sociological generalizations about gender identity followed her completion of training as a professional psychoanalyst.

The psychoanalytic focus on the unconscious and the irrational is a second obstacle to its incorporation into feminist social science. In general, social scientists trained in positivistic and quantitative methodologies have raised questions about the reliability of psychoanalytic data. In addition to this generic concern, feminist social scientists have voiced some reluctance to attribute nonrational motivations to women's behavior. For some, this theoretical approach reproduces sexist beliefs about women's reputed emotionalism and inability to act rationally. Interestingly, sociologists specializing in the relatively new field of masculinity studies have been more open to psychoanalytic theory, perhaps because the emphasis on the irrational motives of men is less worrisome to feminists (e.g., Connell 1995, Segal 1990, Williams 1995).

A third set of criticisms made of psychoanalytic theory by a number of feminist social scientists concerns the possibilities for social and political change. Many object to psychoanalytical emphasis on treating the individual neurosis without addressing the social problem that causes the illness. To the extent that feminist psychoanalytic theory has concerned itself with social change, the focus has been exclusively on changing the family, and not the economy, the political system, or other public institutions that contribute to discrimination against women. Even the changes that psychoanalytic feminists endorse have

been criticized: some feminists reject the co-parenting strategy because it ignores, if not demonizes, family patterns that deviate from the conventional heterosexual nuclear form, such as gay and lesbian families and single-parent families. Although theoretically these family types are not incompatible with feminist psychoanalysis, critics are correct that little attention has been paid to the psychodynamic consequences of being raised in these alternative family settings.

Incorporating diversity is one of the major challenges facing feminist psychoanalytic theory in the future. For the most part, feminist psychoanalysis has focused on white middle class heterosexuals, leading some feminists to reject the theory as irrelevant for understanding the lives of those suffering from the effects of racism, poverty, and heterosexism. Some efforts are being made to broaden the scope of feminist psychoanalytic theory to incorporate these experiences, but more work is needed (e.g., Abel et al. 1997).

A second challenge for feminist psychoanalytic theory is to move beyond the critique of the family as the mainstay of male domination and privilege in society. Other institutions, including schools, the popular media (especially television), and day-care centers, are increasingly important in the development of children's gender identity. Our understanding of childhood socialization in these settings could be enriched with a psychoanalytic account (e.g., Martin 1996).

Finally, feminist psychoanalytic theorists must develop the interpretive tools needed to demonstrate the operation of psychodynamic forces outside the therapeutic environment. Lacanian feminists have been successful at applying psychoanalytic concepts to the interpretation of literary texts. What is needed is a similar effort directed to understanding the unconscious motives that underlie lived engagement in the social world. Sexual harassment, workplace segregation, domestic violence, and obsession with body image are all examples of seemingly intractable social problems that call out for the kind of deeper analysis of male domination that psychoanalytic theory can provide. But this will require new psychoanalytic methodologies that will permit generalization, and a greater willingness on the part of social scientists to incorporate unconscious and irrational motives in their explanations of social behavior.

See also: Feminist Epistemology; Feminist Ethics; Feminist Political Ecology; Feminist Psychotherapy; Feminist Theology; Feminist Theory; Feminist Theory and Women of Color; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Radical Lesbian; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist

Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Gender Differences in Personality and Social Behavior; Gender Ideology: Cross-cultural Aspects; Klein, Melanie (1882–1960); Masculinities and Femininities; Psychoanalysis in Sociology; Psychological Anthropology; Theory: Sociological

Bibliography

- Abel E, Christian B, Moglen H 1997 *Female Subjects in Black and White: Race, Psychoanalysis, Feminism*. University of California Press, Berkeley, CA
- Buhle M J 1998 *Feminism and Its Discontents: A Century of Struggle with Psychoanalysis*. Harvard University Press, Cambridge, MA
- Chodorow N J 1978 *The Reproduction of Mothering: Psychoanalysis and the Sociology of Gender*. University of California Press, Berkeley, CA
- Chodorow N J 1994 *Femininities, Masculinities, Sexualities: Freud and Beyond*. University Press of Kentucky, Lexington, KY
- Connell R W 1995 *Masculinities*. University of California Press, Berkeley, CA
- Daly M 1978 *Gyn/Ecology: The Metaethics of Radical Feminism*. Beacon Press, Boston
- Flax J 1993 *Disputed Subjects: Essays on Psychoanalysis, Politics and Philosophy*. Routledge, New York
- Gardiner J K 1992 Psychoanalysis and feminism: A humanist's view. *Signs* 17(2): 437–54
- Grosz E 1990 *Jacques Lacan: A Feminist Introduction*. Routledge, New York
- Johnson M M 1988 *Strong Mothers, Weak Wives: The Search for Gender Equality*. University of California Press, Berkeley, CA
- Martin K A 1996 *Puberty, Sexuality, and the Self: Girls and Boys at Adolescence*. Routledge, New York
- Mitchell J 1974 *Psychoanalysis and Feminism*, 1st edn. Pantheon, New York
- Segal L 1990 *Slow Motion: Changing Masculinities, Changing Men*. Rutgers University Press, New Brunswick, NJ
- Williams C L 1995 *Still a Man's World: Men Who Do 'Women's Work.'* University of California Press, Berkeley CA

C. Williams

Feminist Theory: Radical Lesbian

Perhaps what most obviously differentiates mainstream liberal feminism from her more radical sisters is that while the former continues to hope that the master's tools can be deployed to dismantle and then rebuild a better version of the master's house, the latter has largely quit this project and jettisoned its tools in favor of building not merely a house but an entire conceptual village whose collective mission it is to embody the experiences, values, and desires of

women. That is, while the primary focus of mainstream liberal feminist theory has been the critical evaluation of patriarchal or male-dominated institutions such as government, medicine, family, and religion in the interest of retooling these institutions to affect greater equality or inclusion for women within them, radical feminism takes its point of departure to be the position that any thoroughgoing critique of patriarchy must include an analysis of the tools themselves.

How, in other words, do concepts like 'woman,' 'man,' 'body,' and 'mind' inform the maintenance of these institutions? How have we defined notions like 'the rational,' 'the desirable,' 'the good,' 'the moral,' and 'the beautiful,' such that being born male virtually guarantees greater access to opportunities and resources? What are the conceptual and ideological anatomies of patriarchies as we find them across culture and history? How do patriarchal ideologies help to maintain patriarchal institutions?

Perhaps, however, the central issue which gave birth to contemporary radical feminism in the late 1960s and early 1970s was the emerging recognition that it could never be enough to agitate for an equality of inclusion within a conceptual framework which defines women's value only in reference to men. The liberal feminist project cannot fail to risk the self-defeat that accrues to retaining a conceptual framework—the master's tools—which systematically inferiorizes women. The merely cosmetic improvement produced in some women's lives (mostly Caucasian, middle-class, and educated like their male counterparts) inevitably fails to address the very real life and death issues which confront many women both locally and globally. Equipped with this insight, radical feminists undertook the conceptual as well as political quest to discover, for example, how the inferiorizing of women is accomplished, and how the domination of women varies with respect to class, ethnicity, culture, ability, age, sexual orientation, and institution. As late feminist philosopher Audra Lorde put it, the master's tools cannot be used to dismantle the master's house. The revolutionary change necessary to produce lasting improvement in all women's lives requires that we fashion our own tools.

Some radical feminists argue, for example, that the notion of equality is itself tethered to a view of human nature which so privileges attributes traditionally associated with men such as aggression, hierarchical forms of organization, rugged individualism, competition, and ethical systems that strongly favor impersonal rule-following over relationality, that women are only able to be included within such conceptual frameworks as honorary men. It is little wonder, they argue, that examples offered to young women by liberal feminists as role models and heroines are modeled after male heroes; the accomplishments of women like Harriet Tubman, or Marie Curie (or movie heroines like those found in *Terminator II*,

Aliens, or *Femme Nikita*) are valued according to a standard that rewards domination over domesticity, conquest over caring, and militarism over mothering, in short masculinity over femininity. This is not to imply that these are not examples of very real valor, but that what is rewarded in them goes further to reinforce a male-centered or masculinist vision of the good than it does to foster a vision centered around values identified as feminine such as cooperation, compromise, collective decision-making, or compassion. It is thus simply not enough, argue radical feminists, for women to settle for reformist programs that aim at equality. For given the conceptual constraints that determine value, such programs will either fail or, even if they succeed, will do so only at the cost of continuing to devalue qualities associated with femininity, and hence with women.

Liberal feminism does not, in other words, necessarily represent the desires or interests of all women, whether we take these desires to be culturally constructed or endowed by nature. Here, however, we arrive at another important juncture in radical feminist theory: if what conceptual analysis shows is that the ways in which we conceive of gender operate consistently to privilege men, ought we to consider these ways a reflection of natural propensity, and then aim to elevate the value of feminine characteristics like nurturing, emotional responsiveness, and the tendency to focus on relationships over abstract autonomy? Or, alternatively, should we eschew traditional femininity as the construction of patriarchal culture, arguing that there are no *essential* qualities that importantly distinguish men from women? If so, at least two other possibilities present themselves: One meaning of *androgyny* favors the promotion of the masculine and the feminine in all of us, acknowledging that while gender may be constructed, such qualities may still be salvageable reinterpreted in a nonsexist context. Another meaning of *androgyny*, however, suggests that we ought to abandon notions of femininity and masculinity altogether as unworkable in favor some other notion of personhood. Here the question that arises concerns what such a notion would look like, and whether it is possible to envision a conception of human being that can comprehend biological differences without merely reproducing an atomistic individual so divorced from the bodily that it becomes subject to the same criticism radical feminists level at liberal feminism. That is, can we envision a concept of 'embodied person' which, irreducible to the determinations of either biology or culture, is also rich enough to account for differences relevant to and affected by both?

Whatever the resolution to the essentialism/anti-essentialism debate among radical feminists, their main critique of liberal feminism remains unaffected, namely, that what the history of patriarchy demonstrates is that the oppression of women is sanctioned through the denigration of whatever qualities are

identified with femininity. Whether we seek liberation through the elevation or the abolition of femininity what remains clear is that reformist programs for inclusion can never by themselves be adequate. We have not, however, exhausted our avenues of critique, for there remains another possibility: Marxist/socialist feminism whose primary aim it is to show how, given the marriage of patriarchy to *capitalism*, it is illuminating to conceive of women not merely as a group but as a *class* whose oppression is maintained first and foremost through economic means. Combined with a radical feminist analysis of patriarchal concepts of motherhood and domestic labor, a Marxist/socialist account shows how divisions of labor are themselves typically sexed in at least two ways: first, because women's *labor qua* procreation (the reproduction of human life itself) is assigned to the 'natural' and hence is not counted as labor at all, and second, because the domestic labor performed almost exclusively by women, (the reproduction of the conditions of life) has itself become naturalized in a capitalist economy that identifies all labor as *paid* or *wage* labor; that is, if it is not paid, it is not labor. Cast as the vehicle for the reproduction of both life and the means to life, 'woman' appears a paradox. 'She' is unexploitable in the sense that her labor is not counted as labor within a capitalist conceptual framework at all, yet 'she' also epitomizes the very prototype of exploitability in that 'her' labors are systematically and maximally undercompensated.

Again, the valuable point can be made without entering the essentialism/anti-essentialism debate: regardless what qualities we assign to nature or nurture, *that* women's oppression is intimately linked to the relationship between her economic status and her status as provider of hearth, home, and progeny *for men* well supports the radical feminist claim that there exist no social institutions in which men do not have a controlling interest, including economy and family. What recommends this approach over liberal feminism, then, is the emphasis on the role of economic class, for whereas the best that liberal feminists can hope for is an equality of competition between men and women within capitalism, the radical Marxist/socialist feminists are free to imagine a world within which neither sex nor class oppression are operative; radical feminist theory is thus utopian in a way that liberal feminism cannot be. However, though no doubt radical Marxist/socialist feminism goes further than liberal feminism or than radical feminism alone, even it ultimately fails the project of genuine emancipation. For while such a critique of capitalism shows something important about the specifically sexed parameters of economic oppression, there remains considerable work to be done to show the relationship between economic and other forms of oppression such as sex, sexual orientation, race, age, ability, and so on.

Given the need for this work, many radical feminists distinguish themselves from both mainstream liberal

feminism and Marxist/socialist feminism by taking a more skeptical, self-reflective approach to theorizing, arguing that it is *at least* an open question whether any patriarchal institution is salvageable in any form. There is little more revolutionary a declaration than that capitalism must be overthrown, but if this is not accompanied by the overthrow of other patriarchal institutions, the status of women could remain unchanged or worsened as men's conditions improve. What revolutionary changes would institutions like the family be required to undergo to achieve equality not only with respect to gender, but with respect to any number of other axes of oppression, including race, ability, economic class, indigenous culture status, age, sexual orientation, and gender expression?

Some radical feminists such as Mary Daly argue that patriarchy must be overthrown or abandoned in favor of *separatist* or women-only cultures and societies. In her 'radical elemental feminist manifesto,' *Quintessence*, Daly describes the fundamental ideology of patriarchal institutions as *rapist*, 'characterized by invasion, violation, degradation, objectification, and destruction of women and nature; the fundamental paradigm of sexism, racism, classism, speciesism, and all other oppressive-isms' (1998, p. 8). Not unlike the radical Marxist/socialist feminist's view that the sexed division of labor lies at the root of economic exploitation, Daly argues that at the root of all forms of oppression lies that sexual domination—the male prerogative to rape—on the basis of which degradation and objectification, in short, *enslavement*, becomes the naturalized condition of virtually all nonmale, nonwhite, nonheterosexual, non-Western, nonhuman, nonaffluent, persons.

If Daly is correct, (a) all forms of oppression and exploitation take sex as their essential archetype. Hence, (b) feminism's emancipatory objectives cannot be achieved without serious attention paid to the ways in which sexism informs the other 'isms' which sustain a patriarchy that (c) is malignant in all of its past and present forms, and whose beneficiaries, primarily Caucasian, affluent, heterosexual males continue to have a stake. Finally, (d) to whatever extent sex oppression is archetypal of all oppression so too must heterosexuality become *compulsory*, that is, a characteristic of the *heteropatriarchal* way we define sex and hence sexuality such that other sexualities are either completely occluded or, when acknowledged at all, identified with the perverse and unnatural. Within compulsory heterosexuality, an action only counts as sex if (a) it involves a penis-penetrating-a-vagina, (b) it is compulsory in the sense that no other option counts as an acceptable choice, and (c) it negatively associates abstinence with frigidity, mental illness, or homosexuality. On this view, then, the domination of sex cannot be accomplished without the domination of sexuality. Given the combined compulsion of (a)–(c), moreover, it is little wonder that Daly characterizes heteropatriarchy as *rapist* urging the view that re-

sistance to patriarchy must be accompanied by resistance to heterosexuality. Radical feminism thus becomes radical *lesbian* feminism as both the rejection of male sexual access and as the radical embrace of the value, lives, and bodies of *women*. Such is the meaning of dismantling the master's house with *new tools*. That is, radical lesbian feminists (or radical feminist lesbians) argue for two distinct but intimately related claims: first that compulsory heterosexuality must be repudiated, and second, that lesbianism must itself be theorized as a positive, life-affirming, morally responsible choice.

One potentially compelling objection to this view is that it is surely possible to embrace the political goals of lesbian feminism without jettisoning sexual relationships with sympathetic men. However, because even the best of men are still disproportionately empowered in heterosexual relationships such intercourse cannot fail to remain compulsory, and hence self-defeating. This is not to say that heterosexual sex either is or is not *inherently* harmful to women, but rather that because male sexual access can be identified as one of the primary sources of women's oppression, cauterizing this source of harm is instrumental to ending heteropatriarchal power. Indeed, whether heterosexual sex is inherently harmful or not is largely beside the point given that from a radical lesbian feminist point of view it is hard to imagine conditions under which the potential for harm could be sufficiently deflated, given men's history of sexual and domestic violence, to make the risk worth taking. Nothing, moreover, requires us to theorize men's 'essential' natures as aggressive or women's as passive; it is enough to simply recognize that history and culture are so saturated with the glorification of male sexual conquest and violence, that the prospects for change, and the likelihood of that change being initiated by men is less than compelling. Not unlike Marx's conception of the permanent revolution of the working class, radical lesbian sexual separatism constitutes much more than a political strategy; it constitutes part of the bedrock of a burgeoning *lesbian culture* in the sense that what propels this revolution is a *positive choice to live one's life among women*, not to live among women until men learn to behave better. To opt merely for the latter is tantamount to abandoning in advance the worth of having chosen the former at all.

Philosophers such as Sarah Hoagland, for example, argue that a constitutive feature of a coherent lesbian ethic includes the repudiation of all forms of domination and subordination. To the extent, then, that heterosexuality is entrenched in and defined by these values, it can find few comfortable bedfellows—literally or figuratively—among radical lesbian feminists. Is a nonoppressive, nondominating heterosexuality possible? In theory, no doubt it is. But, as feminist philosopher Marilyn Frye points out, the practical likelihood of this given the stakes men have and are likely to continue to have in heteropatriarchy

seems illusive at best. Lesbians, she argues, are simply impossible within the heteropatriarchal conceptual framework; indeed, they are like 'arm wrestling ducks.' That is, no arms, no wrestling, no penis, no sex. Hence while gay men can be condemned for having perverse sex, lesbians cannot have sex at all, and without the possibility of sex, there can be no lesbians. As a clever male college student was once heard to put the point: 'There can be no arm-wrestling ducks, but a good arm-wrestling fuck, now that's sex.' Frye goes on to argue that given the biological facts of reproduction as well as the social and political expectations attached to child-rearing, it is hardly a wonder that exercising control over the conditions under which pregnancy, birth, and lactation take place becomes imperative for men who are likely to experience the termination of privilege as the unjust withdrawal of something owed. As Frederick Engels had shown more than a century earlier, men's investment in sexual domination is rooted first and foremost in their investment in the products of their own (hetero)sexuality, namely, the offspring who represent a paradigmatic form of property, a reflection of self-interest, and the potential for future inheritance. To deny men sexual access, then, is in effect to deny them the opportunity to control the conditions of women's reproductive lives, and hence of women's lives in general, particularly in nations or regions where access to contraceptives and abortion is heavily regulated if not banned by heteropatriarchal laws, mores, and religion.

The affirmation of women's lives, bodies, values, and choices forms the second critical axis of any radical lesbian feminist theory. Yet, while this affirmation has often been theorized as necessitating some variety of separation from men—*separatism*—it is less clear what exactly this might mean, for whom it is realistically practicable, and to what extent it requires physical and material independence from both men and patriarchal institutions. This much, however, seems clear: the choice to no longer engage with men sexually constitutes a minimal criterion of separatism whereas the choice to embrace a life composed primarily or even exclusively of 'women-loving-women' goes much further—and is also far more difficult to realize in practice. There are several reasons why this is the case: first, because we live in a capitalist economy, regionally, nationally, and globally, where-withal *is* economic wherewithal. Hence, the creation of any women-only space, permanently (a communal farm, for instance) or temporarily (such as women's music festivals) is an economic undertaking which, in virtue of this, is always at risk of excluding some in favor of those who can afford the price of separatism. As some lesbian feminist women-of-color point out, separatism continues to be a luxury of whiteness to whatever extent economic affluence does. Second, because every institution, economic, social, cultural, medical, to which we have access is imbued with heteropatriarchal values, it remains somewhat obscure

what separatism could mean short of wholesale abandonment. Third, short of the invention of parthenogenesis, a contribution of male-produced sperm is still a requisite of human reproduction.

Such a characterization of separatism may, however, be guilty of a straw fallacy, that is, perhaps separatism appears impracticable only because its parameters have been defined too narrowly or too simplistically. Whether this is so, however, depends precisely on what kind of conceptual work such a notion is supposed to be able to do, for whom, and for what duration. The range of other potentially separatist choices is very wide and includes at least the following: the formation of one's primary affective relationships, sexual or otherwise, with women and girls only, choosing a woman-centered spirituality, adopting and rearing girls within a lesbian culture, creating woman-centered works of art, dance, music, poetry, comedy, film, philosophy, and literature, rejecting other male-centered practices such as non-human animal flesh consumption, hunting, or other hierarchical forms of organization, choosing political affiliations and activities that are consistent with a woman-centered life. The difficulty here, of course, is twofold: first, how ought we to define what counts as woman-centered or lesbian (and do these mean the same thing?)? How do we avoid reproducing the same forms of oppression which so disable us within heteropatriarchy? (Should preoperative transsexual men-to-women be allowed into women-only spaces? Can a woman-centered life include rearing sons? Should radical lesbian feminists form political alliances with gay men? If so, should we sign on to their political causes such as the promotion of legalized marriage for gays and lesbians?)

Second, if 'separatist' is itself subject to evolving cultural, ethnic, and historical interpretation, surely it is as dangerous to treat it ahistorically as it is to treat 'woman,' 'man,' and so on. What, then, does separatism require beyond the refusal of sexual access to men? Does anyone's refusal matter more than others? Can lifelong lesbians who do not identify as feminists be separatists? Can women who have left long-standing heterosexual relationships and now identify as lesbian qualify as separatists in mid-life? How separate is separate? The point, of course, is that many questions attend the endeavor to create new tools, and much more attention needs to be paid to the ways in which oppressions related not only to sex, but to race, ability, age, gender expression, and class intersect with compulsory heterosexuality. There is much to worry and much to celebrate. Nonetheless, if the radical lesbian feminist critique of capitalist heteropatriarchy is correct, this endeavor is more than just important; it is necessary to the emancipation of women.

See also: Evolutionism, Including Social Darwinism; Feminist Ethics; Feminist Theology; Feminist Theory;

Feminist Theory and Women of Color; Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology

Bibliography

- Allen J (ed.) 1986 *Lesbian Philosophy: Explorations*. Institute for Lesbian Studies, Palo Alto, CA
- Card C (ed.) 1991 *Feminist Ethics*. University of Kansas Press, Lawrence, KS
- Daly M 1998 *Quintessence: A Radical Feminist Manifesto*. Beacon Press, Boston
- Frye M 1983 *The Politics of Reality: Essays in Feminist Theory*. Crossing Press, Freedom, CA
- Hartmann H 1990 Capitalism, patriarchy, and job segregation by sex. In: Hansen K V, Phillipson I J (eds.) *Women, Class, and The Feminist Imagination*. Temple University Press, Philadelphia, PA
- Hoagland S 1988 *Lesbian Ethics: Toward New Value*. Institute of Lesbian Studies, Palo Alto, CA
- hooks b 2000 *Feminist Theory From Margin to Center*. South End Press, Cambridge, MA
- Lee-Lampshire W 1999 Spilling all over the wide fields of our passions. *Hypatia* 14(3): 1-16
- Lorde A 1984 *Uses of the Erotic: The Erotic as Power*. In: Lorde A (ed.) *Sister Outsider*. Crossing Press, Trumansburg, NY
- Lugones M 1987 Playfulness, world-traveling, and loving perception. *Hypatia* 2(2): 3-19
- Rich A 1994 Compulsory heterosexuality and lesbian existence. In: Jaggar A (ed.) *Living with Contradictions*. Westview Press, Boulder, CO

W. L. Lee

Fertility and Culture: Anthropological Insights

In its narrowest meaning, fertility refers to the production of a living child that adds one person to a population. Between this biological event and the level of culture, which defines and shapes the events that produce a child, lies an engaging set of issues. The mainstream of anthropology has never been interested in the question of numbers of children in the way that demography has, in its focus on the stocks and flows of population. Conversely, the anthropological penchant for rich, contextualized cultural meaning draws limited interest from demographers seeking parsimonious ways to explain fertility levels. Still, the

two disciplines are finding increasing common ground in the domain of reproduction. This article describes anthropological attempts to derive cultural insight in studies of fertility.

Demography seeks to identify and measure key variables related to population and to demonstrate relations among them. But to frame its questions and explain the results, it has tended in the past to draw on other disciplines, particularly on sociology, economics, and history. In recent years, demographic interests have turned as well to anthropology and to its methods. There have been two principal reasons for this growth of interest. The first was the impact of the Princeton European Fertility Project (Coale and Watkins 1986). The most important finding of this project was that the decline in marital fertility in Europe was explained less by the expected economic factors of income and level of development than by factors such as religion, dialect, ethnicity, and region—factors the project termed ‘cultural,’ although the awkwardness of the nomenclature was recognized. In Belgium, for example, villages experienced fertility decline in consonance with linguistic groupings. Wall-oon villages tended to experience the transition to lower fertility earlier than did Flemish villages, irrespective of their locations or economic characteristics (Lesthaeghe 1977). The implication drawn was that information diffused more readily within culturally or linguistically similar regions. The second reason why demography has turned to anthropology has been an increasing interest in patterns of fertility and mortality in developing countries. Questions were emerging particularly about the social supports for high fertility and about such tremendous variation in reproductive practices that researchers trained largely in the Western quantitative disciplines felt poorly equipped to handle. Because anthropology is the discipline that has so routinely attempted to study non-Western societies, anthropological works have drawn increasing demographic attention.

1. What is Anthropology, and How Has it Been Applied to Demographic Studies of Fertility?

Anthropology seeks to understand why people behave the way they do and to understand their beliefs and values within particular socioeconomic contexts. Anthropological work begins with the premise that the cultural frames from which our understandings of human motivation and action spring are likely partial, if not altogether wrong. They are therefore misleading. This is true, no matter how familiar the culture appears to be to the researcher. In fact, the more familiar the culture appears to be, the greater the risks of being misled and the more wary the analyst should be. To a field whose practitioners start with the expectation of being wrong, most valued are data that embody local people’s own ideas about the world, in contrast to

those based only on the researcher’s predetermined constructs. To this end, open-ended approaches are critical. A key goal is to achieve a subjective understanding: that is, to comprehend how local people’s experience the world and the logic that may underlie their choices. Doing so, it is felt, should lead not simply to better insights into the cultural beliefs of people living in remote places, but to a more encompassing view of reality and hence to better theory.

Whereas anthropology was an exotic oddity 15 or 20 years ago in demographic circles, demographic studies increasingly integrate methods and concepts conventionally associated with anthropology into their enterprise. Many demographic projects have begun to supplement their survey analyses with field projects, which include an anthropologist or individuals trained in ethnographic methods and qualitative methods are increasingly common in demographic project designs. Even the international Demographic and Health Surveys (the DHS) now have a ‘focus group’ component in which statements and explanations, rather than simply numeric and short-answer responses, can be elicited from groups of people.

1.1 Operationalizing Cultural Traits in Statistical Taxonomies

What anthropology has had to offer the demographic enterprise has been largely a function of the state of its own field. When significant demographic interest first turned to anthropology in the middle of the twentieth century, a particular methodological approach, sometimes called the cross-cultural method, dominated American anthropology. This was a kind of statistical taxonomy in which societies were described by traits that were typed and ‘operationalized’—that is, cast in the form of discrete variables—and coded in order to test quantitative hypotheses and look for cross-cultural patterns. The primary data used in this endeavor was the Human Relations Area Files (Murdock 1975), which contained detailed codes of the organizational features of hundreds of ethnic groups. Prevalent as well in American anthropology at this time were questions about the relationship of culture to personality from Freudian perspectives. A number of anthropologists sought to combine this theoretical perspective with the cross-cultural methodology. Whiting et al. (1958), for example argued from HRAF data that life under harsh environmental conditions required parents to protect a child by a long period of breastfeeding and a ban on sexual relations. Such practices, because they led to strong psychological attachments of the child to the mother, were said to necessitate dramatic manhood initiation rituals for boys, to separate them from their mothers and to foster the development of their masculinity. While personality development drew little demographic interest, there was considerable interest in the potentials of such studies to shed light on facts such as the

duration of breastfeeding and postpartum sexual abstinence, which had more direct connections to fertility. Applying this kind of analysis to data on kinship and economic subsistence, demographer Frank Lorimer (1954) argued that populations like those on small islands, with highly constrained and intensively utilized resources, developed types of social organization and customs that checked family size. In Polynesia, he maintained, such customs included coitus interruptus, abortion, infanticide banning further reproduction to the mother of a married son, and encouraging sea voyaging for young men.

1.2 Culture as Ideas and Mental Schemes

The taxonomic approach drew considerable interest in demography. Yet within anthropology itself, it encountered increasing antipathy. Critics charged that the approach, in extracting practices from the contexts in which they drew meaning and treating them as a checklist of discrete, modular traits, effectively froze what was in reality a multifaceted and fluid entity. Particularly with the impact of Clifford Geertz (1973), anthropology's conception of culture shifted decisively away from such approaches and moved toward one of shared mental schemes or symbols that signaled meanings and interpretations of the world. By contrast to efforts to operationalize cultural data in order to break them down into pieces and assimilate them to our own analytical categories, most anthropologists now have the opposite purpose. They stress that culture cannot be broken down into separable pieces. Such pieces would in any case have no meaning apart from a context. Anthropologists also take cultural oddities as possible signs of some facet of a reality that the analyst has not seen and hence has not coded into his or her own cultural ideas.

Anthropologist's efforts to reclaim culture from the domain of operationalizable concepts have not always been easy to explain to other disciplines. The work of demographer John Caldwell, however, took a notable stand in its efforts to make culture figure more explicitly as ideas or attitudes. The most notable example is his description of the proliferation of Westernization: the rise of new attitudes about work, family relations, and fertility practices, particularly through schooling. By comparison to traditional values of authoritarian family structures, a distant and severe father, and marriage as an arrangement between elders, Caldwell (1980) describes infant and primary school textbooks in three African countries as attempting to convey messages of parental closeness to children: kindly, understanding fathers and love-match marriages. So powerful are such ideas that with the shift to mass education, all members of the population, including those who have not attended school, may be affected by the changes generated by a new cultural climate of an educated populace. These ideational changes, he argues, have facilitated a shift

to nuclearization of families and to smaller family size goals.

The evolutionary implications associated with traditional/modern frameworks have fallen out of favor with sociocultural anthropology. Contemporary anthropologists reject any necessary assignment of labels of traditionalism or modernity, and they recognize all societies as inherently fluid and changing. Nonetheless, the move toward seeing culture as a set of shared mental schemes that can shed light on ideas and behaviors has opened up significant possibilities for understanding demographic phenomena.

An engaging example that attempts to link forms of belief to fertility behavior is *Mussallam's* (1983) account of the allowable uses of contraception for men in classical Islam. Islamic practice has often been cast by analysts as opposing the use of contraceptives. *Mussallam*, however, describes a more complex picture. Although contraception was seen as an act that fell short of religious perfection for a man, Islamic practice allowed a wide range of exceptions. Contraception was allowed to preserve health and well-being. A man could contracept to protect his wife from the dangers of childbirth, especially if she were young or sickly, or simply to protect her health and beauty from the wearing effects of childbearing. He could contracept as well to protect a young nursing child from the nutritional competition posed by a new pregnancy. Another salient domain of concern was economic welfare. A man who found himself tempted to engage in theft to meet his economic responsibilities to his children could practice contraception to limit the number of his offspring to a manageable size. Analogous concerns emerged in the desire to safeguard human property. Because a man could not sell a slave after she became the mother to his child, using contraception was seen as a lawful measure to retain liquidity of slave assets. A man could use contraception as well when having sex with another man's slave to prevent his own children from being born into slavery. Related to this was the need to safeguard the welfare of the Islamic community under conditions of warfare. If a man had sex with a woman behind enemy lines, he risked not only losing his own children by procreating but swelling the numbers of the infidels unless he took contraceptive precautions.

1.3 Culture in Flux; Culture in Contestation

Despite the intrinsic appeal of pursuing cultural pathways of logic like this, it is risky to assume that behavior reflects cultural beliefs in straightforward ways. Most contemporary anthropologists hold that people have multiple cultural templates, and that culture is a fluid set of interpretive resources that people can draw on in particular political and historical contexts. Even a body of religious canon, as powerful as its injunctions may seem, has the potential for numerous interpretations.

Among the most lively examples of new research in anthropology that demands a fluid, resourceful vision of culture is the topic of male fertility. Fertility research is by far the largest sub-discipline in contemporary demography. Yet until a very few years ago, women comprised virtually its sole objects of study. Men, if they appeared at all, were shadows, partners-by-implication of those engaged in childbearing. Even if genetic tests could verify the relationships among household members in a survey, this would hardly begin to solve the challenges surrounding the analysis of male fertility. While some men do not know how many children they have or deny connection to offspring born outside of wedlock, other men claim to have fathered children they have never seen. The fact that men so often redefine parenthood can work to the disadvantage of children, but also to their advantage. For example, Townsend (2000) shows that in rural Botswana, where the demands of migrant labor are so taxing, a young man who lives full time with his children is an oddity. At the same time, men provide considerable support for children, even though it is often not their own children but those of other men, whether these are uncles, grandfathers, or distant cousins. Thus, despite the widespread perception in the current international population policy world of massive male disengagement from parenthood responsibilities, Western cultural preoccupations with a biological definition of paternity have masked some critical variations in the cultural assignment of parenthood.

One of the most important points of debate in current socio-cultural theory that is now being brought to demography is the question of societal forces of politics and economics versus individual agency. That is, to what extent are people's choices limited by wider powers beyond their control? Can people, on the other hand, make their own decisions about reproductive actions by resisting or evading state powers? The increasing answer to this question is that both forces are invariably at work, and that cultural ideologies are often at the center of the contestation. Greenhalgh and Li's account (1995) of the heavy-handed Chinese state policy of one child per couple that was implemented in 1979 reveals a dynamic of mutual influence. Greenhalgh and Li show that women began to resist the formal one-child policy, strategizing to have two children, especially if the first was a girl. Women attempted to adopt out daughters, move to another city to give birth to a second child, or remove state-imposed intrauterine devices. They might also attempt to conceal unauthorized pregnancies until it was too late for a state-required abortion, or (as ultra-sound technology spread) engage in sex-selective abortions. By 1991, the state, faced with such resistance, was forced to rewrite policy. It 'encouraged' one child for all and prohibited third children, yet allowed two children if the first was a girl. In sum, individual women may have been liberating themselves indi-

vidually from unwanted daughters, yet their strategies of resistance effectively reinforced the state's gender hierarchy.

The fact that culture is increasingly seen in anthropology as a resource that people may put to use is best examined in a context in which reproductive volition has been seen as low. For this reason, some of the best examples are found in topics related to high fertility: a pattern that has been seen in demographic convention as an outcome of biology and custom and not as the calculus of conscious choice (Coale 1973). In populations said to have what has been called natural fertility (Henry 1961), the demand for children is high, and the number of children women produce is assumed to be limited only by how quickly they can be supplied, given the biological rhythms resulting from specific customary regimes. In sub-Saharan Africa, for example, most people continue to report on surveys that they want 'all the children God gives them,' and typical reactions to suggestions to reduce fertility range from polite interest to outrage. By contrast, in societies where desires for large numbers of children have abated, couples are said to engage in parity-specific behaviors, especially the use of modern contraceptives, in order to control completed family size.

However, ethnographic accounts of practices such as infanticide, child fosterage, child abandonment, adoption, and abortion make it clear that people in all societies exert some form of control over reproduction. Whereas Western models assume that reducing fertility is the best way to address economic troubles, families in many African societies see the diverse skills and social ties that they manage to cultivate through children as the key to their future capacity to cope with economic and political adversity. Adults treat the original biological event of parenthood not as a secure hold on a child's future support, but as the beginning of a long, continuously negotiated relationship in which they try to influence children's obligations. Young children are commonly fostered to 'grannies.' Later, when they are older, they may be sent to urban households that offer training in specialized apprenticeships or a chance to attend a respected school. In short, rather than trying to reduce child numbers, many African families would find this a drastic solution to economic problems. They prefer to have a number of children, cultivating differentiated opportunities for them (Bledsoe 1994). Yet although the effects of certain practices on cutting reproductive costs may be analytically discernable, they do not necessarily constitute people's motivations for taking actions that have effects on fertility or household composition. Most anthropologists insist that it is wrong to reduce family structure to economic or demographic terms alone.

Another example of people's efforts to exert control over childbearing appears in the cultural reinterpretation of contraceptive technologies in sub-Saharan Africa (Bledsoe 1994). In Western cultural per-

ceptions, contraceptives are devices that limit fertility, and people who do not wish to limit births would be unlikely users of devices that would do so. In West Africa, however, people see contraceptives more as devices to ensure the survival of many children by spacing their births carefully than to limit births. Some of the most common contraceptive users are thus breastfeeding women who want to avoid overlapping children, one in the womb and the other nursing, but who have resumed their menses before the breastfeeding child can be weaned. As this suggests, families exert prodigious efforts to control the timing and circumstances of children's births, though, again, not necessarily to limit numbers of children.

1.4 Reconciling Local Cultural Knowledge and Western Science

Pursuing threads of counterintuitive cultural logic that are opened up by scrutinizing demographic patterns can yield not simply observations about culture but empirical truths that have been lost to contemporary scientific view. An example comes again from the use of contraception in West Africa (Bledsoe 2001).

Contemporary Western writings, both scientific and humanistic, depict the ability to bear children as one limited by time. That is, Western culture posits a time-dependent ability to conceive as the key to achieving fertility. Accordingly, the number of children a woman may bear is calculated by dividing the total amount of time she can conceive by the interval required for each pregnancy. The use of contraception plays directly into this presumed calculus. To the extent that the use of contraception or abstinence blocks off time in this potential reproductive span, a woman's fertility will remain low. In the African cultural view, however, the ability to reproduce is limited not by time but by bodily resources. Among the forces that drain bodily resources and erode the ability to reproduce, childbirth is by far the most powerful. Each birth takes a mounting toll on a woman's physical reserves, cumulatively 'aging' her body (or wearing it out, in many local translations), and eventually making further childbearing dangerous, irrespective of her chronological age. In this vision, aging is not the same thing as age. Although aging tends to be equated with the passage of time in Western analyses, its pace, in the African view, is contingent on the physical ordeals the body suffers. Difficult and/or closely spaced births, combined with a poor diet, illness, and strenuous work responsibilities, can exact a disproportionate toll. In such situations, a woman who is trying to have all the children God gives her will almost certainly run up against the anatomical and energy limits of the body before she runs into any temporal boundaries on the ability to conceive. Because contraceptives allow women more opportunity to recuperate between births, and because reproductive capacity is seen as governed by taxing events rather than by time, it is

clear why contraceptives can be used to produce more, rather than fewer, children and even to stem the pace of aging.

The African vision of contingent wear and cumulative physical toll over a reproductive lifetime might appear to be 'just common sense.' It is more accurate, however, to say that our increasing personal distance from the high-and-risky end of the fertility horizon, combined with our unprecedented levels of preventive care and copious backup in case of emergencies, has sharply truncated our vision of the tolls that reproduction can take under conditions that we no longer experience or fully understand.

2. Future Possibilities for Anthropology and Demography in Studies of Reproduction

In the evolving relationship between anthropology and demography, the spectrum of topics continues to broaden. Common interests can be found in the politics of fertility control, gender relations, the negotiation of fertility decisions among spouses, and the re-opening to scrutiny of taken-for-granted associations between fertility and variables such as educational achievement. Anthropology's current interests in the new reproductive technologies holds the potential for studies of the timing and circumstances of entry into parenthood. Finally, it can be noted that although anthropology made itself as a discipline on the study of exotic people living in distant places, increasing numbers of practitioners are coming 'back home,' studying social and cultural phenomena in the urban United States and Europe. This trend may facilitate fresh collaborations with demographers on studies of reproduction in Western societies. Some of the most productive areas of common ground in studies of reproduction for the two disciplines may thus lie in studies of migration. New anthropological interests in global flows of people, ideas, and technologies are coming at the time when demographers are taking up the challenges of studying international population movements and refugees. The opening of the European Union and, in the United States, the removal of national-origin quotas for immigrants has spawned migration across national borders on an unprecedented scale. Such a mix of national ideologies and policies of social support that must underpin these dynamics as people envision a reproductive career invoke are virtually uncharted areas. It will no longer be possible to rest on conventional notions either of a population boundary or of local cultural practices. Clearly, in any case, anthropology's style of engagement with demography and the insights to be gained from such mergers will continue to change as the fields themselves change.

See also: Family Theory and the Realities of Childbearing Behavior; Fertility: Institutional and

Political Approaches; Fertility: Political and Political-Economic Perspectives; Fertility: Proximate Determinants

Bibliography

- Bledsoe C 1994 Children are like young bamboo trees: Potentiality and reproduction in sub-Saharan Africa. In: Kiessling K L, Landberg H (eds.) *Population, Economic Development, and the Environment*. Oxford University Press, New York, pp. 105–38
- Bledsoe C 2001 *The Contingent Life Course: African Challenges to the Culture of Western Science*. University of Chicago Press
- Caldwell J C 1980 Mass education as a determinant of the timing of fertility decline. *Population and Development Review* 6(2): 225–55
- Coale A J 1973 The demographic transition reconsidered. In: *International Union for the Scientific Study of Population* (ed.). *International Population Conference, Liege IUSSP, Liege*, Vol. 1, 53–72
- Coale A J, Watkins S C (eds.) 1986 *The Decline of Fertility in Europe*. Princeton University Press, Princeton, NJ
- Geertz C 1973 *The Interpretation of Cultures: Selected Essays by Clifford Geertz*. Basic Books, New York
- Greenhalgh S, Li J L 1995 Engendering reproductive policy and practice in peasant China: For a feminist demography of reproduction. *Signs: Journal of Women in Culture and Society* 20(3): 601–41
- Henry L 1961 Some data on natural fertility. *Eugenics Quarterly* 8: 81–91
- Lesthaeghe R J 1977 *The Decline of Belgian Fertility, 1800–1970*. Princeton University Press, Princeton, NJ
- Lorimer F 1954 *Culture and Human Fertility*. UNESCO, Paris
- Malthus T R 1798 *An Essay on the Principle of Population*. J. Johnson, London
- Murdock G P 1975 *Outline of World Cultures*, 5th edn. Human Relations Area Files, New Haven, CT
- Mussallam B F 1983 *Sex and Society in Islam: Birth Control Before the Nineteenth Century*. Cambridge University Press, Cambridge, UK
- Townsend N W 2000 Male fertility as a life time of relationships: Contextualizing men's biological reproduction in Botswana. In: Bledsoe C, Lerner S, Guyer J I (eds.) *Fertility and the Male Life Cycle in the Era of Fertility Decline*. Oxford University Press, Oxford, UK
- Whiting J W M, Kluckhohn R, Anthony A S 1958 The function of male initiation ceremonies at puberty. In: Maccoby E E, Newcomb T, Hartley E (eds.) *Readings in Social Psychology*. Holt, New York, pp. 359–70

C. Bledsoe

Fertility Change: Quantum and Tempo

In the measurement of fertility the term 'quantum' usually refers to the average number of births occurring over the lifetime of a cohort of women (Ryder 1980, Pressat 1985). This number is called the cohort

total fertility rate. More generally, quantum can refer to the average numbers of births that occur in a group of women under certain defined circumstances, such as marriage, or its measurement can be limited to births of a certain order, such as first births. The term 'tempo' usually refers to the average age at which births occur in a cohort of women who have reached the end of the childbearing years (Ryder 1980, Pressat 1985). More generally, tempo can refer to the average time elapsed between a birth and a specific preceding event (e.g., marriage) or its measurement can be limited to births of a certain order.

Estimation of the quantum and tempo of childbearing for a cohort of women is straightforward provided the required data are available and accurate. These variables quantify key dimensions of a cohort's reproductive history but they have the considerable disadvantage of representing past experience: women currently aged 50 did most of their childbearing two to three decades ago when they were in their 20s and early 30s. Trends in cohort measures therefore do not reflect current conditions, and they are not often used to monitor recent developments because fertility can change substantially in a short period of time. Instead, an alternative measure, the period total fertility rate (PTFR), is used far more widely as an indicator of current fertility. It is defined as the average number of births a woman would have if she were to live through her reproductive years (ages 15–49) and bear children at each age at the rates observed in a particular year or period. It is a hypothetical measure because no real group of women has experienced or will necessarily experience these particular rates. The advantage of the PTFR is that it gives up-to-date information on levels and trends in fertility. The PTFR is a quantum indicator of period fertility, and the mean of the age-specific fertility schedule is a tempo measure.

The simplicity and wide availability of the PTFR have contributed to a neglect of some deficiencies in this fertility indicator. The demographic literature on the measurement of fertility includes a number of criticisms of the PTFR (Ní Bhrolcháin 1992). The most important of these is the distortion of the PTFR that occurs when the timing of childbearing changes. During years in which women delay childbearing, period fertility rates are depressed relative to cohort fertility; in years when childbearing is accelerated, period fertility is raised (Hajnal 1947, Ryder 1956, 1964, 1980, 1983). The quantum and tempo dimensions of period fertility interact.

To demonstrate how these tempo distortions operate, it is helpful to examine a set of simple illustrations as presented in Fig. 1. In these hypothetical illustrations, every woman is assumed to have exactly one birth so that the cohort total fertility rate equals one for all cohorts, and these births are assumed to take place at exactly the same age within each cohort. Figure 1(a) depicts the reference situation in which this age is set at exactly 20 years. Since there are no tempo effects,

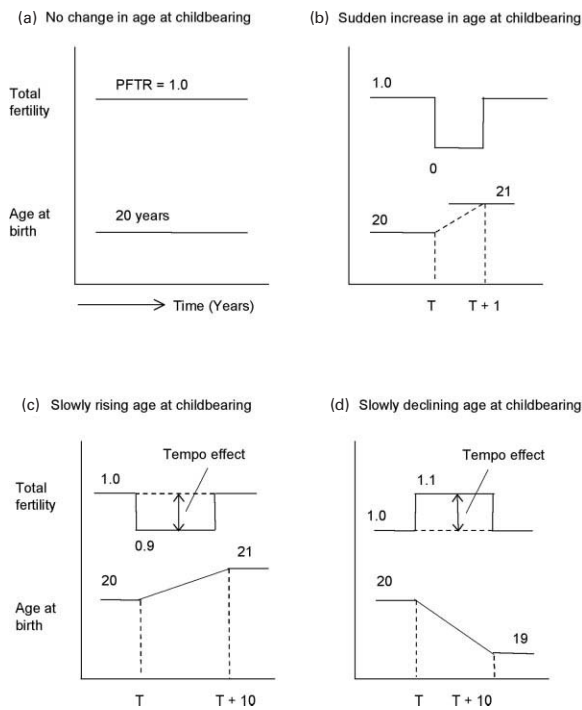


Figure 1

Hypothetical patterns of childbearing, with and without tempo effects

period fertility as measured by the PTFR is also constant and equal to one birth per woman. Assume next that the age at childbirth is raised by one year starting at time T , as shown in Fig. 1(b). In this case, women who would have given birth at age 20 immediately following T have to delay their childbearing for a year until they reach their 21st birthday. As a consequence, no births occur during the year between T and $T+1$ and the PTFR then equals 0. In other words, one year's worth of births is lost because the mean age at childbearing increased by one year. In reality, the mean age at childbearing rarely changes suddenly and rapidly. A more typical situation is summarized in Fig. 1(c), which illustrates the effect of a slow rise in the age at childbearing. Starting at time T , successive cohorts delay childbearing so that the mean age at first birth rises linearly from 20 to 21 years over a 10-year period from T to $T+10$. In this case the PTFR drops from 1 to 0.9 at time T , remains at 0.9 from T to $T+10$, and finally returns to 1 after time T . The PTFR is distorted by 10 percent or 0.1 birth per woman during the interval T to $T+10$. Over the 10-year period, one year's worth of fertility is lost from the PTFR as a result of a one-year increase in the mean age at childbearing. Conversely, as shown in Fig. 1(d), if the mean age at first birth declines linearly from 20 to 19 years over the same period of 10 years, the PTFR

rises from 1 to 1.1 at time T , remains at 1.1 from T to $T+10$, and returns to 1 birth per woman after T . Bongaarts and Feeney (1998) present a fuller discussion of these effects.

The existence of timing distortions is readily documented when the age at childbearing is declining significantly. In that case, implausible results are often obtained for the birth-order components of the PTFR, computed as the sum over age of the period-specific f_{pj} fertility rates of the second kind by age and birth order (see *Demographic Techniques: Rates of the First and Second Kind*). For example, the component for births of order one ($PTFR_1$) equals the average number of first births women would have by age 50 if they were to bear first births at the age-specific rates of the second kind observed in a given year or period. In most years during the late 1940s and early 1950s, the $PTFR_1$ in the US exceeded 1 (Whelpton 1954), a fact which seems to imply that women had more than one first birth on average. This is impossible and these $PTFR_1$ estimates must therefore be reinterpreted. The main explanation of why $PTFR_1$ exceeded 1 during the baby-boom years is that the age at childbearing declined, resulting in a temporary inflation of the PTFR and its additive birth-order components:

$$PTFR = \sum_n PTFR_n \quad (1)$$

An extensive analysis of the effects of variations in the childbearing tempo on period fertility indicators was undertaken by Ryder in a series of influential papers (Ryder 1956, 1964, 1980, 1983). Ryder proposed 'translation' equations to calculate the period fertility measures from corresponding cohort measures when the timing of cohort fertility is changing. These procedures have not found wide acceptance for two main reasons. First, in his work on the translation issue, Ryder assumes that the tempo and quantum of cohort fertility are the determinants of the PTFR and other period fertility measures. However, extensive empirical analysis of this issue has demonstrated that this is not the case (Ní Bhrolcháin 1992). Period effects are clearly the prime source of variation in fertility rates, and the completed fertility of cohorts is largely determined by the period fertility rates during the cohort's reproductive years, rather than the other way around. Second, the two dimensions of aggregate cohort fertility—quantum and tempo—are not independent in practice. When cohorts reduce their fertility, they do this primarily by reducing childbearing at higher birth orders. As a result, the mean age at childbearing for all births to the cohort declines even if the timing of individual births does not change. In other words, a decline in the cohort quantum can lead to changes in the cohort's mean age at childbearing that do not cause tempo distortions. Such distortions only occur if the mean ages at births of specific orders rise or fall. Ignoring this effect—as

Ryder does in much of his writing—therefore gives biased results except when cohort fertility is constant. Fortunately, this second problem can be solved by analyzing fertility trends separately for each birth order rather than for overall fertility.

In a recent study Bongaarts and Feeney (1998) propose a procedure for removing tempo effects from the PTFR. The approach is an outgrowth of Ryder's original translation equation. The above objections to Ryder's approach are addressed by assuming that fertility is only affected by period effects, and by applying adjustments separately to the birth-order components (PTFR_n) of the total fertility rate rather than to the total fertility rate itself. Under these conditions, it is possible to estimate a tempo-free total fertility rate in any given year (or period) from the conventional PTFR_n at each birth order n with the formula

$$PTFR'_n = PTFR_n / (1 - m_n) \quad (2)$$

where PTFR_n is the observed total fertility rate component for birth order n , PTFR'_n is the tempo-free total fertility rate component for birth order n , and m_n is the annual *change* in the mean age of the age-specific fertility schedule for birth order n . By dividing the observed total fertility rate by $(1 - m_n)$ at any given birth order n , we obtain an estimate of the total fertility rate that would have been observed if there had been no change in the timing of childbearing. Adding up results for the various birth orders gives the overall tempo-free total fertility rate

$$PTFR' = \sum_n PTFR'_n \quad (3)$$

The PTFR' equals the distortion-free quantum component of the PTFR, and the difference PTFR'–PTFR equals the tempo distortion.

According to this argument, an annual increase of one-tenth of a year in the mean age at childbearing ($m_n = 0.1$) reduces the PTFR_n by 10 percent below its tempo-free level, because PTFR_n = 0.9PTFR'_n. Similarly, an annual decline in the mean age at a rate of just 0.1 years per year ($m_n = -0.1$) inflates the PTFR_n by 10 percent. Clearly, rather modest changes in the timing of childbearing at any birth order can produce substantial changes in observed fertility.

It should be emphasized that the above simple formula for removing tempo distortions assumes that the shape of the age-specific period fertility rates at each birth order does not change over time. Kohler and Philipov (2001) have suggested a more general equation that incorporates variance effects. Their formula reduces to the simpler version presented here when the shape of the fertility schedule is invariant.

The existence of tempo distortions of period fertility measures has recently become an important issue in contemporary developed countries. In most of these societies, fertility has dropped below the replacement

level. If future fertility remains at these low levels, populations will decline in size and age rapidly. The fact that ages at childbearing have been rising quite rapidly since the 1970s implies that the PTFRs of these countries contain a downward tempo distortion. The size of this distortion can be estimated with the above equation, but the required data—age-specific fertility rates by single age and birth order for successive years—are available only for a restricted number of countries. Bongaarts (1999) estimates the tempo effects for seven post-transitional societies in the late 1980s and finds downward distortions ranging up to 0.4 births per woman. This effect may appear modest, but in some cases it is the primary reason why the PTFR is below the replacement level.

Assessment of fertility levels and trends should include an analysis of the tempo effects because the conventional PTFR gives a misleading estimate of the actual rate of childbearing when the age at childbearing is changing. This is the case in many post-transitional populations. Childbearing is being delayed and women are actually bearing more children than is indicated by the observed PTFRs. This distortion in the PTFR will continue for as long as the mean age at childbearing rises, but once the deferment ends the distortion is removed, and the PTFR will then rise unless there are offsetting declines in the quantum component of period fertility.

See also: Family Size Preferences; Family Systems and the Preferred Sex of Children; Fertility: Institutional and Political Approaches; Fertility: Proximate Determinants; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations; Reproductive Rights in Developing Nations

Bibliography

- Bongaarts J 1999 Fertility decline in the developed world: Where will it end? *American Economic Review* **89**: 256–60
- Bongaarts J, Feeney G 1998 On the tempo and quantum of fertility. *Population and Development Review* **24**: 271–91
- Hajnal J 1947 The analysis of birth statistics in the light of the recent international recovery of the birth-rate. *Population Studies* **1**: 137–64
- Kohler H-P, Philipov D 2001 Variance effects in the Bongaarts-Feeney formula. *Demography* **38**: 1–16
- Ní Bhrolcháin M 1992 Period paramount? A critique of the cohort approach to fertility. *Population and Development Review* **18**: 599–629
- Pressat R 1985 *The Dictionary of Demography*. Basil Blackwell, Oxford, UK
- Ryder NB 1956 Problems of trend determination during a transition in fertility. *Milbank Memorial Fund Quarterly* **34**: 5–21
- Ryder NB 1964 The process of demographic translation. *Demography* **1**: 74–82
- Ryder NB 1980 Components of temporal variations in American fertility. In: Hiorns RW (ed.) *Demographic Patterns in Developed Societies*. Taylor and Francis, London, pp. 15–54

Ryder N B 1983 Cohort and period measures of changing fertility. In: Bulatao R A, Lee R D (eds.) *Determinants of Fertility in Developing Countries*. Academic Press, New York, pp. 737–56

Whelpton P K 1954 *Cohort Fertility: Native White Women in the United States*. Princeton University Press, Princeton, NJ [Reissued 1973 Kennikat Press, Port Washington, NY]

J. Bongaarts

Fertility Control: Eastern Europe

This article reviews the decline of fertility in the countries of Eastern Europe, broadly defined to include all regions of the former Soviet Union. Direct data on contraceptive utilization which have recently become available are examined, together with indirect measures of fertility limitation.

1. Total Fertility

Fertility levels in Eastern Europe and the former Soviet Union span most of the range of variation observed around the world (see Fig. 1). As recently as the mid-1970s, Uzbekistan, Turkmenistan, and Tajiki-

stan registered total fertility rates (TFRs) from six to seven, well within the range of high-fertility Third World countries. At earlier dates (in the 1950s) this was also true of Albania and Azerbaijan. However, the vast majority of the countries in question qualify as low-fertility countries, having experienced TFRs on the order of two to three for most of the period since World War II. In general, all of the countries in question have experienced fertility decline during the postwar period. This includes the period after the dissolution of the Warsaw Pact and the breakup of the former Soviet Union.

It is possible to group the countries under consideration according to their phase of the 'Demographic Transition' from high, natural fertility to low, controlled fertility (Anichkin and Vishnevskiy 1994, Vishnevskiy and Volkov 1983). For the purposes of the present discussion, 'low fertility' is taken to refer to fertility at or below the 'replacement level,' which is the level of fertility that would be consistent with a stationary population in the long run; a TFR of 2.1 is considered conventionally to be equivalent to the replacement level. The vast majority of East European countries, together with the European republics of the former USSR, were in the intermediate stages of the transition prior to World War II, and completed the transition during the postwar period. However, certain of these countries, such as Estonia, had actually

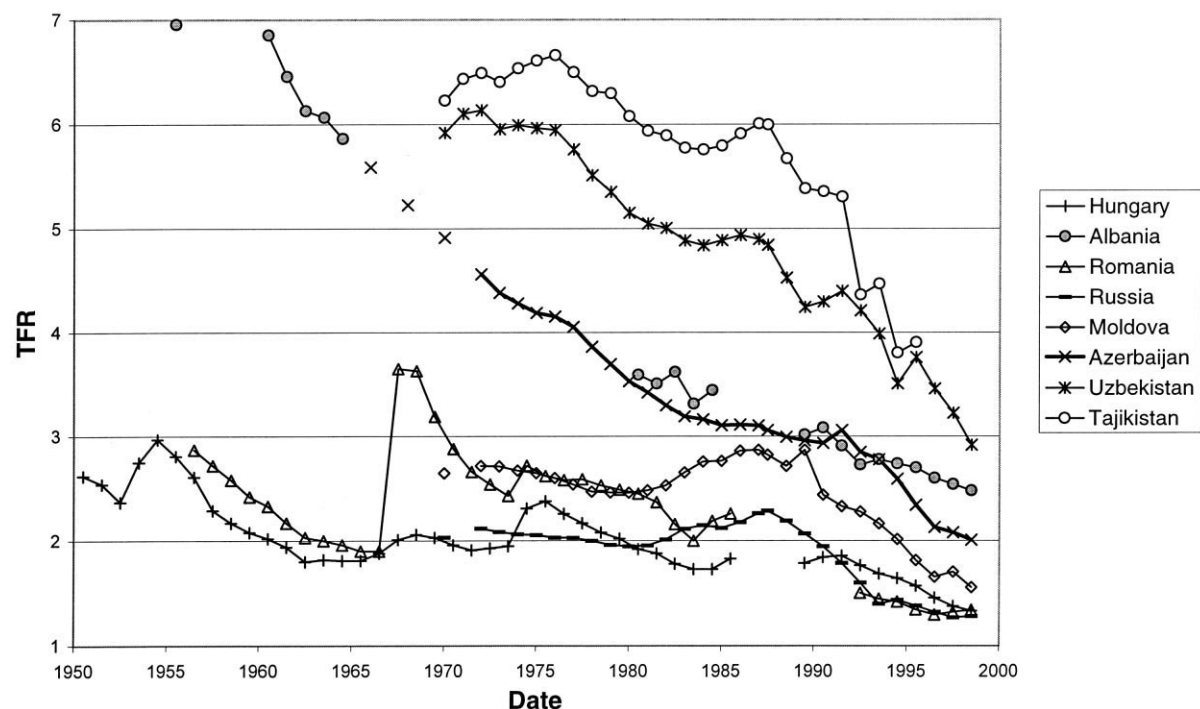


Figure 1

Total fertility rates in selected countries of Eastern Europe and the former Soviet Union

achieved below-replacement fertility prior to World War II. Azerbaijan has completed the transition from high to replacement-level fertility since 1959. The trend in Albania's fertility level resembles closely that of Azerbaijan, although Albania's fertility has not fallen quite to the replacement level. At the opposite extreme, in the indigenous population of former Soviet Central Asia, fertility decline did not take hold until the mid-1970s. These populations remain in the early phases of the demographic transition, exhibiting sustained fertility decline.

There is little evidence of any fundamental difference between the fertility levels and trends of the countries which comprised the former Soviet Union, and other East European countries (see *Fertility Trends in the Formerly Socialist Countries of Europe*). Fertility levels in the European former Soviet republics have fallen below replacement. Russia and Latvia, whose respective TFRs were 1.28 and 1.09 as of 1998, figure among the lowest-fertility countries in Europe. The only major difference between these regions (former USSR and East Europe) at the time of writing is that the former Soviet Central Asian populations, in which fertility decline is a comparatively recent phenomenon, have not yet attained the low fertility levels which typify the European populations. The fertility levels of most of the former Soviet republics discussed here are adjusted for underregistration of births according to the procedures of the International Programs Center of the US Bureau of the Census.

Both the low-fertility formerly Socialist countries of Eastern Europe and the republics of the former USSR, whose governments were pronatalist in orientation, experimented with policy measures to stimulate child-bearing (David and McIntyre 1981, David and Skilogianis 1999). The East European countries were the first to introduce these measures, and did so in the late 1960s and the 1970s. In most cases (e.g., Bulgaria, Hungary, Poland), these measures involved financial incentives such as extended maternity leave, cash awards, and housing subsidies. Typically, the introduction of such measures was followed by a modest rise in fertility of relatively brief duration, and a longer-term slackening or cessation of fertility decline. These pronatalist policies were generally unsuccessful in bringing about increases in fertility to levels substantially above replacement. Only Romania achieved a major rise in TFR, which rose spectacularly from below replacement to over 3.5, then stabilized in the neighborhood of 2.5 for more than ten years. The Romanian achievement owed much to the absence of modern methods of family planning coupled with a draconian ban on abortions, enforced to the extent of subjecting women to regular gynecological examinations to ensure that they were not employing any modern means of fertility limitation. Once these restrictions were lifted, Romania's fertility declined to a level close to that of Russia, one of Europe's lowest.

The Soviet Union introduced a major set of pro-

natalist incentives in 1981. Shortly afterwards, fertility began to rise in the European republics of the USSR, together with Armenia and Georgia. Intriguingly, the fertility levels of these countries converged to approximately the replacement level in the mid-1980s. In Azerbaijan, fertility decline was halted temporarily in the mid-1980s. Concurrently, levels of fertility rose somewhat in Tajikistan, and to a lesser extent in Turkmenistan and Uzbekistan. Although various legislative articles were implemented to enhance the effect of the pronatalist measures throughout the 1980s, they succumbed ultimately to inflation and to the collapse of the Soviet economy. By 1990, the TFRs of the European Soviet republics had returned to approximately their 1980 levels, while fertility decline had resumed in Soviet Central Asia. After the breakup of the USSR, fertility levels continued to decline in all former Soviet republics (Vishnevskiy 1999).

2. Contraceptive Utilization

Direct data on utilization of contraceptive methods in the countries of Eastern Europe and the former Soviet Union is available in many cases from surveys conducted at various points since World War II, especially recently. Table 1 summarizes the data available from the International Demographic Database maintained by the US Bureau of the Census. Two aggregate categories of contraceptive methods are distinguished. The category of 'All' methods of contraception includes traditional methods such as abstinence and withdrawal, whereas 'Modern' methods include the contraceptive pill, IUDs, condoms, sterilization, and a variety of other chemical and mechanical interventions. According to the figures, in all of the East European and former Soviet countries with available data on both forms of contraception, a majority of the population of reproductive-age women were practicing contraception in one form or another at all dates of observation. However, the same cannot be said with regard to modern means of contraception: In most instances, less than half the population were using modern methods.

The data in Table 1 depart from conventional expectations in that the East European countries register levels of modern contraceptive utilization that are usually no greater than those of the former Soviet countries. In several instances, the East European levels are substantially lower than those of former Soviet countries. The levels of modern contraceptive utilization exhibited by Bulgaria and Romania in the 1970s are comparable to that observed in Zaire in 1991. Although this is to some extent a function of time, as access to modern contraceptives improved after the dissolution of the Warsaw Pact and the former USSR, the level of modern contraceptive utilization registered by Romania in 1993 remained well below what is typical for industrialized countries.

Table 1

Percentage of reproductive-age women using contraception

Country	Year(s)	All methods	Modern methods
Bulgaria	1976	76.0	8.0
Czech Republic	1993	68.9	44.8
Hungary	1958	58.0	18.0
Hungary	1974	74.0	44.0
Hungary	1993	84.4	74.6
Poland	1972	60.0	12.0
Poland	1977	75.0	26.0
Romania	1978	58.0	5.0
Romania	1993	57.3	13.9
Slovakia	1991	74.0	42.0
Slovenia	1989	91.6	56.8
Russia	1994	66.8	48.6
Belarus	1995	50.4	42.1
Estonia	1994	70.0	56.0
Latvia	1995	67.8	55.5
Lithuania	1994–5	65.9	22.1
Moldova	1997	73.7	50.1
Kazakhstan	1995	59.1	46.1
Kazakhstan	1999	66.1	54.4
Kyrgyzstan	1997	59.5	48.8
Uzbekistan	1996	55.6	51.3
Congo (Kinshasa)	1991	8.0	2.0
France	1972	64.0	21.0
France	1998	79.5	73.8
Germany	1992	75.0	72.0
United States	1955	58.8	47.8
United States	1988	74.3	69.0

Source: US Bureau of the Census, International Programs Center.

With regard to the former Soviet countries, the data on contraceptive utilization also depart from standard expectations in that the high-fertility Central Asian countries exhibit levels of modern contraceptive utilization as high as, or higher than, a number of European former Soviet republics with much lower fertility levels. Perhaps this reflects the emphasis placed on fertility reduction in these republics in the population policies of the former Soviet government, as well as those of several international organizations that have been active in promoting contraception in Central Asia after the breakup of the USSR.

The above discussion raises the question of the extent to which contraceptive utilization as measured in available surveys relates to realized fertility levels. The answer to this question is positive with respect to modern means of contraception: use of modern contraception has a decidedly negative and statistically significant correlation ($r = -0.534$) with the total fertility rates registered by the countries for which data are available at the appropriate dates. However, when traditional means of contraception are included, the

relationship deteriorates into a minor positive correlation ($r = 0.153$) which is not statistically significant.

3. Overall Fertility Limitation

The limited coverage of, and unexpected results obtained with, the available direct data on contraceptive utilization in the East European and former Soviet countries justifies the use of indirect estimation methods to assess the extent of fertility reduction in a fuller range of countries and dates. Table 2 contains estimates of Coale's m (Coale and Trussell 1974) as of 1995 for those countries for which there were sufficient data available to permit calculation of this measure. The results leave an overall impression of greater conformity with general expectations than obtains for direct measures of contraception. In particular, the Czech Republic exhibits a higher level of fertility limitation, as reflected by a higher value of m than any of the former Soviet countries, while Hungary and Slovakia register levels comparable to the Slavic former Soviet countries (Russia, Ukraine, Belarus), along with Moldova. These former Soviet republics, in turn, register higher levels of fertility limitation than any of the Transcaucasian or Central Asian former Soviet republics. The principal departure from standard expectations is that the Baltic countries of the former USSR (Estonia, Latvia, Lithuania) register lower values of m than one would anticipate. This may partly reflect the fact that these countries had achieved

Table 2

Index of fertility limitation (m) for former Soviet and certain East European countries, 1995

Country	Index
Hungary	2.294
Czech Republic	2.575
Slovak Republic	2.252
Russia	2.365
Ukraine	2.379
Belarus	2.402
Moldova	2.175
Estonia	1.625
Latvia	1.918
Lithuania	2.026
Georgia	1.872
Armenia	2.026
Azerbaijan	1.921
Kyrgyzstan	1.300
Kazakhstan	1.772
Uzbekistan	1.977
Turkmenistan	1.384
Tajikistan	0.769

Note: Figures calculated from data maintained by the US Census Bureau, International Programs Center

high levels of fertility control earlier during the former Soviet period, and that the recent further reductions experienced in these countries have caused the resultant marital fertility schedules to deviate from the model age pattern of fertility regulation.

See also: Eastern European Studies: History; Eastern European Studies: Society; Fertility Control: Overview; Fertility: Proximate Determinants; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations; Reproductive Rights in Affluent Nations; Revolutions of 1989–90 in Eastern Central Europe; Soviet Studies: Society

Bibliography

- Anichkin A A, Vishnevskiy A G 1994 Three types of fertility behavior in the USSR. In: Lutz W, Sherbov S, Volkov A (eds.) *Demographic Trends and Patterns in the Soviet Union before 1991*. Routledge, London
- Coale A J, Trussell J T 1974 Model fertility schedules: Variations in the age structure of childbearing in human populations. *Population Index* 40(2): 185–258
- David H P, McIntyre R J 1981 *Reproductive Behavior: Central and East European Experience*. Springer, New York
- David H P, Skilogianis J 1999 *From Abortion To Contraception: A Resource to Public Policies and Reproductive Behavior in Central and Eastern Europe from 1917 to the Present*. Greenwood Press, Westport, CT
- Vishnevskiy A G 1999 Demographic processes in the post-Soviet states *Revue Baltique*, No. 13: 23–37
- Vishnevskiy A G, Volkov A G 1983 *Воспроизводство населения СССР* (Reproduction of the Population of the USSR). Nauka, Moscow

W. W. Kingkade

Fertility Control: Overview

The primary immediate, or proximate, biological and behavioral determinants of the levels and patterns of human fertility are (a) the extent and effectiveness of contraceptive use; (b) the degree to which abortion is practiced; (c) the intensity and duration of breastfeeding behavior; and (d) the degree of exposure to sexual intercourse, often proxied by various measures of the extent of marriage. The levels of the various proximate determinants are, in turn, functions of social structure, which differs across populations and over time (Davis and Blake 1956).

The term ‘fertility control’ can be taken to refer to any or all of the proximate determinants. However, while marriage patterns clearly affect fertility, it is often thought that marriage behavior is motivated and explained by factors that are not related directly to a desire to control fertility. The same could be con-

sidered true regarding breastfeeding behavior. Thus, some demographers would define ‘fertility control’ only in terms of those behaviors known to be motivated by a desire to control fertility: contraception and abortion. The primary focus of this article is to provide an overview of these two determinants of fertility.

1. A Variety of Reasons Humans Have Desired to Control Fertility

Judging from what is known about nineteenth- and twentieth-century preliterate populations, as well as from a variety of written evidence left behind by ancient through early modern societies, it appears that men and women have been motivated to control fertility—so that there would not be too few, nor too many children—for at least the last several thousand years (Himes 1970). Additional evidence of the (post-natal) desire to control fertility is seen in the resort to infanticide, child negligence, and child abandonment in the ancient world and throughout much of the Middle Ages and beyond (see *Infanticide and Induced Abortion*). The focus here is on the various reasons humans have desired to limit fertility.

Health concerns have always been an issue; pregnancy and childbearing still present serious medical risks to women in much of the developing world, as they have until fairly recently in the currently industrialized world. Even if continuous pregnancy and childbirth did not prove fatal, the sickness, pain, and suffering involved were motivation enough for wanting to postpone, avoid, or terminate the next pregnancy. A rapid pace of childbearing presented risks to surviving children too, particularly the youngest, since it taxes parents’ ability to care for their family. For example, birth spacing was seen as desirable in order to prolong breastfeeding, often considered important for a child’s health and development.

In addition, there has often been the motivation to prevent or terminate pregnancy that resulted from extramarital or premarital sexual relations, either because such relations were socially sanctioned or because the consequent childbearing was deemed undesirable. Offspring born as a consequence of such sexual relations could prove extremely inconvenient to the fathers (and their legitimate children) and devastating to the mothers.

Within marriage, there has also been the desire to space and limit the number of births, for various social and economic reasons (see e.g., McLaren 1990). In terms of the timing of fertility, there was a desire to avoid childbearing during the busy agricultural seasons, since women would be working hard in the fields. In other contexts, parents who wished to avoid dividing their estates into too many parts may have wanted to limit the number of their descendants. Later on, upwardly mobile middle-class and working-class families may have chosen to have smaller families

because of increasing costs of raising children or in order to invest more in the education and health care of each child. While there have been many reasons for men and women to want to control their fertility in different periods and in different cultures, that desire is found in all societies and is very old indeed.

2. A Variety of Fertility Control Methods

Over the thousands of years that humans have desired to exercise control over their fertility, various methods have been developed, many through trial and error, some partially effectively, some highly effective, and some not effective at all. Some methods have been medically harmful to the women who used them, while others harmless. In this section, descriptions of the types of methods used, their historical significance, mechanisms of action, degrees of efficacy, and where possible, extents of use in the past, are presented.

Regarding efficacy, it is important to keep in mind that a contraceptive method which does not guarantee that any one woman will not conceive while the method is applied, can nonetheless have a profound effect on the overall probability of conceiving in a population. To see this, consider the following: It is estimated that if 100 couples exposed to vaginal intercourse do not use any form of contraception, approximately 85 of the women will become pregnant within one year. A contraceptive method such as *coitus interruptus*—one which is not considered by modern standards to be among the most effective—dramatically reduces the annual risk of pregnancy from 85 percent to approximately 19 percent. In this article, references will be made, when possible, to contraceptive failure rates which indicate the proportion of women who would become pregnant during the first year of typical use among couples in the United States. Typical use refers to method use which may be imperfect or inconsistent. The failure rates cited should always be compared with the probability of pregnancy when no method is used—that is, approximately 85 percent. The reader should also keep in mind that ‘perfect-use’ failure rates—the proportion of women who would become pregnant within one year of correct and consistent use of a specific method of contraception—are often substantially lower than typical-use failure rates (Hatcher et al. 1998).

2.1 Methods Which are Completely Ineffective

The application of ineffective methods provide evidence for the widespread desire and hope for fertility control, rather than for human ability to translate those hopes into effective action. A long list of amulets, symbolic magic, and rituals that have no contraceptive

or abortive effect have nonetheless been known and practiced in ancient and contemporary populations. For example, written evidence from the Middle Ages in Europe describe a belief that lying or sitting down upon fingers will prevent conception for a number of years that is equal to the number of fingers. In addition, it has long been believed that the passivity of women during sexual intercourse, or their ‘holding back’ an orgasm will prevent conception. A wide variety of herbal preparations have also been used with the intent of fertility control; doubt remains as to whether any of these potions do indeed have contraceptive or abortive effects in humans (Riddle 1997, Santow 1998).

2.2 Behavioral Methods of Contraception

This section reviews methods of preventing contraception which do not require the use of any type of physical or chemical substance, but which rather rely on the modification of sexual and reproductive behavior. In this category fall some of the oldest effective methods known. Reduced coital frequency, even if practiced without regard to the timing of sexual intercourse during the menstrual cycle, will affect the monthly probability of conception to an extent which is roughly proportional to the percentage decline in coital frequency (Trussell 1979). Clearly, sexual abstinence, or continence—that is, refraining from vaginal intercourse—reduces fertility. David and Sanderson (1986) argue that sexual restraint of these kinds were important factors in bringing about the nineteenth-century decline in fertility in the United States.

Fertility awareness methods of periodic abstinence, by which the fertile days during the cycle are correctly identified and avoided, will be more effective than a simple overall reduction in coital frequency. Various types of methods are in use today (e.g., Calendar Rhythm Method, Basal Body Temperature method, Cervical Mucus charting), but historically, these are fairly new developments. It was not until the early 1930s that the timing of ovulation within the menstrual cycle was correctly identified and a chart was worked out for use by women desiring to identify the days it was ‘safe’ to have sexual intercourse. Typical-use failure rates of periodic abstinence hover around 25 percent. Prior to the second quarter of the twentieth century, attempts at periodic abstinence, based on incorrect beliefs about the timing of the fertile period, while often discussed, were generally ineffective.

Prolonged and intensive lactation have the effect of delaying the return of ovulation after a live birth. There is evidence that certain preliterate societies, as well as the ancient Egyptians, were aware of the fertility-inhibiting function of breastfeeding, and practiced prolonged lactation for the purpose of increasing the spacing between births. Such practices were com-

mon in the currently developed nations in the 1900s, and they are still in use there among various religious groups. In addition, throughout history, it has often been believed that semen would spoil the milk of lactating women. Customs of post-partum abstinence during lactation, more common in polygynous societies, enhance the fertility-inhibiting effect of breastfeeding, whether or not they are practiced for the purpose of fertility control (see *Fertility Control: Prevalence and Consequences of Breastfeeding*).

Modifications of sexual practices, such as oral sex, anal sex and *coitus reservatus*—vaginal penetration without ejaculation—may have been used for the purposes of preventing conception. Probably the ancient Peruvians, Greeks, and Hebrews, among others, performed such practices. However, the behavioral modification invoked most universally, both in terms of time and space, and most certainly for the purposes of fertility control, is *coitus interruptus*, or withdrawal, whereby male ejaculation occurs outside the vaginal cavity, following vaginal penetration.

Evidence indicates that preliterate populations, the ancient Hebrews, ancient Islamic societies, and others used withdrawal. The Bible mentions *coitus interruptus* in *Genesis* 38:7–10 in the story of Er and Onan. Numerous references to the practice can be found afterwards, some in favor, many denouncing it. Santow (1995) provides a compelling argument for the demographic importance of *coitus interruptus*, even centuries before the modern transition to low fertility. There is no doubt that its use was very important in bringing about the decline in fertility in many parts of Europe during the last part of the nineteenth century and early part of the twentieth century. Even in countries like the USA, where its prevalence declined relatively early, *coitus interruptus* was among the most frequently cited, if not the most cited, method of birth control reported in studies undertaken in the 1930s. In France and Belgium, significant reliance on withdrawal continued into the 1970s. Today, *coitus interruptus* remains a frequently-used method of birth control only in some parts of Southern and Eastern Europe, as well as in parts of the developing world, such as West Asia (Santow 1993).

2.3 Physical, Mechanical, and Chemical Means of Contraception

Various techniques have been developed over time to remove semen from the vagina after ejaculation. Physical methods of sperm removal cited in early literature include such means as the woman squatting and straining, or jumping following intercourse (Himes 1970). Probably the most widely known and used technique for sperm removal in past times was douching, referred to, for example, in Hindu writing of around 1,500 BC. Some primitive douches have an additional spermicidal effect, such as those which use

solutions of vinegar, salt, lemon juice, or tannic acid. Charles Knowlton, an early American birth control advocate, recommended postcoital douching with a solution of zinc and alum in his 1832 book *The Fruits of Philosophy*. David and Sanderson (1986) suggest that douching was a method frequently used by urban American birth controllers in the early nineteenth century, perhaps in conjunction with reduced coital frequency and abstinence. Woycke (1988) also writes that douching was popular in nineteenth-century Germany, partly because it could be seen as a having a legitimate hygienic function, even by those ambivalent about the moral rectitude of pregnancy prevention. Typical-use failure rates of douching are probably close to 40 percent (Kestelman and Trussell 1991).

Other physical and mechanical means of contraception include those that attempt to prevent entry of sperm into the cervix. Barrier methods described by ancient Greek writers include smearing of oils on the cervix. Similar strategies are discussed in ancient Indian writings. A description of the use of crocodile dung as a pessary appears as early as 1850 BC in the ancient Egyptian Petri papyrus. In other references, alternative sticky substances, such as honey or melted beeswax were suggested for use as pessaries (Himes 1970).

A tuft of wool or cotton, frequently referred to in some Jewish theological writings (Talmud) as a *mokh*, was probably placed in the vagina before coitus or after coitus, in order to absorb the semen and prevent it from entering the cervix (Feldman 1995). In certain contexts, the wool or cotton was used as a vehicle for some type of sticky substance or primitive spermicide. Modern-day sponges are small and pillow-shaped, made of polyurethane, imbued with spermicide, and have typical-use failure rates of approximately 20 percent among nulliparous women, 40 percent among parous women.

Another important barrier method of birth control, the male condom, also has a long history. In use in various forms by preliterate peoples as well as early Egyptians for noncontraceptive purposes (e.g., for decoration), condoms became associated with prostitution during the sixteenth century, when they were used primarily for protection against sexually transmitted diseases. Even today in some contexts, condoms carry a negative stigma associated with promiscuity or prostitution. At least from the second century AD, condoms were often made from animal membranes. It was not until the middle of the nineteenth century that the development of the vulcanization of rubber allowed for production of less expensive condoms. However, even rubber condoms were probably too expensive in the mid-nineteenth century for widespread use by the lower classes. In the 1930s, the improved technology that allowed for latex condoms made condoms even cheaper and more appealing for use. Their popularity spread, so that

surveys of the 1930s show condoms competing with withdrawal for popularity in the USA. At the beginning of the twenty-first century, unlike in the past, condoms are high quality and reliable, so that breakage is rare. However, typical-use failure rates approximate 14 percent, largely because of inconsistent use.

Various other barrier methods, forerunners of the modern-day cervical caps and diaphragms, included the hollowed-out rinds of lemons, discussed, for example, by Casanova in the eighteenth century. The vaginal barriers cover the cervix in order to prevent the entrance of sperm. The rubber cervical cap and diaphragms were developed and perfected by German physicians and scientists in the nineteenth century—the cap by F. A. Wilde in 1838 and the diaphragm by W. Mensinga in 1880. The use of the latter method spread from Germany to Holland, and then to England, where it became known as the ‘Dutch Cap.’ Use of the diaphragm and cap have been supplemented with various spermicidal elements, such as quinine compounds, and later, modern spermicidal creams, introduced into the interior cavity of the objects before being inserted into the vagina or placed on the cervix. Promoted by birth control propagandists both in England and in the USA in the 1920s and 1930s—Dr. Marie Stopes in England advocated use of the cervical cap, Margaret Sanger the diaphragm in the USA—use of these vaginal barrier methods was limited largely to the middle classes, but apparently well accepted in this sector of the population. Modern-day diaphragms, used in conjunction with spermicidal cream or jelly, have typical-use failure rates in the range of 20 percent. Failure rates for the cervical cap are generally higher, especially among parous women.

2.4 Hormonal Methods

The search for an oral contraceptive has a long history in herbal formulations. However, only in the mid-1920s did scientists discover the antifertility effects of estrogenic compounds, and it was not until 1960 that the United States Food and Drug Administration approved the first hormonal oral contraceptive. Modern-day combined oral contraceptives, among the most extensively studied medications ever prescribed, are safe; in many countries, but not in the USA, oral contraceptives are available without a doctor's prescription. The pills are highly effective and prevent pregnancy primarily by suppressing ovulation through the combined actions of estrogen and progestin. Typical-use failure rates are approximately 5 percent.

Following its introduction in 1960s, there was rapid, widespread acceptance of the pill in the US, especially among young women. By 1965, 24 percent of all married contraceptors were relying on the pill, and by 1970, the pill was the single most popular method used

by married American women. Increasing reliance on the pill was accompanied by a reduction in use of barrier methods, rhythm, and withdrawal (Westoff and Ryder 1977). In many other currently industrialized nations as well, pill use has eventually overtaken older methods of contraception at various times, depending on when it was introduced and made easily available. Today, the pill is the most widely-used reversible method of contraception in the Western developed countries. It is important to recall, however, that the major declines in fertility in the currently industrialized nations began in the late nineteenth and early twentieth centuries, well before the development of the oral contraceptive, and even before the mass production and distribution of barrier methods such as condoms and diaphragms.

For a variety of reasons, some women want to avoid hormonal contraceptives which contain estrogen. In these cases, progestin-only methods of contraception, such as injectibles, implants, and pills can be safe, highly effective alternatives. In the USA, Depo-Provera injections and Norplant implants have typical-use failure rates of less than 1 percent. The progestin-only pill, or minipill, is probably only slightly less effective than combined oral contraceptives (Hatcher et al. 1998).

Emergency contraceptive pills (ECPs)—consisting of progestin only or a combination of estrogen and progestin—can be used within 72 hours following unprotected intercourse. Emergency contraception acts to prevent pregnancy, and is not an abortifacient. If 100 women had unprotected intercourse once during the second or third week of their menstrual cycle, about eight would become pregnant. Following treatment with ECPs, only one or two would become pregnant. Thus, there is a reduction of about 75 percent–88 percent in the probability of pregnancy resulting from use of ECPs (<http://ec.princeton.edu/>).

2.5 Intrauterine Devices (IUDs)

Physicians began experimenting with intrauterine pessaries for women in the middle of the nineteenth century. However, use of early IUDs, made of silk and silver wire, often led to serious pelvic infections, and their use was thus considered by most physicians to be unwise up through the middle of the twentieth century. The use of plastic instead of wire, begun in the late 1950s, improved the safety of IUD use.

The modern-day IUD is inserted by a physician into the uterus, with a string that should protrude from the uterus to aid in removal. The device can remain in the uterus and provides contraceptive protection for a number of years. The IUD is thought to work primarily by preventing fertilization; it is not currently thought to be an abortifacient. Highly effective, the method has typical-use failure rates that generally lie below the 2 percent level. Some types of IUDs can also

be used as emergency contraception, whereby they are inserted into the uterus within a few days of unprotected intercourse.

2.6 Contraceptive Sterilization

Primitive attempts at producing sterility in men have taken the form of drastic actions such as the castration of male servants chosen to safeguard female residents of imperial palaces in ancient China circa 1,100 BC. In an attempt to make women infertile, surgical removal of the ovaries was used by the ancient Egyptians, and retroflexion of the uterus to block the cervical canal was used by certain tribes of Java, as reported in the late nineteenth century (Finch and Green 1963).

During the nineteenth and twentieth centuries, more effective and safer methods of sterilization were developed with improvements in asepsis and anesthesia. Today, female sterilization is accomplished through various methods of mechanically blocking the fallopian tubes to prevent the sperm and egg from uniting. Male sterilization (vasectomy), medically even simpler, safer, and less expensive than female sterilization, involves blocking the *vas deferens* in order to prevent passage of sperm into the ejaculated seminal fluid. Although it is sometimes possible to reverse the procedures, neither male nor female sterilization is considered reversible. Highly effective, both methods have typical-use failure rates of less than 1 percent.

In the first half of the twentieth century in the US and in Europe, sterilization was thought to be appropriate primarily in the eugenic context of those thought mentally unfit to be parents. However, since the 1960s, there has been widespread acceptance of voluntary sterilization, so that now roughly 50 percent of married contraceptors of reproductive age in the USA rely on either female or male sterilization. Worldwide, female sterilization is the dominant form of contraception used today.

2.7 Induced Abortion

A popular notion of the difference between contraception—the prevention of pregnancy—and abortion—the termination of pregnancy—is a modern-day distinction which did not usually exist in the past. At least during the early modern period, if not before, and to the end of the nineteenth century, abortion performed prior to quickening was viewed traditionally by European women as just another point on the continuum of fertility-control practices, and not generally less legitimate or acceptable than the rather indelicate means of contraception available at the time (McLaren 1990).

Abortion was probably more commonly used than was contraception before modern times (Himes 1970). Various traditional and folk methods were employed

throughout the ages, and some are still in use today, where legal or safe abortion is not widely available. Examples include jumping off tables or down stairs, engaging in vigorous exercise, and performing abdominal massages. Purportedly emmenagogic or abortifacient potions and pills made of lead, arsenic, phosphorus, gin, quinine, or different herbal extracts, to name a few, were often either poisonous and potentially lethal to the pregnant woman or, at most partially effective (Santow 1998, Woycke 1988). Insertion into the uterus of objects such as hat pins, knitting needles, and infamously, coat hangers, is dangerous, primarily because of risk of infection and hemorrhage. Sometimes these mechanical abortions are self-induced, sometimes they are performed by abortionists of varying reputation.

Over time, improvements in techniques and asepsis allowed for the development of safe and effective surgical abortion. By the 1930s, the dilation and curettage method was recognized as being as safe as full-term delivery. Today, legally induced abortion is significantly safer medically than continuing a pregnancy to term. Introduced to the USA in 1967, vacuum aspiration is currently the most commonly used method of abortion there. Recent developments in medical abortion have the potential of reducing the need for these types of surgical abortion. For example, mifepristone (formerly known as RU486), in combination with prostoglandins, is used widely to abort early pregnancies in several European countries, and has been found to be safe and effective.

The contribution of abortion to the decline in fertility in currently industrialized countries cannot be quantified, but there is strong reason to believe that it is significant. Probably no country has been able to attain replacement-level fertility without the recourse to induced abortion. Generally, use of abortion increases during the early stages of fertility decline, and then decreases as more effective contraception is adopted by a larger proportion of the population (Ross and Frankenberg 1993; see also *Infanticide and Induced Abortion*).

3. Recent Trends in Contraceptive Use

Accurate information on the use of abortion is not available in many countries, even when abortion is legal, but especially when it is not (Henshaw 1990). Tracking trends in this sort of data is even more difficult. Therefore, this section focuses only on levels and trends in contraceptive use. One important measure of contraceptive use is contraceptive prevalence, defined as the percentage of couples (with the woman of reproductive age) who are currently using contraception.

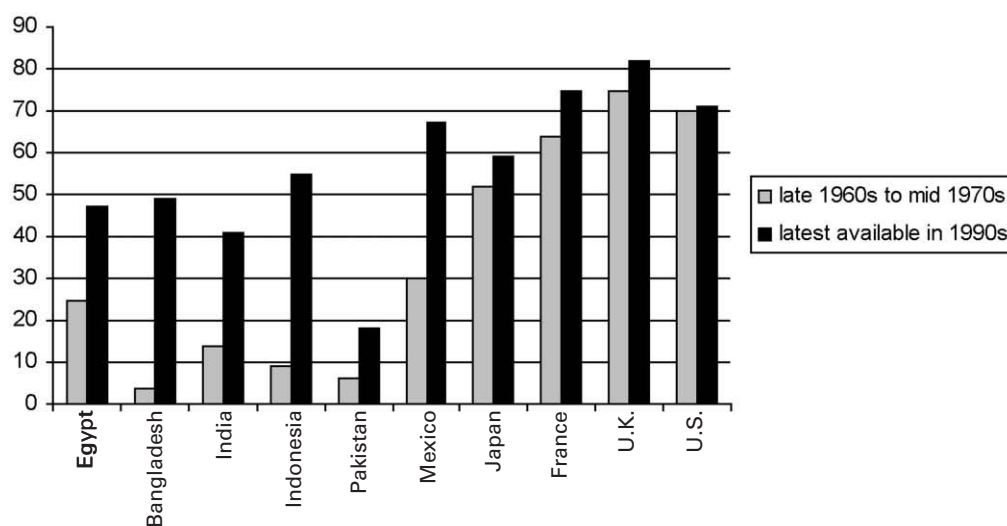
Systematic surveys of contraceptive prevalence and method use began to be conducted only in the 1960s in developing countries, as well as in most developed

Table 1Contraceptive use^a by method type and by region of the world, circa 1993

World and regional averages	Contraceptive prevalence ^b	Percentage distribution of couples, by type of method used						
		Sterilization		Pill	IUD	Condom	Other supply methods ^c	Nonsupply methods ^d
		Female	Male					
World	58	19	4	8	13	4	3	8
Less developed regions	55	21	4	6	14	2	2	5
Africa	20	2	< 1	7	4	1	2	4
Asia ^e	60	24	5	5	17	3	2	4
Latin America and the Caribbean	66	28	1	14	7	4	2	9
Oceania	29	9	< 1	5	1	1	6	7
More developed regions	70	9	5	17	6	14	2	19
Japan	59	3	1	< 1	2	46	1	6
Europe	72	3	2	20	8	10	2	26
North America	71	24	14	15	1	10	4	3
Australia/New Zealand	76	26	12	23	5	6	1	4

Source: United Nations 1999.

^a Contraceptive use among currently married women of reproductive age, including, where possible, those in consensual unions. ^b Contraceptive prevalence is the percentage currently using contraception among couples with the woman of reproductive age. ^c Other methods requiring supplies or medical services, including injectibles, diaphragms, cervical caps, and spermicides. ^d Including periodic abstinence or rhythm, withdrawal, douche, total sexual abstinence if practiced for contraceptive reasons, folk methods and other methods not separately reported. ^e Excluding Japan.

**Figure 1**

Trends in contraceptive prevalence, selected developing and developed countries (Sources: United Nations (1996, 1999))

countries. Increased concern over rapid population growth and high fertility levels in post-World War II developing countries led to the employment of large-scale surveys of contraceptive use that allow for documentation of prevalence and trends in those regions that is more comprehensive and up-to-date than in currently developed regions. Table 1 presents contraceptive prevalence rates in different regions of

the world, based on the most recent data that is currently available. Data in the table refer roughly to the period around 1993.

Contraceptive prevalence was approximately 58 percent for the world as a whole in the years around 1993, but there are important differences between regions. Overall, less developed regions have somewhat lower proportions who use some method of

contraception (55 percent) than do more developed regions (70 percent). The very high level of prevalence in China, currently estimated at 83 percent, raises substantially the average for the less developed regions because of China's large population size. It is estimated that in less developed regions, excluding China, contraceptive prevalence is only 42 percent. In Africa, particularly Western Africa, the prevalence level is lowest in the world (8 percent). Most countries in Latin America and the Caribbean have moderate to high levels of contraceptive use, with nearly all countries reporting prevalence rates of at least 40 percent. Among the more developed regions, the range in prevalence levels is much smaller than among the less developed regions.

Eighty-seven percent of couples who currently use contraception worldwide rely on supply methods—that is, methods which require supplies or medical services. The major supply methods used by couples today are female and male sterilization, the pill, the IUD, and condoms. Supply methods actually comprise a larger percentage of total contraceptive use in developing regions (91 percent) than in developed regions (74 percent). This is probably a function of the role in developing countries of family planning programs, which began massive supply work just when modern methods became available in the 1960s (see *Family Planning Programs: Development and Outcomes*). Only in certain areas such as Southern and Eastern Europe, and parts of West Asia, are nonsupply methods, primarily *coitus interruptus* and rhythm, still widely used among married couples.

Worldwide, and in less developed regions overall, female sterilization is the method most often used. In China and in India, between a quarter and a third of all married women of reproductive age are sterilized. In developed regions overall, the pill is the most commonly used supply method among married couples, but sterilization is the most dominant method in North America and Australia/New Zealand. Condom use is more common in the more developed regions, particularly in Japan, than in the less developed regions, while IUD use is more common in the less developed regions, especially in China. In the more developed regions, IUD use is relatively common in Northern and Eastern Europe. Use of vaginal barrier methods is negligible everywhere.

Nearly all developing countries for which trend data is available have experienced increases in contraceptive prevalence between the years 1970–2000. It has been estimated that contraceptive prevalence for less developed countries has risen by approximately 41 percentage points since the period 1960–5, when it stood at only 14 percent. Increases in prevalence rates among developed countries have been much more modest than among developing countries, in part because the latter started from lower levels (see *Fertility Control: Eastern Europe*). Figure 1 provides a few examples of changes in prevalence rates for a

selection of less and more developed countries. In terms of method mix, rising use of supply methods, particularly female sterilization, is the dominant trend in more developed as well as in less developed regions, but the experiences of individual countries is varied. In many more developed countries, increasing use of supply methods has been accompanied by decreasing use of non-supply methods.

See also: Fertility: Proximate Determinants; Reproductive Rights in Developing Nations

Bibliography

- Bongaarts J, Potter R G 1983 *Fertility, Biology, and Behavior: An Analysis of Proximate Determinants*. Academic Press, New York
- David P A, Sanderson W C 1986 Rudimentary contraceptive methods and the American transition to marital fertility control, 1855–1915. In: Engerman S L, Gallman R E (eds.) *Long-Term Factors in American Economic Growth*. University of Chicago Press, Chicago, pp. 307–90
- Davis K, Blake J 1956 Social structure and fertility: An analytic framework. *Economic Development and Cultural Change* 4: 221–35
- Draper E 1972 *Birth Control in the Modern World: The Role of the Individual in Population Control*, 2nd edn. Penguin Books, Harmondsworth, Middlesex, UK
- Feldman D M 1995 *Birth Control in Jewish Law*. New York University Press, New York
- Finch B E, Green H 1963 *Contraception Through the Ages*. Peter Owen, London
- Hatcher R A, Trussell J, Stewart F, Cates W Jr, Stewart G K, Guest F, Kowal D 1998 *Contraceptive Technology*, 17th edn. Ardent Media, New York
- Henshaw S K 1990 Induced abortion: A world review, 1990. *Family Planning Perspectives* 22: 76–89
- Himes N E 1970 *Medical History of Contraception*. Schocken Books, New York
- Kestelman P, Trussell J 1991 Efficacy of the simultaneous use of condoms and spermicide. *Family Planning Perspectives* 23: 226–32
- McLaren A 1990 *A History of Contraception from Antiquity to the Present Day*. Blackwell, Oxford, UK
- Ranum O, Ranum P 1972 *Popular Attitudes toward Birth Control in Pre-Industrial France and England*. Harper & Row, New York
- Reed J 1983 *The Birth Control Movement and American Society, From Private Vice to Public Virtue*. Princeton University Press, Princeton, NJ
- Riddle J M 1997 *Eve's Herbs: A History of Contraception and Abortion in the West*. Harvard University Press, Cambridge, MA
- Ross J A, Frankenberg E 1993 *Findings from Two Decades of Family Planning Research*. The Population Council, New York
- Ross J, Stover J, Willard A 1999 *Profiles for Family Planning and Reproductive Health Programs 116 Countries*. The Futures Group International, Glastonbury, CT
- Santow G 1993 Coitus interruptus in the twentieth century. *Population and Development Review* 19: 767–92
- Santow G 1995 Coitus Interruptus and the control of natural fertility. *Population Studies* 49: 19–43

- Santow G 1998 Eve's Herbs: A history of contraception and abortion in the West. *Population and Development Review* **24**: 869–75
- Trussell J 1979 Natural fertility: Measurement and use in fertility models. In: Leridon H, Menken J (eds.) *Natural Fertility*. Ordina Editions, Liege, Belgium
- Trussell J, Kost K 1987 Contraceptive failure in the United States: A critical review of the literature. *Studies in Family Planning* **18**: 237–83
- United Nations 1996 *Levels and Trends Contraceptive Use as Assessed in 1994*. United Nations, New York
- United Nations 1999 *World Contraceptive Use 1998 Wall Chart*. United Nations, New York
- Westoff C F, Ryder N B 1977 *The Contraceptive Revolution*. Princeton University Press, Princeton, NJ
- Woycke J 1988 *Birth Control in Germany 1871–1933*. Routledge, London

B. S. Okun

Fertility Control: Prevalence and Consequences of Breastfeeding

The fertility inhibiting consequences of breastfeeding have been known for centuries. Among the ancient Greeks, Aristotle noted that women who nursed their infants for extended periods of time were able to delay the onset of postpartum menstruation and conception. Yet only recently have demographers begun fully to appreciate breastfeeding's impact on global fertility levels. Since the 1960s a large number of scientific studies from many parts of the world have consistently demonstrated that breastfeeding is closely associated with prolonged periods of postpartum amenorrhea, long birth intervals, low completed fertility, and optimal health and wellbeing for the mother and infant. Breastfeeding is now recognized as an important determinant of fertility, especially in the developing world, where modern contraceptive usage is low and breastfeeding rates are high. In this article we review the recent transition in thinking regarding the contraceptive role of breastfeeding.

1. Early Demographic Research

Beginning in the 1950s researchers hypothesized that variation in breastfeeding behavior contributed to fertility differences among populations. Studies showed that the survival of one child is positively associated with the mother's succeeding birth or pregnancy interval. In Senegal, researchers noted that the interval between one birth and the next increased by at least nine months when the age at weaning of the first of those children increased by one year. Also, average birth intervals were 25 months when Senegalese mothers breastfed their children for 12 months

or less, but were as long as 41 months when they breastfed for up to 35 months. Still other studies demonstrated that in the absence of any breastfeeding the average duration of postpartum amenorrhea is one or two months with a corresponding birth interval of just over one year. By the 1970s the weight of the accumulated demographic evidence led one researcher to conclude that, in developing countries, breastfeeding prevents more pregnancies than all modern forms of contraception combined (for details see Rosa 1975).

By the mid-1970s other researchers using linear regression techniques found that when regressing the average duration of postpartum amenorrhea on the average duration of breastfeeding, each additional month of breastfeeding increases amenorrhea by somewhat less than one month. These types of empirical findings helped form the basis for what has become known in demographic circles as the 'proximate determinants of fertility' framework. This modeling approach tries to account for variation in fertility between populations by quantifying a limited and exhaustive list of variables having a direct bearing on fertility. The period of time in which nursing mothers are not menstruating or ovulating and are unable to conceive following a pregnancy is referred to as lactational infecundability. As the name implies, breastfeeding is assumed to be the leading or sole cause of the delayed return to the fecundable state in nursing mothers. Used widely as an analytical tool, the proximate determinants of fertility framework has demonstrated, worldwide, that lactational infertility is an important determinant of the total fertility rate between populations. For example, in an analysis of World Fertility Survey data (1974–84) using the proximate determinants framework, demographers estimated the relative contributions of breastfeeding and artificial methods of contraception to the restriction of fertility in 29 developing countries in Africa, the Americas, and Asia. The analysis showed that in Africa the amount of potential fertility reduced by modern contraceptive use is relatively low, whereas breastfeeding inhibits an average of four births per women. In Asia, breastfeeding remains important for fertility control, but overall, modern contraceptive use is more important than in Africa. In Mexico, Central America, and South America, modern contraceptive use plays the predominant role with only one-fifth of the total fecundity reduction due to breastfeeding. In poorer developing countries like Bangladesh, Peru, and Haiti, however, breastfeeding still plays the dominant role in reducing fecundity. Similar results have been found using more recent data from the Demographic and Health Surveys (DHS) for currently developing countries. Results using the proximate determinants framework in 25 DHS countries showed that for each additional month increase in the duration of breastfeeding there is a one-half to three-fourths month increase in the duration of postpartum amen-

orrhea (for detailed treatment see Goldman et al. 1987, Guz and Hobcraft 1991, Thapa et al. 1992).

Others have developed models incorporating the underlying physiological processes of postpartum infertility. In doing so they have pointed out an important limitation of the proximate determinants approach. The proximate determinants framework uses the total duration of breastfeeding to assess postpartum lactational infecundability, when in fact amenorrhea typically ends well before breastfeeding is terminated. Thus simply showing empirically that the duration of breastfeeding is highly correlated with the duration of postpartum amenorrhea in a regression model reveals little about the underlying causal relationship between the two. Furthermore, physiological models provide a means of assessing the contribution of full, unsupplemented breastfeeding, not simply the total duration of breastfeeding used in linear regression approaches. Results suggest that for each month of exclusive breastfeeding, anovulation and/or amenorrhea would on average be postponed by *more* than one month. Nonetheless, the strong statistical relationship between the duration of breastfeeding and the duration of postpartum amenorrhea, whether measured at aggregate or individual levels, is remarkably robust, and indicates a strong underlying biological link between the two (see details in Habicht et al. 1985, Santow 1987).

2. Neuroendocrine Control of Postpartum Lactational Infertility

Thanks to a series of clinical studies carried out in a small homogeneous group of breastfeeding mothers in Edinburgh, Scotland, the neuroendocrine regulation of lactational infertility is now fairly well understood. When breastfeeding is initiated, afferent nerve impulses pass directly along nerve axes leading from the mother's nipple to the hypothalamus in the brain, directly disrupting the release of the hormones responsible for ovulation. Disruption continues until the neural suckling stimulus declines sufficiently to allow for normal hormonal release to be restored. When suckling is reduced or stopped, ovulation, fertilization, implantation, embryogenesis, and pregnancy are once again biologically possible. Therefore, it is the suckling stimulus itself that plays the key role in the neuroendocrine control of postpartum lactational infertility, not the direct effect of delayed weaning. Only in cases where weaning or the death of a child terminates breastfeeding prior to the resumption of menses can the total duration of breastfeeding, *per se*, be considered causally related to the return of postpartum menstruation.

Apart from elucidating the underlying neuroendocrine mechanisms of postpartum infertility, the Scottish studies also revealed significant new insights into the relationship between postpartum infertility and

specific aspect of infant feeding behavior. The research showed, for example, that both the number and the duration of suckling episodes played key roles in suppressing ovarian function. None of the Scottish women in the study ovulated if they breastfed six or more times and more than 60 minutes a day. The introduction of supplemental foods prompted a decline in both suckling frequency and duration with suckling duration declining somewhat earlier than frequency.

The more abrupt the introduction of supplements the more rapid the decline in duration compared to frequency and the more rapid the resumption of ovarian function. Women who maintained regular night-time feeds tended to breastfeed with a greater frequency and longer duration and to remain anovulatory for a longer period of time than those who did not. Overall, supplementation undermined and reduced the intensity of the infant's suckling stimulus (however measured) and led to an earlier resumption of ovarian activity and menstruation than if exclusive breastfeeding were maintained (for details see Tay et al. 1996).

3. Hazard Model and Life Table Studies of the Fertility Inhibiting Consequences of Breastfeeding

To date only a handful of large-scale, longitudinal studies, including those in India, Taiwan, Bangladesh, Guatemala, Indonesia, and the Philippines, have been undertaken to investigate, either directly or indirectly, the determinants of postpartum lactational infertility in nonclinical settings. The strength of these studies derives, in part, from their prospective data collection frameworks and their measurement of nursing behavior. The investigations lasted for periods of one to three years or longer in some cases. They prospectively collected information on infant nursing behaviors, supplemental feeding practices, maternal and infant health, the duration of postpartum amenorrhea and birth interval, as well as a rich array of other socio-demographic and control variables.

An additional strength of these prospective studies arises from their use of an alternative body of statistical techniques known as event history or hazard model analysis. These statistical procedures have been developed since the 1980s specifically to handle duration or event variables, censoring (when a subject, for whatever reason, leaves a prospective study before it has ended or the event of interest does not occur before the study ends), and other problems inherent in longitudinal study designs. Hazard model analysis provides the ideal statistical framework for unambiguously quantifying the effects of nursing behavior on the duration of postpartum amenorrhea in non-clinical settings while simultaneously controlling for the joint effects of confounding variables (see *Event-history Analysis in Continuous Time*).

To cite one example, women living in rural Indonesia were interviewed monthly about their current menstrual status and their average daily breastfeeding pattern. The women were classified into four groups based on nursing intensity. Univariate life-table analysis revealed that low-intensity nursing mothers resumed menses at 11.5 months, while high-intensity nursers resumed menses at 20.4 months. Medium-low and medium-high intensity nursing groups resumed menses at 14.0 and 17.7 months, respectively. A series of multivariate proportional hazard models showed that, when controlling for confounding factors, the higher the intensity of suckling the slower the rate of return of menses. Further multivariate analysis also showed that partial breastfeeding (with supplements) significantly increased the risk of resumption of postpartum menstruation when controlling simultaneously for nursing intensity and other confounding factors.

In the Philippines, research also demonstrated that partial (supplemented) breastfeeding significantly increases the risk of early return of postpartum menstruation. Hazard model analysis showed that factors such as active suckling, the use of two breasts versus one, and breastfeeding on demand vs. a fixed nursing schedule significantly decreased the risk of early return of postpartum menstruation. By contrast, the feeding of milk other than breast milk and nonnutritive or low caloric nonmilk liquids increased the risk of early resumption of postpartum menstruation.

In the final analysis, these longitudinal studies and the clinical research from Edinburgh share an emergent consensus concerning breastfeeding and fertility. The consensus is that, in spite of the wide range of variability in infant feeding patterns and practices across populations and individuals, it is the sum total of all direct and indirect breastfeeding activity affecting the intensity of the suckling stimulus, and thus the strength of the neural input to the brain, that is the most important determinant/mediator of postpartum infertility. (For references and details see Jones 1989, 1990, Popkin et al. 1993, Zohoori and Popkin 1996.)

4. Breastfeeding and Family Planning

The emergent scientific consensus regarding the reproductive consequences of breastfeeding represents a marked departure in thought and sophistication from that of just a few decades earlier. After carefully reviewing the evidence at hand, a group of researchers meeting in Italy in the late 1980s issued a statement regarding the contraceptive role of breastfeeding. Known as the Bellagio Consensus Statement, the announcement noted that, if three conditions prevail, postpartum mothers could expect a pregnancy rate of less than 2 percent. These three conditions are: a woman must remain amenorrheic since delivery, she must be fully or nearly fully breastfeeding (little or

no supplements), and she must be within six months of delivery. This is referred to as the lactational amenorrhea method (LAM) of family planning (for details see Kennedy et al. 1989, 1990).

5. Conclusion

The accumulation of scientific evidence since the 1960s has confirmed what many, especially women, have known intuitively over vast stretches of human history and prehistory: that breastfeeding acts as nature's way of keeping human fertility rates low and of guaranteeing enough time between births for the child to be adequately nourished and developed and for the mother to sufficiently recover from the rigors of child bearing and rearing. Prospective research undertaken in both clinical and real-life settings has confirmed the contraceptive effects of breastfeeding. Knowledge gained from these studies has led to the development of guidelines to enable breastfeeding mothers to use lactation as a means of fertility control, at least during the first six months postpartum, or possibly up to 12 months postpartum, if mothers are actively pursuing breastfeeding and are receiving appropriate family planning counseling.

Worldwide it is estimated that more than 90 percent of mothers with infants still breastfeed, at least for a short period of time. If women were to be counseled on LAM use, this could potentially provide a natural way to prevent pregnancy immediately after birth, much as it has for most of humankind's evolutionary history. Unfortunately, however, the initiation, intensity, and duration of breastfeeding are declining due to urban lifestyles and other changes in the postindustrial years. The recent Demographic and Health Surveys have shown that today in many developing countries young women who are educated, urban, and affluent are abandoning breastfeeding in favor of bottle-feeding. Often these women are the trend setters in their communities, and their influence is spreading to the other urban and rural poor women who are the least likely to have access to or use modern contraceptives and whose infants have the most to gain from breast milk.

In recent years, however, the advent of the HIV pandemic presents a potentially more dangerous public health problem than declining rates of breastfeeding. Even though HIV, the virus that causes AIDS, can be transmitted by breastfeeding, WHO and UNICEF have advised that where other infectious diseases and malnutrition are the primary cause of infant deaths, breastfeeding should remain the standard advice to pregnant women, including those who are known to be HIV-infected, because their baby's risk of becoming infected through breast milk is likely to be lower than its risk of dying of other causes if deprived of breastfeeding. Whether or not this advice can or should be implemented is currently a matter of strong debate. Only time and more research will be

able to resolve these important social, medical and ethical issues (see details in Kennedy et al. 1990, Family Health International 1992, World Health Organization Global Programme on AIDS 1992, Trussell et al. 1992; see also *HIV and Fertility*).

See also: Family Theory and the Realities of Child-bearing Behavior; Family Theory: Complementarity of Economic and Social Explanations; Fertility Control: Overview; Fertility: Institutional and Political Approaches; Fertility: Proximate Determinants

Bibliography

- Bongaarts J, Potter R G 1983 *Fertility, Biology and Behavior*. Academic Press, New York
- Family Health International 1992 Breastfeeding. *Network* 13
- Goldman N, Westoff C F, Paul L E 1987 Variations in natural fertility: The effect of lactation and other determinants. *Population Studies* 41: 127–46
- Gray R, Leridon H, Spira A (eds.) 1993 *Biomedical and Demographic Determinants of Reproduction*. Clarendon Press, Oxford, UK
- Guz D, Hobcraft J 1991 Breastfeeding and fertility: A comparative analysis. *Population Studies* 45: 91–108
- Habicht J-P, Davanzo J, Butz W P, Meyers L 1985 The contraceptive role of breastfeeding. *Population Studies* 39: 213–32
- Jones R E 1989 Breast-feeding and post-partum amenorrhoea in Indonesia. *Journal of Biosocial Science* 21: 83–100
- Jones R E 1990 The effect of initiation of child supplementation on resumption of post-partum menstruation. *Journal of Biosocial Science* 22: 173–89
- Kennedy K I, Fortney J A, Bonhomme M G, Potts M, Lamp-
tey P, Carswell W 1990 Do the benefits of breastfeeding
outweight the risk of postnatal transmission of HIV via
breastmilk? *Tropical Doctor* 20: 25–9
- Kennedy K I, Rivera R, McNeilly A S 1989 Consensus state-
ment on the use of breastfeeding as a family planning method.
Contraception 39: 477–96
- Knobil E, Neill J D, Ewing L L, Greenwald S, Markert C L,
Pfaff D W (eds.) 1994 *The Physiology of Reproduction, Volume*
2, 2nd edn. Raven Press, New York
- Leridon H, Menken J (eds.) 1977 *Natural Fertility*. Ordina
Editions, Liege, Belgium
- Leridon H 1977 *Human Fertility: The Basic Components*.
University of Chicago Press, Chicago
- Popkin B M, Guilkey D K, Akin J S, Adair L S, Udry J R,
Flieger W 1993 Nutrition, lactation, and birth spacing in
Filipino women. *Demography* 30: 333–52
- Rosa F W 1975 The role of breastfeeding in family planning.
Protein Advisory Group Bulletin 5: 5–10
- Tay C C K, Glasier A F, McNeilly A S 1996 Twenty-four hour
patterns of prolactin secretion during lactation and the
relationship to suckling and the resumption of fertility in
breast-feeding women. *Human Reproduction* 11: 950–5
- Thapa S, Kumar S, Cushing J, Kennedy K 1992 Contraceptive
use among postpartum women: Recent patterns and pro-
grammatic implications. *International Family Planning Per-
spectives* 18: 83–92
- Thapa S, Short R V, Potts M 1988 Breast feeding, birth spacing
and their effects on child survival. *Nature* 335: 679–82
- Trussell J, Grummer-Strawn L, Rodriguez G, VanLandingham
M 1992 Trends and differentials in breastfeeding behavior:

evidence from the WFS and DHS. *Population Studies* 46:
285–307

Wood J W 1994 *Dynamics of Human Reproduction: Biology,
Biometry, Demography*. Aldine De Gruyter, New York

World Health Organization Global Programme on AIDS
(WHO/GPA) 1992 Current and Future Dimensions of the
HIV/AIDS Pandemic: A Capsule Summary. WHO, Geneva

Zohoori N, Popkin B M 1996 Longitudinal analysis of the
effects of infant-feeding practices on postpartum amenorrhea.
Demography 33: 167–80

R. E. Jones

Fertility: Institutional and Political Approaches

Much of fertility analysis is based on the model of an autonomous actor like the one assumed in micro-economic theory, characterized by a preference set against which responses to price and other signals are determined. In fertility studies, the actor is a woman (rarely a man) of reproductive age or a composite 'couple.' Fertility surveys relating individual child-bearing patterns to individual characteristics, family-size desires, and practice of birth control help to maintain this focus. Fertility, however, is also a societal process and outcome. It is affected by the powerful close-range forces at work within the micro-society of the family. Moreover, both individuals and families are enmeshed in larger social groups—clans, villages, and other communities, castes, social classes, religious groups, nations—which may affect their fertility preferences and are the locus of institutional constraints within which their fertility-related decisions are taken. The ways in which fertility is influenced by these broader social organizational and institutional settings, and by the patterns of social control that operate within them, are the subject of this article.

1. Institutional Determinants of Fertility

Institutions are enduring clusters of behavioral rules, known but often unwritten, governing human actions and relationships in recurrent situations. They are not rigid constraints requiring compliance, though there may be external or self-imposed sanctions for violation. They are not unchanging: they are constantly being made and remade by those coming into contact with them—and sometimes by deliberate efforts at policy design. They have to do with 'structure,' but not only in the sense of tangible reality. Cultural values and ideational systems may also be institutionalized.

Marriage and the family are familiar examples of institutions—more precisely, the sets of behavioral

rules and expectations characterizing marriage and the family. Clearly, both are closely linked to fertility outcomes. Other institutions that potentially influence fertility, some more distantly, include the body of family and property law, whether customary or codified; the formal authority structure in the community and local government; the stratification system, and the mobility paths (class- and gender-specific) that it supports; public systems of interage transfers, such as child allowance and pension schemes; and the labor market, particularly as it affects employment of women. The forms such institutions take, emerging from each society's unique history, support a 'fertility regime'—a distinctive pattern of family formation and array of fertility preferences, behaviors, and outcomes. This can be illustrated by some examples of different regimes.

1.1 Social Regulation of Fertility in Preindustrial Agrarian Societies

Many societies facing static land resources and productivity, and recurrent high mortality found institutional means of keeping population growth low, with fertility well below its maximum level and able to respond promptly to mortality crises. In English parishes, for example, prior accumulation of property was required before a person could marry and set up an independent household; in periods of economic hardship, saving took longer and marriage was delayed. A similar kind of community-level social control of household formation existed in preindustrial Japan. In China, social pressures, particularly from the family or lineage, applied not to marriage (which for women was near to universal) but to fertility within marriage, through expectations of lengthy interbirth intervals. Other institutions—'service' in England (sending children to work for a period as servants in other households), adoption in China—helped to smooth family consumption and production imbalances (see Smith 1981, Lee and Wang 1999; *Family Systems in Europe*).

In most traditional societies it is possible to discern distinctive patterns of institutional regulation of fertility that yield somewhat comparable demographic outcomes. The patterns vary with the nature of the family system, with the strength or weakness of territorially-based communities, with the kind of economy, and with the structure and roles of the larger political and administrative systems in which the local society is set. Very high fertility is not a feature of such societies: where it occurs it usually indicates the disruption of customary, institutionalized behavior.

1.2 Institutional Factors in Demographic Transition

Early accounts of and explanations for demographic transition—the secular declines in mortality

and fertility from high to low levels—were largely cast in institutional terms. Social institutions like those just discussed constituted a normative order that sustained high fertility in the face of the prevalent high-mortality risks. As those risks lessened and as technological advances created new economic opportunities, that normative order was slow to change. Hence the pressures and incentives facing individuals tended for a time to maintain the earlier patterns of fertility. Ultimately, however, families discerned their economic interest in having fewer children and acted accordingly. Institutional control was supplanted by rational individual or family-level decision making (see Szreter 1993).

More recent explanations of the transition do not emphasize this supposed shift from institutional to personal and conscious control. Fertility decisions are seen as rational (or boundedly rational) throughout—in the sense of being related to the actor's knowledge and perceived constraints and options at the time. Fertility is also institutionally conditioned throughout: those constraints and options are set by the society's institutional configuration as well as by its members' individual backgrounds and circumstances. This is the case no less in periods of rapid social and economic change than in the 'stable' pre-industrial situation. Indeed, such change is in large part a change in that institutional setting (see North 1990). For example, industrial development creates new avenues of social mobility, but ones that require education and often relocation to urban centers; the social capital of kinship and other ties within the community is supplanted by human capital, transforming the economics of the family. High fertility is increasingly seen as inconsistent with the new conditions. (For various accounts of demographic transition as institutional change see Davis 1963, Caldwell 1982, McNicoll and Cain 1990, *Fertility: Political and Political-Economic Perspectives*.)

Demographic transition in turn induces institutional change. The faster population growth occurring with sustained high (or even rising) fertility but falling mortality early in the transition is a factor promoting change to more intensive agricultural systems (Boserup 1965). A rapid decline from high fertility levels, such as took place in many countries since the 1960s, substantially lowers child dependency rates (its effect in increasing old-age dependency is felt only later) and provides a boost to development performance.

1.3 Institutional Factors Perpetuating Low Fertility

The social pressures that support high fertility do not long survive in the conditions of urban industrialized societies. Other institutional influences come to bear on fertility with the opposite effect. The labor market offers new career opportunities to women, but ones difficult to combine with childbearing. With women's

greater economic independence the institution of marriage erodes; the individual rather than the family becomes the basic unit of society in matters such as public finance. The state intrudes further in the education of children, lessening parental roles in socialization; more generally, peer influence becomes more significant, intergenerational influence less so. At the same time, the cost of childraising—both the direct cost of education and the opportunity cost of forgone income and leisure time—increases markedly in the new knowledge-based economy, much of it having to be borne by the parents.

There is little agreement on how low fertility will become in modern affluent societies, and whether there are inherent forces that might cause it to rebound (see *Demographic Transition, Second*). Some observers foresee an emerging retreat from individualism and restoration of family and community roles as a natural cyclical phenomenon (Fukuyama 1999). Others point to innate biological drives that may sustain fertility. But many see little likelihood of a return to near-replacement levels. Once the lingering effects of earlier high fertility have worked their way through the age distribution, however, the prospect of natural population decrease of 20 percent or more per generation is likely to provide a strong rationale for pronatalist interventions by the state in the design of social and economic arrangements (see Chesnais 1996, Sect. 3.3).

1.4 Linking Institutions to Fertility

Fertility at its simplest is a summary measure of reproductive outcome. Institutional arrangements in a society sometimes deal directly with that outcome, as in systems rewarding or penalizing births in a family beyond a certain prescribed number (China's one-child policy, for instance). But fertility is also a process—a series of individual decisions and biological events—providing many opportunities for indirect institutional influence on outcomes.

The biosocial process leading to a birth can be decomposed into its elements, the so-called intermediate variables or proximate fertility determinants such as age at marriage, duration of breastfeeding, and use of contraception. Each such element is potentially subject to separate institutional influence. Where an intermediate variable is distant from the fertility outcome, the effect on fertility may be merely incidental. For instance, the institutional shift from arranged marriage to love marriage is likely to be accompanied by a rise in average ages of marriage and first birth, tending to depress fertility. Labor market changes affecting women's employment may lead to shortened breastfeeding and hence (through shortened amenorrhea) to higher fertility—to be offset eventually by greater use of contraception (see Davis and Blake 1956).

If the intermediate variables are strongly interdependent, with individuals or couples having some given family size goal in view, subject to economic circumstances, then the main institutional determinants of fertility are those that affect that goal or its attainment. The intermediate variables in this case lose much of their behavioral interest. Where the preferences of one spouse differ from the other, changes in power relations within the family, or in the broader gender system in the society, may be a significant source of fertility change. Greater gender equality is widely believed to promote lower fertility. Other institutional effects on fertility goals would be those that change the perceived costs and benefits of children to parents or, more fundamentally, alter people's concept of the family and its role in the society. Effects on the attainment of fertility goals would include any overt social or political pressures bearing on reproductive behavior (see McNicoll 1994, Potter 1983).

2. Origins and Maintenance of Fertility Regimes

The theory of institutions investigates the origins of conventions and other institutionalized behavior, drawing particularly on transaction-cost economics and game theory. Institutional change bears an analogy to technological change, in being characterized by path dependence and possibilities of multiple equilibria. Some of this theory illuminates fertility regimes.

2.1 Transaction Cost Approaches to the Family

Transactions among members of a society entail 'transaction costs' associated with advertising or acquiring information, negotiating terms, and enforcing contracts. Social institutions, such as a banking system, may lower such costs and thus make market transactions more efficient. It is also possible to economize on transaction costs by internalizing some transactions within an enduring group or organization, members of which come to know and trust each other and thus do not need to fully specify the contractual terms. The firm is such an organization; so is the family (see Ben Porath 1980). The kinds of transactions that are best dealt with in such organizations are those where the identity of the parties matters. Examples are exchange of labor services in situations where effort cannot readily be measured, and parental investment in a child's education in expectation of a much later return.

The family, from this perspective, is the most important institution for nonmarket transactions. Its form is influenced by the emergence of market institutions, such as savings institutions, that lessen its transaction-cost advantages. The implicit contract between parent and child is one such intrafamily

transaction subject to change over the broader course of institutional development in the society, with strong implications for fertility. (See Pollak 1985.)

2.2 *Path Dependent Institutional Change*

Fertility, however, is not sufficiently described as an output of the family treated as a miniature firm-like organization. Family and gender systems require a fuller institutional specification to accommodate their distinctive patterns and fertility outcomes. The persistence of such systems over many generations in pre-industrial societies suggests that they were stable solutions to fundamental problems of organizing reproduction and dealing with child and old-age dependency. Such problems include how children are to be protected and socialized, how sexuality is regulated and parenthood affirmed, how claims to property are maintained and transmitted, and how the dependent elderly are supported. Each admits of many different solutions. Institutionalized, some combinations of those solutions dominate over time, partly through characteristics of the resulting society—among them, its reproductive success—and partly perhaps by fortuity. Belief systems develop alongside such structures, generating explicit rules of practice and sanctions for violation. Ties beyond the immediate family are an essential means of risk-spreading in high mortality conditions, and the reciprocal relationships thus established acquire cultural validity as a ‘moral economy.’ Social groups or statuses advantaged by a given structure have evident interests in its maintenance, entrenching a system of social control.

Such an account suggests that variant family and gender systems are stable institutional equilibria, and their resulting fertility regimes are to some degree fortuitous. Different regimes subjected to the same pressures for fertility change (new resources, new technologies, larger surviving birth cohorts) would respond differently, some more resistant to change than others. Variant paths of fertility transition are in fact observed.

The insight is quite general. A society’s institutional structure is historically contingent and thus in some measure accidental. Institutional change, like change in technology, is path dependent: the direction of change is constrained not merely by its existing state but by its history. Institutions, in part the residues of past social action, limit the possibilities of subsequent social action (see David 1993). But change is also influenced by hopes and expectations about the future, which are not just reflections of immediately perceived options but are formed out of information about the world and the opportunities it seems to present. A role for expectations introduces the possibility of bandwagon effects in social change, as individuals, in making their own decisions, take account of how they anticipate others will decide. Fertility outcomes derive from both history and expectations.

Fertility policy, of course, is also path dependent. Governments do not have a blank slate on which to make policy changes in accordance with the best advice from experience elsewhere; they deal with the politically feasible, meaning a situation inherited from the past.

2.3 *Institutions and Culture*

One strand of economic theorizing on fertility assumes a largely ‘institution-free’ world, in which, in Becker’s (1976) words: ‘all human behavior can be viewed as involving participants who maximize their utility from a stable set of preferences and accumulate an optimal amount of information and other inputs in a variety of markets.’ The need to specify preferences, or tastes, introduces a domain of values, hence presumably a cultural system, but one kept in the background. A second strand, institutional economics in the manner of Ben Porath, or North, derives from the analysis of transactions and iterated games. It is structurally more sophisticated but at its core assumes the same model of a utility-maximizing agent. Institutions, by implication, are economic institutions—or at least institutions amenable to economic analysis.

A consequence of this situation is that an institutional perspective on fertility is often contrasted with a cultural or ideational perspective, as if the two were entirely distinct. The antithesis is false. Even in the study of economic development, where resources and technologies are dominant realities, economic incentives and constraints are necessarily interpreted through a cognitive lens. Controversies over the role of the Protestant ethic or the Confucian ethic in promoting development are not just about abstract values, but about rules of conduct. For fertility, the relevant systems of social control typically combine tangible rewards and sanctions with the symbolic. Ideational systems provide a structure of meaning and a framework against which to evaluate life-course opportunities and options, to determine the scope of tradeoffs in economic calculation, and (internally) to assess gratification or guilt. Where those ideational structures are widely shared, behavior that accords with them can be described as institutionalized. (For discussions of the role of cultural systems and ideational change in fertility determination see Kreager 1986, Kertzer and Fricke 1997. See also *Fertility Transition: Cultural Explanations; Family Theory: Complementarity of Economic and Social Explanations.*)

3. *Governance and Fertility Change*

Since societies, and nations in particular, have an evident interest in their demographic futures, fertility (along with migration) is an object of policy attention.

For most societies most of the time it is not high on the policy agenda, but it typically becomes so in situations of very high or very low fertility. However, governments are routinely engaged in maintaining and modifying systems of law and administration which, though having no deliberate demographic intent, have a significant influence on fertility.

3.1 Incidental Effects of Government Action

There is a wide consensus on the kinds of political and administrative setting that are conducive to economic growth. In establishing such a setting a government is likely also, but incidentally and perhaps unintentionally, to be promoting fertility change along the path of demographic transition. Individuals in seeking economic opportunity are likely to see low fertility as advantageous. A striking illustration of this is seen in those Western countries, including the United States, which attained close to replacement-level fertility early in the twentieth century in the face of strong government disapproval of birth control (see Watkins 1991).

In modern welfare states a range of other government measures bears unintentionally on fertility. Elaborate social policy objectives are pursued through family law, employment regulation, and tax expenditures. Recognition of individual rights and entitlements, translated into social policy, sometimes works against the interests of families. The general social interest in reproduction—the issue of bearing and raising the next generation—is left to the private actions of individuals and families, with no assurance that the latter will aggregate to the social interest (see Demeny 1986).

3.2 Political and Administrative Pressures to Lower Fertility

Many governments pursuing economic development have sought to lower fertility by offering economic incentives to participants (or to recruiters of participants) enrolling in birth control programs. The effects, when discernible at all, have been small. More forceful measures, however, have sometimes attained significant results. Examples are China's campaign of the 1970s for later marriage, longer birth intervals, and fewer births per woman, and China's subsequent one-child policy; and India's sterilization program during the period of rule under the declaration of national emergency in 1975–7. In these campaigns the government's political authority and administrative apparatus were drawn on to implement central policy directives that might otherwise have lain within the duties of a relatively powerless line ministry. In some other countries more routine family planning programs have also sometimes made use of the local civil administration in securing clients (see McNicoll 1997).

The personal costs of coercive measures are high. The political costs are also potentially large: they were,

for example, a factor in the change of government in India in 1977. Even authoritarian governments have limited political capital to allocate over the policy spectrum, and may be loath to expend much of it on birth control programs. A program's unpopularity at the grass-roots level can provide a source of solidarity that spills over into other kinds of opposition to government authority.

3.3 Institutional Innovation in Very Low Fertility Settings

Under a continued rate of fertility well below two children per woman, a country will experience declining population size and rapid population aging. It is likely to seek ways of raising fertility closer to replacement level. Direct state subsidies in the form of child endowments are expensive and have not proven very effective. The amounts offered typically are dwarfed by the actual costs of childraising, especially for more affluent parents; moreover, large windfall gains to those recipients who would not have required the subsidy are inescapable. Symbolic rewards have fared even worse. More promising are measures that increase the compatibility of women's (and more generally, parents') employment and childrearing. Examples are flexible working hours, provision of childcare, job-sharing, and telecommuting. These may sometimes be in the employer's interest; in other cases their cost would have to be allocated between employer and government. A third category would cover efforts to reaffirm parental interest in childrearing by providing parents with a greater tangible return to them. (For proposals of this kind see Coleman 1993 and Demeny 1987.)

The diverse efforts of the many societies likely to be seeking to raise their fertility if very low birth rates persist will provide a natural experiment that will help to determine the relative efficacy and costs of alternative measures.

See also: Demographic Transition, Second; Fertility: Political and Political-Economic Perspectives; Fertility: Proximate Determinants

Bibliography

- Becker G S 1976 *The Economic Approach to Human Behavior*. University of Chicago Press, Chicago
- Ben Porath Y 1980 The F-connection: Families, friends, and firms and the organization of exchange. *Population and Development Review* 6: 1–30
- Boserup E 1965 *The Conditions of Agricultural Growth*. Aldine, Chicago
- Caldwell J C 1982 *Theory of Fertility Decline*. Academic Press, London
- Chesnais J C 1996 Fertility, family, and social policy in contemporary Western Europe. *Population and Development Review* 22: 729–39

- Coleman JS 1993 The rational reconstruction of society—1992 presidential-address. *American Sociological Review* 58: 1–15
- David P A 1993 Historical economics in the long run: Some implications of path dependence. In: Snooks G D (ed.) *Historical Analysis in Economics*. Routledge, London
- Davis K 1963 The theory of change and response in modern demographic history. *Population Index* 29: 345–66
- Davis K, Blake J 1956 Social structure and fertility: An analytic framework. *Economic Development and Cultural Change* 4: 211–35
- Demeny P 1986 Population and the invisible hand. *Demography* 23: 473–87
- Demeny P 1987 Pronatalist policies in low fertility countries: Patterns, performance, and prospects. In: Davis K, Bernstam M S, Ricardo-Campbell R (eds.) *Below-Replacement Fertility in Industrial Societies: Causes, Consequences, Policies*. Cambridge University Press, Cambridge, UK
- Fukuyama F 1999 *The Great Disruption: Human Nature and the Reconstitution of Social Order*. Free Press, New York
- Kertzer D I, Fricke T 1997 *Anthropological Demography: Toward a New Synthesis*. University of Chicago Press, Chicago
- Kreager P 1986 Demographic regimes as cultural systems. In: Coleman D, Schofield R (eds.) *The State of Population Theory: Forward from Malthus*. Blackwell, Oxford, UK
- Lee J, Wang F 1999 *One Quarter of Humanity: Malthusian Mythology and Chinese Realities, 1700–2000*. Harvard University Press, Cambridge, MA
- McNicoll G 1994 Institutional analysis of fertility. In: Lindahl-Kiessling K, Landberg H (eds.) *Population, Economic Development, and the Environment*. Oxford University Press, Oxford, UK
- McNicoll G 1997 Governance of fertility transition: Reflections on the Asian experience. In: Jones G W, Caldwell J C, D'Souza R M (eds.) *The Continuing Demographic Transition*. Clarendon Press, Oxford, UK
- McNicoll G, Cain M (eds.) 1990 *Rural Development and Population: Institutions and Policy*. Oxford University Press, Oxford, UK
- North D C 1990 *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, Cambridge, UK
- Pollak R A 1985 A transaction cost approach to families and households. *Journal of Economic Literature* 23: 581–608
- Potter J E 1983 Effects of societal and community institutions on fertility. In: Bulatao R A, Lee R D (eds.) *Determinants of Fertility in Developing Countries*. Academic Press, New York
- Smith R M 1981 Fertility, economy, and household formation in England over three centuries. *Population and Development Review* 7: 595–622
- Szreter S 1993 The idea of demographic-transition and the study of fertility change: A critical intellectual history. *Population and Development Review* 19: 659–701
- Watkins S C 1991 *From Provinces into Nations: Demographic Integration in Western Europe, 1870–1960*. Princeton University Press, Princeton, New Jersey

G. McNicoll

Fertility of Single and Cohabiting Women

Historically, fertility outside formal marriage has been referred to variously as bastardy, illegitimacy, or out-of-wedlock childbearing, with such terms typically

connoting the inferior status of children born in such circumstances. More recently, the use of these terms has declined, in part because they presume an answer to the question of whether 'inferior status' is a necessary consequence of being born to an unmarried woman or whether some of this disadvantage might stem from other causes. Instead, terms such as nonmarital fertility or childbearing by single or cohabiting women, which carry more neutral connotations, have seen increasing usage.

1. Measurement

Measuring nonmarital fertility requires knowledge of at least two demographic statuses—fertility and marriage—and ideally a third—whether a birth occurs within a cohabiting union. By and large, identifying births as occurring within or outside formal marriage is a straightforward matter given sufficiently detailed data. However, it is useful to identify potential inconsistencies across alternative definitions of nonmarital fertility; for example, 'the fertility of single or cohabiting women' vs. 'childbearing to unwed parents.' Consider a birth conceived within a marital union in which the father dies before the child is born. Viewed one way, the child was conceived to married parents; viewed another way, the mother was single at birth by virtue of the death of her spouse. Similar issues arise for legitimated births or births to separated parents. In the latter case, because the birth occurs prior to the dissolution of a formal marital union, one might view it as a marital birth. However, the child may have been conceived by the woman and a man other than her husband, either before or after marital separation; thus, another definition might classify such a birth as nonmarital. Furthermore, note that the period between separation and divorce can be lengthy, with the timing of a formal divorce decree often a response to various factors, including the outcome (e.g., childbearing) under study. Consequently, some researchers have adopted an operational definition that shifts emphasis from formal to *de facto* marital status, thus including as nonmarital those births that occur between separation—marking the informal but behavioral dissolution of a union—and divorce—marking the formal dissolution of a union.

Additional measurement issues arise when attempting to distinguish between nonmarital births to cohabiting women and to women in neither a marital nor cohabiting union. It is sometimes difficult to determine when or whether a cohabiting union has begun, a problem similar to that posed by common law marriage in historical studies. These issues have led some to argue that cohabitation might be better conceptualized not as a discrete status but rather in terms of a continuum capturing the extent to which a couple might be said to have formed a union. Nevertheless, estimates from alternative data sources are typically in substantial agreement, providing evidence, albeit of an

indirect nature, suggesting that female respondents can report accurately on their union status at birth, even retrospectively.

As a general accounting principle, one might suppose that the number of births to males and females should, in the aggregate, be identical as when data are available from official birth registers. Still, this principle need not hold under some definitions of nonmarital fertility (e.g., when both parents are married, but not to each other) given the types of data typically available to the analyst. Moreover, comparisons of survey and registration data often reveal that marriage and fertility data supplied by male respondents are markedly inferior in quality to those from female respondents. One consequence has been that nearly all research on nonmarital fertility to date has focused on the nonmarital fertility of women. Although such an emphasis is often sensible, in that childrearing responsibilities continue to fall most heavily on women, it has nevertheless produced an asymmetry in research findings, with little being known about the men who have fathered children outside marriage.

2. Demographic Trends

Nonmarital fertility has increased substantially in both Europe and the USA. Official statistics for the USA suggest that, as of 1998, roughly one in three births (32.8 percent) were to unmarried women, a level about three times higher than that in 1970, at which time one in ten births (10.7 percent) were to unmarried women. These levels are mirrored in much of Europe. More than every third birth occurs outside marriage in France, the UK, Finland, and the former East Germany, and roughly every second birth is nonmarital in the Scandinavian countries of Iceland, Sweden, Norway, and Denmark (Kiernan 2001).

In the USA, nonmarital first births have been and continue to be prevalent among teenage women; however, the composition of nonmarital births (counting all parities) increasingly consists of those at higher birth orders to nonteenage women. As of the early 1990s, second and higher-order nonmarital births accounted for 48 percent of births to unmarried white women and fully 60 percent of births to unmarried black women. Nonmarital fertility in the USA has also increased substantially among women over age 19. In 1970, one-half of all nonmarital births were to teenage women, but by the early 1990s, nonmarital births to women aged 20 and older accounted for more than two-thirds of all nonmarital births. These period trends have been accompanied by marked declines in the likelihood of marriage following a nonmarital first birth, with some attributing this decline in marriage to changes in men's economic circumstances and others to period change in both men's and women's economic circumstances.

Nonmarital fertility in the USA continues to exhibit striking differentials by race and ethnicity. In 1998,

slightly more than one in four white births (26.3 percent) were nonmarital, while nearly seven in ten black births (69.0 percent) were nonmarital. These differences also extend to births in cohabiting unions. Births to cohabiting white women rose sharply in the last two decades of the twentieth century, accounting for virtually all of the observed increases in white nonmarital fertility (Bumpass and Lu 2000). By contrast, births to cohabiting black women accounted for a far smaller proportion of nonmarital black births. Fewer than one out of five of recent nonmarital births occurred to cohabiting black women and, over time, births within cohabiting unions have constituted a declining fraction of nonmarital births to black women (Wu et al. 2001).

Nonmarital fertility in the USA and Europe differs in several important respects. The overlap between teenage and nonmarital fertility, characteristic of the USA, is largely absent in the European context. In addition, births within cohabiting unions constitute a far larger fraction of European than of US nonmarital births. Fewer than half of US nonmarital births are to parents cohabiting at the time of the birth (Wu et al. 2001), whereas the overwhelming proportion of nonmarital births in Europe are to cohabiting couples (Kiernan 2001). As a result, recent cohorts of US children are likely to spend more years, on average, growing up without a father present in the child's household than are their European counterparts.

One might expect that the decoupling of both sexual activity and childbearing from marriage and the accompanying erosion of norms proscribing nonmarital fertility would imply that nonmarital fertility has increasingly diffused beyond poor and disadvantaged populations to the non-poor and more highly educated segments of society. The available empirical evidence is in fact quite mixed on this point. On the one hand, the prevalence of cohabiting births in Europe, particularly Scandinavia, has been interpreted by some as consistent with this view. On the other hand, the expected diffusion is difficult to detect in the USA and Great Britain, where nonmarital fertility continues to occur predominantly among those with the least education and where there is accumulating evidence of diverging trends in the sociodemographic characteristics of married and unmarried mothers (see, e.g., Wu et al. 2001, Haveman et al. 2001).

3. Functionalist, Historical, and Policy Literatures

The distinction between social and biological parenthood implicit in the discussion of measurement of nonmarital fertility was an explicit emphasis of early functionalist authors in anthropology and sociology. A classic statement is Malinowski's:

The most important moral and legal rule concerning the physiological state of kinship is that no child should be

brought into the world without a man—and one man at that—assuming the role of sociological father, that is guardian and protector, the male link between the child and the rest of the community. I think that this generalization amounts to a universal sociological law, and as such I have called it ... *The Principle of Legitimacy*. (Malinowski 1930, pp. 137–8)

Accumulating empirical evidence of violations of this principle led Goode (1960) to argue that it was more likely to be followed or enforced when fathers were high in status, as when a father possessed substantial income or assets, or when his stake in the child was high.

By contrast, historians have typically regarded nonmarital fertility as a rough empirical proxy for premarital or illicit sexual activity, although many historians have also emphasized the ebbs and flows in both the incidence of ‘bastardy’ in different historical periods and the tenuous connection between levels of nonmarital fertility and the intensity of social or moral disapproval. Moreover, while public and moral authorities often reacted to nonmarital childbearing as concrete evidence of fornication and adultery, real concern appears to have often centered on the financial burdens posed to the community by children born out of wedlock. Wells (1979), in particular, concludes that local authorities in colonial America were by and large ‘unwilling to punish children for the sins of their parents,’ noting, for example, that ‘both William Franklin, the last royal governor of New Jersey, and Alexander Hamilton were able to achieve political eminence ... despite the fact that their parents were not legally married.’

The question of whether the sins of parents should be visited on children remains highly relevant to contemporary policy debates. A continuing policy dilemma is that social and economic resources directed to children born outside marriage also in effect provide assistance to parents—it is often difficult to target social policies and provisions in ways that benefit children but not parents. This concern has been less evident among policy-makers in Europe, with many European nations having had a longer and more generous tradition of social welfare provisions. It has, however, played a central role in US policy debates, where concern that welfare programs such as Aid to Families with Dependent Children (AFDC) may have had the unintended consequence of increasing nonmarital births was a key factor in the dismantling of these programs in favor of those intended to discourage nonmarital childbearing.

The specific issue of whether more generous AFDC benefits have produced increases in nonmarital births in the USA has spawned an extensive research literature. The available evidence, constructed largely from cross-sectional US data (see Moffitt (1998) for a review), finds support for such an association, with most researchers finding that this association is small in magnitude. However, a persistent puzzle generated by these cross-sectional findings is that they appear

inconsistent with period trends in nonmarital births and welfare generosity, with nonmarital births increasing and the inflation-adjusted value of AFDC and other benefits falling between 1975 and 1995. One possibility is that there is in fact no real inconsistency between the time-series and cross-sectional evidence, in that falling AFDC benefits may have acted to reduce nonmarital births, but that other factors may have operated in the opposite direction to increase nonmarital births. Surprisingly little work has been done with respect to this issue, save for Moffitt (2001), who presents evidence in support of this possibility.

4. Socioeconomic Determinants

A consistent finding in the US literature (see, e.g., McLanahan and Sandefur 1994) is that the risk of a nonmarital first birth is associated with family structure, with the standard finding being that these risks are higher for women who resided in a single-mother family at some point during adolescence. This empirical association appears remarkably robust, for example, to adjustments for unmeasured factors shared by siblings or the tendency for women with nonmarital first births to have initiated sexual activity slightly earlier than other women.

A difficulty with many of these studies is their reliance on a ‘snapshot’ of family structure, typically measured at age 14 or 16, which makes problematic the interpretation of what exactly it is about family structure that may account for the higher observed risk of a nonmarital first birth. Somewhat surprisingly, evidence to date suggests no support for widespread expectations concerning the intergenerational transmission of nonmarital childbearing—that being born outside marriage raises the risk that a young woman will herself bear a child outside marriage (Wu and Martinson 1993, Wu 1996, Haveman et al. 2001). Instead, what appears to be transmitted across generations is the propensity to bear a child at an early age (Haveman et al. 2001). Another consistent finding is that low income and low educational attainments of parents appear substantially associated with premarital first birth risks (Haveman and Wolfe 1994, McLanahan and Sandefur 1994, Wu 1996, Ermisch 2001). Three studies (Wu and Martinson 1993, Haveman and Wolfe 1994, Wu 1996) found that turbulent family environments, in the form of repeated changes in family structure, are associated with higher risks of a premarital first birth.

5. Consequences for Women and Children

Nonmarital childbearing is commonly thought to carry adverse consequences, both for the women bearing the child and for the children born into such families. An early view capturing this sentiment (Campbell 1968) held that:

The girl who has an illegitimate child at the age of 16 suddenly has 90 percent of her life's script written for her. She will drop out of school; even if someone else in her family helps to take care of her baby, she will probably not be able to find a steady job that pays enough to provide for herself and her child; she may feel impelled to marry someone she might not otherwise have chosen. Her life choices are few, and most of them are bad.

Although this view remains highly influential in many respects, the research community has increasingly regarded it with some skepticism. For example, Furstenberg et al. (1987) reported substantial variability in the life-course outcomes for women who began childbearing while in their teenage years, many of whom had births while unmarried. These findings have caused many to question whether a woman's life script is in fact '90 percent written' as a consequence of teenage (and nonmarital) childbearing. Wu et al. (2001), using nationally representative US data, have documented similar variability for women with nonmarital births, suggesting that the results of Furstenberg et al. may also hold for more recent cohorts of women. Another basis for skepticism is the realization by researchers that previous studies may have overstated the magnitude of the negative effects of teenage childbearing. Indeed, a considerable controversy has arisen over whether teenage childbearing might be held to have any causal relationship to the later social and economic disadvantages observed for teenage mothers (Geronimus and Korenman 1993, Hoffman et al. 1993a, 1993b).

Although a large empirical literature has examined outcomes for children who resided at some point with a single parent, this literature rarely distinguishes between whether such families were formed through divorce or nonmarital fertility. Thus, surprisingly little is known concerning outcomes for children born to unmarried mothers, aside from the fact that these children are substantially more likely to grow up in poverty. As noted above, several studies have found no empirical evidence in support of the transmission of nonmarital childbearing across generations. Korenman et al. (2001) found that the higher prevalence of unintended births among unmarried mothers relative to married mothers appears to account for little of the observed differential in parenting behaviors, infant health, or child development between marital and nonmarital births. Haveman et al. (2001) have found that children born to unmarried mothers are significantly less likely to complete high school than children born to married parents.

Even less is known about the consequences of being born to cohabiting parents. Empirical studies of Great Britain and the USA (Ermisch 2001, Wu et al. 2001) have found that children born within a cohabiting union are more than twice as likely to experience the loss of a parent—most typically the father—through the dissolution of their parent's cohabiting or marital union as children born within marital unions, but this

instability appears far less true in the continental European context (Kiernan 2001). As a result, it is difficult to know whether outcomes for children born within a cohabiting union would more closely resemble those of children born to single mothers or married parents. It is also worth emphasizing that what is observed—the presence or absence of a marital or cohabiting union—is but a crude proxy for the economic, social, and emotional commitments of the parents to the child, commitments which may, in turn, vary substantially over time.

6. Conceptual and Theoretical Challenges

Unlike empirical research, theoretical work to date has taken seriously the 'two-sex' nature of nonmarital fertility—for example, how nonmarital fertility might respond to both men's and women's characteristics. Notable examples include Wilson's (1987) argument concerning a dearth of 'marriageable' males in the US black community and Willis' economic models of nonmarital fertility (Willis and Haaga 1996, Willis 1999), in which different distributions in the relative wages of men and women may produce behavioral equilibria in which men move from a pattern of high to low paternal investment in the children they father. Neal (2000) has proposed a model in which the prevalence of nonmarital fertility can vary substantially if the choice set for women is expanded to consist not only of giving birth within or outside marriage, as assumed by both Wilson and Willis, but also of the option of foregoing a birth entirely.

The study of nonmarital fertility poses other unusual challenges for both theoretical and empirical researchers. One is that nonmarital fertility lies at the intersection of several joint (and potentially endogenous) processes by which families are formed—childbearing, marriage, and, increasingly, the formation of cohabiting unions. Many researchers have only begun to grapple with the additional analytic issues that arise when including cohabitation in their behavioral and statistical models. A second challenge concerns the likely life-course dynamics underlying nonmarital fertility. Advancing knowledge is complicated by the fact that many key aspects relevant to this process are typically unobserved by the researcher, including the timing of initial (and ensuing frequency) of sexual activity and how the resolution of a nonmarital pregnancy is weighed by the woman and by other relevant actors. A final challenge is that recent trends in nonmarital fertility may well reflect an evolving response by individuals to fundamental changes in the long-term contract represented by marriage, a contract that has itself changed in response to other social forces, including the nearly universal levels of sexual activity prior to marriage, the increasingly widespread diffusion of cohabitation, changes in the nature and levels of labor force participation of men and women,

and continuing high levels of divorce and union instability.

See also: Families and Households, Behavioral Demography of; Family Theory and the Realities of Childbearing Behavior; Fertility Control: Overview; Fertility: Institutional and Political Approaches; Fertility: Proximate Determinants; Sexual Behavior: Sociological Perspective; Teen Sexuality; Teenage Fertility

Bibliography

- Bumpass L L, Lu H-H 2000 Trends in cohabitation and implications for children's family contexts. *Population Studies* **54**: 29–41
- Campbell A A 1968 The role of family planning in the reduction of poverty. *Journal of Marriage and the Family* **30**: 236–45
- Ermisch J 2001 Cohabitation and childbearing outside marriage in Britain. In: Wu L L, Wolfe B (eds.) *Out of Wedlock: Causes and Consequences of Nonmarital Fertility*. Russell Sage Foundation, New York, pp. 109–39
- Furstenberg F F Jr., Brooks-Gunn J, Morgan S P 1987 *Adolescent Mothers in Later Life*. Cambridge University Press, Cambridge, UK
- Geronimus A, Korenman S 1993 The socioeconomic costs of teenage childbearing: Evidence and interpretation. *Demography* **30**: 281–96
- Goode W J 1960 Illegitimacy in the Caribbean social structure. *American Sociological Review* **25**(1): 21–30
- Haveman R, Wolfe B 1994 *Succeeding Generations: On the Effects of Investments in Children*. Russell Sage Foundation, New York
- Haveman R, Wolfe B, Pence K 2001 Intergenerational effects of nonmarital and early childbearing. In: Wu L L, Wolfe B (eds.) *Out of Wedlock: Causes and Consequences of Nonmarital Fertility*. Russell Sage Foundation, New York, pp. 287–316
- Hoffman S D, Foster E M, Furstenberg F F Jr. 1993a Reevaluating the costs of teenage childbearing. *Demography* **30**: 1–13
- Hoffman S D, Foster E M, Furstenberg F F Jr. 1993b Reevaluating the costs of teenage childbearing: Response to Geronimus and Korenman. *Demography* **30**: 291–6
- Kiernan K 2001 European perspectives on nonmarital childbearing. In: Wu L L, Wolfe B (eds.) *Out of Wedlock: Causes and Consequences of Nonmarital Fertility*. Russell Sage Foundation, New York, pp. 77–108
- Korenman S, Kaestner R, Joyce T 2001 Unintended pregnancy and the consequences of nonmarital childbearing. In: Wu L L, Wolfe B (eds.) *Out of Wedlock: Causes and Consequences of Nonmarital Fertility*. Russell Sage Foundation, New York, pp. 259–86
- Malinowski B 1930 Parenthood, the basis of social structure. In: Calverton V F, Schmalhausen S D (eds.) *The New Generation: The Intimate Problems of Modern Parents and Children*. Macaulay, New York, pp. 113–60
- McLanahan S S, Sandefur G 1994 *Growing Up with a Single Parent*. Harvard University Press, Cambridge, MA
- Moffitt R A 1998 The effect of welfare on marriage and fertility. In: Moffitt R A (ed.) *Welfare, The Family, and Reproductive Behavior: Research Perspectives*. National Academy Press, Washington, DC, pp. 50–97

- Moffitt R A 2001 Welfare benefits and female headship in U.S. time series. In: Wu L L, Wolfe B (eds.) *Out of Wedlock: Causes and Consequences of Nonmarital Fertility*. Russell Sage Foundation, New York, pp. 143–72
- Neal D 2000 The economics of family structure. Paper presented at the Workshop on Problems of the Low-Income Population. Institute for Research on Poverty, University of Wisconsin-Madison, WI
- Wells R V 1979 Illegitimacy and bridal pregnancy in colonial America. In: Laslett P, Oosterveen K, Smith R M (eds.) *Bastardy and its Comparative History*. Harvard University Press, Cambridge, MA, pp. 349–61
- Willis R J 1999 A theory of out-of-wedlock childbearing. *Journal of Political Economy* **106**: S33–64
- Willis R J, Haaga J G 1996 Economic approaches to understanding nonmarital fertility. In: Casterline J B, Lee R D, Foote K A (eds.) *Fertility in the United States: New Patterns, New Theories*, pp. 67–86
- Wilson W J 1987 *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. University of Chicago Press, Chicago
- Wu L L 1996 Effects of family instability, income, and income instability on the risk of a premarital birth. *American Sociological Review* **61**: 386–406
- Wu L L, Bumpass L L, Musick K 2001 Historical and life course trajectories of nonmarital childbearing. In: Wu L L, Wolfe B (eds.) *Out of Wedlock: Causes and Consequences of Nonmarital Fertility*. Russell Sage Foundation, New York, pp. 3–48
- Wu L L, Martinson B C 1993 Family structure and the risk of a premarital birth. *American Sociological Review* **58**: 210–32

L. L. Wu

Fertility: Political and Political-Economic Perspectives

Since the rise of modern demography in the early twentieth century, fertility has largely been the province of demographic research. In the 1980s other fields—in particular anthropology, but also gender studies, social history, and cultural studies—began to take an active interest in childbearing as sociocultural and political-economic process. Concerned less with the numbers of children born than with the macro- and microlevel dynamics surrounding the reproduction of families, societies, and politics, these perspectives have flourished along the borders between sociocultural anthropology, population studies, gender studies, and cultural studies of science and technology. This article traces the development of three of the most vigorous streams of such research: the culture and political economy of fertility, the politics of reproduction, and Foucauldian perspectives on 'population.' Although these fields of research are multidisciplinary, because anthropologists have been at the forefront of these intellectual developments, and because of limitations of space and knowledge, this

article emphasizes anthropological work, especially that of US-based scholars.

These three perspectives are rooted in different sets of substantive concerns and theoretical preoccupations. This article describes each analytic framework in turn, tracing its intellectual and political sources, its analytic agenda and research concerns, and its status at the end of the twentieth century. It then turns to some of the methodological, ethical, and political dilemmas faced by scholars conducting research on reproductive politics and political economy. Whereas conventional research places social and economic forces at the heart of fertility change, the perspectives reviewed here emphasize instead the political, political-economic, cultural, and linguistic or discursive dimensions of reproductive transformation. With their political and cultural interests, qualitative research methodologies, and concern about researcher reflexivity, these perspectives do not so much supplant as supplement the more conventional quantitative research of demography. Their emergence on the intellectual scene in the 1980s and 1990s marks a significant broadening of interest in reproduction across the human sciences. As anthropologists and other students of social life have turned their attention to reproduction, the study of fertility has grown in theoretical sophistication. At the same time, human reproduction, long peripheral to theory in the human sciences, has assumed a more central place in social, cultural, and political theory. The appearance of these fresh ways of understanding human fertility may turn out to be one of the more exciting developments in social theory in the late twentieth century.

1. The Culture and Political Economy of Fertility

The first of these broadly political perspectives, the culture and political economy of fertility, grew out of the dissatisfaction of anthropologists in the mid- to late 1980s with conventional demographic approaches to fertility. A highly quantitative social science, demography has relied on relatively simple theoretical formulations to guide its research. The field's main contribution to theorizing about fertility, demographic transition theory, posits a universal fertility 'transition,' or decline from high to low levels of childbearing, that follows urbanization, industrialization, the spread of education, and the fall in infant mortality associated with the worldwide process of 'modernization.' These changes are thought to alter the economics of childbearing, lowering desired family size. Developed in the mid-1940s, demographic transition theory was part of the larger body of modernization theory that dominated social science research on Asia, Africa, and Latin America for roughly the first 25 years after World War II. In the 1970s, the consensus surrounding transition theory began to

erode; by the early 1990s it was no longer an area of active theory building. What undermined faith in transition theory was the massive European Fertility Project, which failed to show the expected connections between the theory's socioeconomic variables and the historic decline in European fertility. The only 'variable' consistently related to fertility was culture, defined operationally as language, ethnicity, or geographical region. Accordingly, in the 1980s and 1990s, demographers increasingly turned their attention to 'culture,' understood primarily as ideational change, more specifically, diffusion of new ideas about the acceptability of modern contraception.

As the disciplinary experts on 'culture,' anthropologists were invited to join the search for the forces underlying fertility change. Yet anthropological notions of culture were embedded in disciplinary traditions that rooted individual behaviors such as childbearing in larger-scale structures and processes ranging from family economy and community, kin and class relations all the way to global processes such as the historic spread of capitalism and, today, the transnationalization of cultural, social, and economic life. Out of one such tradition, anthropological political economy, known now as the anthropology of globalism, emerged a culture and political economy of fertility that sought to broaden, politicize, and historicize research on childbearing across space and time. This body of work traces its intellectual lineage to Marxian models of global capitalism that emerged as critiques of modernization theory in the 1960s and 1970s, namely, dependency theory, world-systems theory, and the anthropological political economy of Eric R. Wolf. More recently, this perspective has drawn inspiration from theories of disorganized capitalism and flexible accumulation advanced by critical geographers such as David Harvey, as well as work on transnationalism by anthropologists such as Ulf Hannerz and Arjun Appadurai.

From the vantage point of these larger global political economies of social life, demographic research on fertility carves out too narrow and too micro a slice of social organization. It also underestimates the importance of power and politics, employs a relatively static notion of culture, has difficulty capturing human agency, and takes too unilinear, evolutionary, and progressivist a view of history. Demographic work also tends to adopt a Eurocentric stance according to which fertility transition is caused by and in turn causes further Westernization and reproductive Westernization is beneficial for all. Underlying these latter notions is the problematic assumption that Europe and its offshoots are superior to the rest of the world and the source of all significant change.

A culture and political economy of fertility takes as its object the location of reproductive subjects at the conjuncture of historically developed local, regional, national, transnational, and global processes. It directs attention to the embeddedness of community inst-

institutions shaping fertility in structures and processes operating at supracommunity levels, and to the historical roots of those macro-micro linkages. Students of the culture and political economy of fertility read demographic history as saying there are many reproductive transformations, each driven by a combination of forces that are to some extent institutionally, culturally, and temporally specific. Accordingly, they focus on specific transformations in different parts of the world, using generalizations derived from case studies as the building blocks of general theory. Aiming to create 'whole demographics' of particular sites, they attend to political and economic as well as social and cultural forces underlying reproductive change, stressing the interrelations among these. Emphasizing the political negotiation of all aspects of social life, the culture and political economy of reproduction places gender, class, race, and other forms of inequality at the heart of reproductive dynamics. It views culture as a cluster of negotiated meanings and motivations that, taken together, form larger cultural logics that shape reproductive aspirations and practices. In this perspective people are seen as active agents of their own reproduction, who ongoingly and often strategically tinker with biology in what Caroline Bledsoe (1990) calls the 'social construction of demographic reality.'

The social constructedness of reproductive outcomes illuminates the limitations of the biological notion, 'fertility,' though students of reproductive political economy generally continue to use the term because it facilitates communication with other population specialists. In the culture and political economy of fertility, history is indeterminate and contingent because it is actively constructed by situated actors. People everywhere are seen as part of an interconnected world in which non-Euro-Americans actively shape not only their own childbearing, but also the reproductive dynamics of the world at large. Clearly, the implied research agenda of such a perspective is very large. It includes such topics as the reproductive impact of the extension of capitalist relations of production, the rise of states interested in counting, surveilling, and policing their families, and the spread of family planning institutions and ideologies to the most remote corners of the third world.

While a great many authors incorporate political-economic insights into their work, the most explicit theoretical statements of the political-economic embedding of reproductive practices can be found in Greenhalgh (1990), contributions in Greenhalgh (1995), and Schneider and Schneider (1996). Exemplary book-length case studies include Kligman (1998) on political demography and the medicalization of reproductive repression in Ceausescu's Romania; Schneider and Schneider (1996) on the capitalist-driven, class-specific reproductive transitions of Sicily; Jeffery and Jeffery (1997) on the classed and gendered political economy of reproduction in rural India;

Inhorn (1994) on the devastating consequences of infertility for Egyptian women; and Kertzer and Hogan (1989), and Kertzer (1993) on the economics and church and state politics of infant abandonment and family change in Italy.

With its global reach and historical stretch, the culture and political economy of fertility presents a vast array of important research problems, few of which have yet been resolved for any time and place. In the early 1990s, however, as anthropologists working on demographic topics began to link up with those working on the political dynamics of reproduction, this field of research began to lose its distinct character and merge with other, more political, feminist, and poststructuralist (or Foucauldian) perspectives on population.

2. The Politics of Reproduction

A second stream of innovative research, the politics of reproduction, coalesced into an identifiable body of work in the early 1990s. A product of the encounter between the social sciences and post-1960s feminism, with its critical views of the reproductive status quo, this work takes issue with the biomedicalization of reproduction in the industrialized world, and with the ideologies and practices of population control in the third world, both of which treat women as objects of control rather than subjects or active agents of their own reproductive lives. The politics of reproduction is not only an intellectual field of activity, although many of its participants are university based. It is better described as a scholar-activist movement aimed at transforming the politics of reproduction in a way that places control over whether, when, and how to have children in the hands of women themselves. The development and internationalization of this feminist project on reproduction were greatly stimulated by the events surrounding the United Nations International Conference on Population and Development (ICPD) held in Cairo in 1994. At that meeting women's health advocates gained significant influence over international population policy, transforming its agenda from one based on demographic concerns with population numbers to one focused on health and equity concerns with women's sexual and reproductive well-being and rights. Work on reproductive politics, then, is rooted in both intellectual and political developments.

Two relatively distinct clusters of research on reproductive politics can be distinguished: multidisciplinary critiques of international population control activities and broadly anthropological analyses of the local/global dynamics of reproductive practices, policies, and politics. The multidisciplinary critique of international population policy combines broadly neo-Marxian concerns with contemporary capitalism and class relations with feminist concerns

with the consequences of large-scale political and economic developments for poor women. Perceiving shared political and economic interests linking transnational pharmaceutical corporations, the international population control establishment, and third world population policymakers, this view holds that the top-down, technocratic, too often high-pressured interventions of third world family planning programs sacrifice women's health and safety in the name of a neo-Malthusian ideology that wrongly blames third world poverty on population growth. The ideology of population control, it suggests, obscures the true cause of growing poverty and inequality in the world. The real source of these problems is not population growth, but the post-1973 global spread of neoliberal capitalism, with the attendant privatization of state sectors of the economy and retreat of the state from support for social programs.

Some writers believe that new, post-Cairo policies have appropriated the feminist language of ICPD to advance a newly antifeminist population agenda. Whatever their position on current policy, these writers agree that the only effective solutions to these problems are to empower women economically, install popular-based health care systems that give women a voice in their own health care, and guarantee women's sexual and reproductive rights as human rights. Important statements of these critiques and proposals are, from a southern feminist perspective, Correa (1994); US-based feminist angles, Dixon-Mueller (1993) and Hartmann (1995); and, a neo-Marxist feminist approach, Bandarage (1997). Petchesky and Judd (1998) illuminate the diverse understandings of sexual and reproductive rights held by grassroots women around the world, crucial information to guide the transformation of international population policy into policies guaranteeing women's sexual and reproductive rights.

The anthropological work steps back from these charged policy debates to reflect on the complex intersections of global-to-local politics of reproduction in different cultural settings. Scholars working in this vein are interested less in the number of children women have than in the dynamics by which child-bearing is negotiated by actors situated differently in hierarchies of gender, class, race, sexuality, and so forth. Starting with Friedrich Engels' insight about the mutual determination of reproduction and production, this work draws on the social anthropology of kinship, marriage, and parenting, adding to it a new interest in women's agency in negotiating the contradictory forces shaping their reproductive lives. Viewing reproduction as an aspect of the distribution of power in interconnected local and global arenas, this perspective pays particular attention to ways people use local cultural logics and social relations to incorporate, revise, resist, and sometimes even reshape larger-scale processes. Its aim is to produce accounts of reproductive politics that are culturally grounded,

ethnographically nuanced, and theoretically multi-layered. Key problems for research include the stratified nature of reproduction, in which certain ethnic, class, gender, sexual, and national categories of people are empowered to reproduce and nurture while others are not; the micropolitics attending the spread of Western biomedicine and its reproductive technologies worldwide; and new cultural understandings of kinship and social life that are produced as people imagine and enact the creation of the next generation. Important theoretical statements and broad-based collections on reproductive politics include Ginsburg and Rapp (1991, 1995) and Franklin and Ragone (1998). Recent case studies include Rapp's (1999) study of prenatal testing in the US, Franklin's (1997) account of assisted conception technologies in Britain, and Strathern's (1992) reflections on the implications for English kinship of the new reproductive technologies.

At the end of the twentieth century, the politics of reproduction was a burgeoning field of research activity. As new technologies for 'perfecting' reproduction continue to proliferate and spread around the world, and the politics of Cairo work themselves out in diverse transnational and southern settings, scholarly-activist interest in the gendered, raced, classed, and north-south politics of reproduction can only be expected to grow.

3. Foucauldian Perspectives on 'Population'

Emerging in the mid- to late 1990s, a third stream of political research reflects the extraordinary influence of the French 'poststructuralist' philosopher and social critic, Michel Foucault, on the human sciences in general and anthropology in particular. Foucault's writing have left their mark on both of the above perspectives, but in the work discussed here his voice is more pervasive, shaping not only some of the terms of discourse, but the central problems addressed. Given limitations of space, this article covers only a few of his most provocative ideas on the interlacing of power and population.

In *Discipline and Punish, History of Sexuality*, and other works, Foucault (1978, 1979) identified a new, modern form of power, biopower, which takes as its object life itself. Biopower operates by constituting its objects of knowledge, to which it then responds with the development of new knowledges, technologies, and arts of government. Far from being 'neutral' and power-free, the biological, medical, and social sciences are seen as parts of the apparatus of modern power. Through their scientific discourses, analytic categories, and intellectual practices, the natural and human sciences help to construct 'fertility,' 'the body,' 'population,' and 'the social,' in the process subjecting people to new, distinctly modern forms of power that affect how they constitute themselves as subjects.

Given the sciences' close relationships with bureaucratic states in most settings, scientific knowledges of reproduction have often served the political objectives of regimes in promoting such goals as socioeconomic modernity, eugenic purity, and national power. In Foucault's view, modern power is embodied in discourses and practices that are not concentrated in any institution, such as the state or medical establishment, but widely dispersed throughout the social system. Yet hegemony is never complete, for 'where there is power, there is resistance,' leading to unstable configurations of power that are always open to contestation. Even this brief review of some Foucauldian concepts makes clear that Foucault's ideas unsettle the received terms of debates over reproduction. Among other things, they make language a central focus of analysis, redefine power, problematize the categories in which reproduction is understood, and turn the sciences, long thought to be beyond politics, into objects of political study in their own right.

Anthropologists and other students of reproduction have taken Foucault's broad, Eurocentric ideas, grounded them historically and culturally, and given the human subject of biopower more agency as a political actor. In the 1990s, two interrelated programs of research on reproduction have emerged, one focusing on the individual body, the other on the social body, or population as a whole. These are the two poles around which biopower is said to operate. Foucauldian work on the individual reproductive body focuses largely on the discourses and practices of biomedicine, which 'medicalizes' ever more phases of the female reproductive life cycle, turning what were once 'normal' aspects of procreation into 'abnormal' medical conditions, which it then proceeds to treat. Viewing the female body as a site of biomedical domination and resistance, researchers have explored the medicalization of pregnancy, childbirth, and menopause, and the accommodations and subversions of women for whom medical interventions are simultaneously seductive and dangerous. Key collections of anthropological work on such issues are Lock and Kaufert (1998), which examines the 'microphysics of power' and its operation in everyday life, and Davis-Floyd and Sargent (1997), which traces how authoritative knowledge about childbirth is produced, displayed, resisted, and challenged in different social, political, and clinical contexts.

While the anthropological research tends to work in the present and from the bottom up, focusing on the daily encounters between ordinary subjects and the agents of reproductive power, it exists in productive conversation with research in cultural studies of technoscience. The technoscience studies more often work historically and from the top down, tracing the development of the biological life sciences and their construction of the discourses, practices, and technologies over which contemporary reproductive contests are waged. Two exemplary studies of this sort are

Oudshoorn's (1994) account of endocrinology and the hormonally constructed, sex-differentiated body, and Clarke's (1998) history of the scientific study of reproduction in biology, medicine, and agriculture, which resulted in the technoscientific production of modern means of contraception.

Foucauldian work on the population as a whole focuses on the role of demography, sociology, social work, social hygiene, urban planning, and other fields in constituting inappropriate ('too high' or 'too low') fertility as a 'social problem.' Working in and with state bureaucracies, these fields first identify fertility as an object of knowledge and management, and then foster the creation of new social technologies—the census, the social survey, the urban plan, the family planning program—designed to manage it at the level of the reproductive body, the home, the family, the city, and the nation-state. A relatively new research topic, the disciplinary and bureaucratic production of 'population' has so far yielded one major monograph, Horn's (1994) study of modernity and social reproduction in 1920s and 1930s Italy. Yet the Foucauldian notion of 'population' is an exciting and important construct that is certain to produce more such studies in the years ahead.

While these three political and political-economic perspectives grew out of different sets of intellectual and political concerns, by the late 1990s, as population had moved closer to the center of social theory, the lines between them, always thinly drawn, had begun to fade and disappear. This blurring of boundaries can be expected to continue into the future, giving rise to a broad politics of reproduction and population that is multilevel in scope, multidisciplinary in approach, and politically engaged in the policy implications of the research.

4. Methodological Issues; Ethical and Political Dilemmas

With their interests in the political and cultural dimensions of reproduction, these perspectives require research methods that are more akin to those of anthropology than of demography. Conventional demographic research relies on the social and demographic survey, aimed at reducing sociodemographic behavior to a series of discrete, measurable variables whose relationships can be mathematically modeled. Such research produces knowledge that, although empirically useful, is thin and undercontextualized, individual in level, and collected in categories developed by the analyst. Led by John Caldwell, since the 1970s a small but growing number of demographers have recognized the shortcomings of survey methods for gathering nuanced data on multiple levels of social reality. In place of large-scale surveys, they have turned to 'microdemographic' research in the field. While microdemographic fieldwork marks an impor-

tant departure from the survey method, such fieldwork is usually short in duration, often conducted through translators, and rarely informed by cultural and political theories of the sort discussed above.

In contrast to these demographic approaches, anthropological work on political and cultural process entails what Tom Fricke (1997) calls 'thick, cultural demography.' In thick demography, or reproductively thick ethnography, the ethnographic ear of the anthropologist is coupled with the textual eye of the cultural analyst and the documentary eye of the historian, and all are informed by in-depth knowledge of the language, culture, history, and social relations of the area studied. Thick reproductive ethnography is based on long-term ethnographic fieldwork in one or a small number of locales. Such fieldwork generally involves the use of a variety of specific methods: participant observation; small-scale censuses and/or surveys; structured and semistructured interviews; the gathering of life histories, textual materials, and so forth. The forces of interest are interconnected processes that exist on multiple levels of social reality: social, institutional, cultural. These processes are embedded in local systems of meaning and are negotiated by local actors differently placed with regard to class, clan, and other systems of social inequality. Concerned with gathering local knowledge in culturally relevant categories, the research aims to 'situate fertility,' that is, to show how it makes sense given the historically developed sociocultural and political-economic context in which it is embedded (Greenhalgh 1995). Situating fertility is an interpretive task well accomplished through narrative modes of explanation. Such forms of explanation offer 'readings' of cultural texts, oral expressions, or social situations that give coherence to something that was poorly understood.

Students of reproductive politics and culture are especially concerned with promoting self-reflexivity in their work. By self-reflexivity is meant explicit attention to the politics of one's own research practices. Attention to one's own politics entails recognition by the researcher that, despite every effort to be objective, the social scientist inevitably is a political actor whose scholarly representations have the power to create a reproductive 'other' (often 'traditional,' non-Western, high fertility) who is different from, and somehow less than, one's (scientific, Western, low fertility) self. This recognition is based not only on critiques of objectivity in science and social science generally, but also on critical studies of actual reproductive research. Two such studies have been particularly influential. Mahmood Mamdani's (1972) early account of the Khanna birth control study in northern India suggests that the scientific apparatus underlying the large-scale promotion of family planning may be based on village research that is marked by profound misunderstanding of village life on the part of foreign researchers and deliberate fabrication of lies on the part of

bemused yet polite villagers. Mamdani found that demographic research of the sort conducted in Khanna strips facts of their social context, producing analytic results and policy suggestions that contradict the vital interests of local people. Agnes Riedmann's (1993) more recent and theoretically motivated study of demographic fieldwork among the Yoruba of Nigeria reveals the exercise of power at all stages of demographic research, from gaining entree to the study population to deciding on the questions to be asked, negotiating the interview process, analyzing the data gathered, and writing up the research results. Funded by first-world donors, Reidmann argues, 'world-system demography' serves first-world interests of prediction and control, while neglecting the concerns of third-world study populations and willfully disrupting traditional cultures and societies. In short, population research serves as a tool for cultural imperialism in which Western values of family and fertility are aggressively imposed on third-world study populations, legitimizing the purposeful limitation of family size. While these arguments misread the motivations of first-world researchers and understate the agency and resistance of third-world study populations, what cannot be gainsaid is that the stakes involved in fertility research are very high, and that research *sans* reflexivity involves a risk that the research itself may exacerbate inequality in the name of enhancing welfare.

Research like this generates suspicions about the power relations embodied in the numbers demographers use. In response to such concerns, some have called for a 'demography beyond, or even without, numbers.' In place of number collection, Nancy Scheper-Hughes (1997) proposes a critical interpretive methodology that ongoingly questions the neutrality of the research process, ponders the multiple and contradictory meanings that attach to the 'facts,' and recognizes the partiality of all truths—biomedical, demographic, and anthropological. Other students of reproductive culture and politics believe that numbers are useful and important—after all, they point out, without numbers one has no measure of the levels and trends of fertility, the subject of primary interest. Yet even the most vocal advocates of quantification insist that the numbers be gathered in a culturally sensitive manner and used in ways that are sensitive to the political implications of population research.

Closely related to these political problems are ethical dilemmas that attend research on such intimate topics as reproductive politics and culture. Ironically, in some ways the ethnographic methods that are preferred by scholars in these fields are more exploitative than the positivistic methods of survey research, for the ethnographer intrudes deep into his or her informants' private lives, leaving them vulnerable first to exposure and later to misinterpretation or even distortion of their viewpoints. The dilemma is that the more sensitive the method, the greater the ethical risk

of exploitation of researched by researcher. While there are no easy solutions to this dilemma, researchers can at least strive for ethical accountability by acknowledging that their own knowledge is both partial and positioned—that is, influenced by their social locations of class, gender, race, sexuality, nationality, and so forth. In research that spans difference—national, racial, class, and so forth—researchers can employ collective decision making about procedures, and commit themselves to give something back to the community in which they work. Petchesky and Judd (1998) describe an exemplary project of this sort. Finally, at the point that research becomes the basis for policymaking or other forms of political action, they can engage in a politics of coalition and affinity, in which apparently incompatible elements join together to promote a shared cause. Given the policy and programmatic consequences of reproductive research, ongoing reflexivity about its politics and ethics is of great importance. These issues have been worked and reworked with particular care in Brazil, where scholar-activists attentive to the political and cultural dimensions of reproduction have been unusually active in the formulation of population policy (Pitanguy 1994, Barroso and Correa 1995). Students of reproduction everywhere have much to learn from them.

See also: Family Theory and the Realities of Childbearing Behavior; Family Theory: Competing Perspectives in Social Demography; Family Theory: Complementarity of Economic and Social Explanations; Family Theory: Economics of Childbearing; Fertility: Institutional and Political Approaches; Fertility of Single and Cohabiting Women; Fertility: Proximate Determinants; Reproductive Rights in Developing Nations

Bibliography

- Bandarage A 1997 *Women, Population and Global Crisis: A Political-Economic Analysis*. Zed, London
- Barroso C, Correa S 1995 Public servants, professionals, and feminists: The politics of contraceptive research in Brazil. In: Ginsburg F D, Rapp R (eds.) *Conceiving the New World Order*. University of California Press, Berkeley, CA
- Bledsoe C 1990 The politics of children: Fosterage and the social management of fertility among the Mende of Sierra Leone. In: Handwerker W P (ed.) *Births and Power: Social Change and the Politics of Reproduction*. Westview Press, Boulder, CO
- Clarke A E 1998 *Disciplining Reproduction: Modernity, American Life Sciences, and the Problems of Sex*. University of California Press, Berkeley, CA
- Correa S 1994 *Population and Reproductive Rights: Feminist Perspectives from the South*. Zed, London
- Davis-Floyd R E, Sargent C F (eds.) 1997 *Childbirth and Authoritative Knowledge: Cross-Cultural Perspectives*. University of California Press, Berkeley, CA
- Dixon-Mueller R 1993 *Population Policy and Women's Rights: Transforming Reproductive Choice*. Praeger, Westport, CT
- Foucault M 1978 *The History of Sexuality, Vol. 1: An Introduction*. Pantheon Books, New York
- Foucault M 1979 *Discipline and Punish: the Birth of the Prison*. Vintage Books, New York
- Franklin S 1997 *Embodied Progress: A Cultural Account of Assisted Conception*. Routledge, London
- Franklin S, Ragone H (eds.) 1998 *Reproducing Reproduction: Kinship, Power, and Technological Innovation*. University of Pennsylvania Press, Philadelphia
- Fricke T 1997 Culture theory and demographic process: Toward a thicker demography. In: Kertzer D I, Fricke T (eds.) *Anthropological Demography*. University of Chicago Press, Chicago
- Ginsburg F, Rapp R 1991 The politics of reproduction. *Annual Review of Anthropology* 20: 311–43
- Ginsburg, F D, Rapp R (eds.) 1995 *Conceiving the New World Order: The Global Politics of Reproduction*. University of California Press, Berkeley, CA
- Greenhalgh S 1990 Toward a political economy of fertility: Anthropological contributions. *Population and Development Review* 16: 85–106
- Greenhalgh S (ed.) 1995 *Situating Fertility: Anthropology and Demographic Inquiry*. Cambridge University Press, Cambridge, UK
- Hartmann B 1995 *Reproductive Rights and Wrongs: The Global Politics of Population Control*, rev. edn. South End Press, Boston
- Horn D G 1994 *Social Bodies: Science, Reproduction, and Italian Modernity*. Princeton University Press, Princeton, NJ
- Inhorn M C 1994 *Quest for Conception: Gender, Infertility, and Egyptian Medical Traditions*. University of Pennsylvania Press, Philadelphia
- Jeffery R, Jeffery P 1997 *Population, Gender and Politics: Demographic Change in Rural North India*. Cambridge University Press, Cambridge, UK
- Kertzer D I 1993 *Sacrificed for Honor: Italian Infant Abandonment and the Politics of Reproductive Control*. Beacon, Boston
- Kertzer D I, Hogan D P 1989 *Family, Political Economy, and Demographic Change: The Transformation of Life in Casalecchio, Italy, 1861–1921*. University of Wisconsin Press, Madison, WI
- Kertzer D I, Fricke T (eds.) 1997 *Anthropological Demography: Toward a New Synthesis*. University of Chicago Press, Chicago
- Kligman G 1998 *The Politics of Duplicity: Controlling Reproduction in Ceausescu's Romania*. University of California Press, Berkeley, CA
- Lock M, Kaufert P A (eds.) 1998 *Pragmatic Women and Body Politics*. Cambridge University Press, Cambridge, UK
- Mamdani M 1972 *The Myth of Population Control: Family, Caste, and Class in an Indian Village*. Monthly Review Press, New York
- Oudshoorn N 1994 *Beyond the Natural Body: An Archaeology of Sex Hormones*. Routledge, London
- Petchesky R P, Judd K (eds.) 1998 *Negotiating Reproductive Rights: Women's Perspectives Across Countries and Cultures*. Zed, London
- Pitanguy J 1994 Feminist politics and reproductive rights: The case of Brazil. In: Sen G, Snow R C (eds.) *Power and Decision: The Social Control of Reproduction*. Harvard Center for Population and Development Studies, Cambridge, MA
- Rapp R 1999 *Moral Pioneers: Fetuses, Families, and Amniocentesis*. Routledge, New York
- Riedmann A 1993 *Science that Colonizes: A Critique of Fertility Studies in Africa*. Temple University Press, Philadelphia

- Scheper-Hughes N 1997 Demography without numbers. In: Kertzer D I, Fricke T (eds.) *Anthropological Demography*. University of Chicago Press, Chicago
- Schneider J C, Schneider P T 1996 *Festival of the Poor: Fertility Decline and the Ideology of Class in Sicily, 1860–1980*. University of Arizona Press, Tucson, AZ
- Strathern M 1992 *Reproducing the Future: Anthropology, Kinship and the New Reproductive Technologies*. Routledge, New York

S. Greenhalgh

Fertility: Proximate Determinants

1. Background

Fertility in human populations is determined by a broad variety of factors. Specifying the relative importance among them as well as understanding the mechanisms through which they affect fertility long eluded scientists. Davis and Blake (1956) brought about an invaluable advancement by creating a conceptual analytical framework which distinguishes 'intermediate' (direct) from social, economic, political, environmental, and other (indirect) fertility determinants (see *Fertility: Political and Political-Economic Perspectives*; *Demographic Measurement: General Issues and Measures of Fertility*; *Fertility Transition: Cultural Explanations*; *Fertility Transition: Economic Explanations*; *Fertility Transition: China*; *Fertility Transition: East Asia*; *Fertility Transition: Southeast Asia*). The principal revelation was twofold. The distinction between the two types of determinants and the penetrating insight that societal factors affecting fertility can operate only indirectly through the intermediate or proximate determinants, which directly modify fertility by affecting conception, gestation, and successful parturition. In other words, social, economic, and cultural circumstances affect human behavior regarding patterns of entering into and remaining in or exiting from stable sexual unions, and practicing breast-feeding, contraception, or induced abortion. It is this behavior which produces the level of fertility and its changes generate fertility trends.

Further crucial advancements were achieved two decades later. Bongaarts (1978, 1982)

(a) identified a simplified set of proximate fertility determinants (seven rather than eleven);

(b) established which of these are central in creating fertility differentials and trends (four out of the seven); and

(c) devised formulas to measure these factors from generally available data, and how to quantify their respective impact on fertility levels and trends in a comprehensive reasonably simple model.

The direct relationship between the proximate determinants and fertility is embodied in the fact that

the direction of change of fertility is given by the direction of change of the respective proximate determinant. If, for instance, breast-feeding declines, fertility increases. An increase in contraceptive prevalence or the practice of induced abortion will engender a fertility decline. To generalize, if a proximate determinant changes then fertility necessarily changes also, and the change is in a predictable direction, provided the other proximate determinants remain constant. This is not true for an indirect determinant. An increase in educational attainment may lead to a decline in breast-feeding which in turn signifies increased fertility. A similar rise in education in another population may cause contraceptive prevalence to increase which in turn signifies a fertility decline. Also, both of these mechanisms may be operating simultaneously in which case they cancel each other out and fertility will not change. Detailed analyses are required to determine how changes in the social, economic, cultural, religious, and environmental circumstances affect proximate determinants so that their effect on fertility trends can be understood.

2. The Proximate Determinants

The proximate determinants can best be identified by examining the events that most immediately influence the duration of the reproductive period and the rate of childbearing within it. A more detailed discussion follows a brief overview.

The beginning and end of the reproductive period are biologically determined. The potential reproductive years start at menarche, the woman's first ovulation which reveals itself in the first menstruation. The childbearing years end with the cessation of ovulation, i.e., the onset of permanent sterility, which occurs before or at menopause. The effective reproductive period is equal to the time women spend in stable sexual unions, marriages, and consensual unions. Consequently, the level of fertility is dependent directly on the length of time women spend in stable sexual unions.

The next important circumstance shaping the level of fertility is the average number of children borne by women during their reproductive life span. It is inversely related to the length of the periods between births. Short periods between births imply high fertility, i.e., a relatively large number of children born per woman, and vice versa.

The period between births has several components. A birth is followed by postpartum infecundability, a fecundable or ovulatory period, and a subsequent pregnancy. The length of the postpartum infecundability interval after a birth during which a woman cannot conceive is determined mainly by the length and intensity of breast-feeding. This period lasts through the first postpartum ovulation which is at the same time the beginning of the fecundable interval,

also called the ovulatory interval or the waiting time to conception. The length of this period is determined by fecundability, the monthly probability of conception, by the monthly risk of conception in the absence of contraception, and by the effectiveness of any contraception practiced. Once conception takes place it can be followed by a full-term pregnancy of nine months, or by an incomplete pregnancy terminated by a spontaneous or induced intrauterine death. Postpartum infecundability after an intrauterine death tends to be much shorter than following a birth.

The overview of the reproductive process identifies the following proximate determinants:

- (a) The beginning and discontinuation of stable sexual unions, marriages, and consensual unions.
- (b) The onset of sterility.
- (c) Postpartum infecundability.
- (d) Fecundability.
- (e) Contraception.
- (f) Spontaneous intrauterine mortality.
- (g) Induced abortion.

2.1 The Beginning and Discontinuation of Stable Sexual Unions

Fertility is related directly to the duration of time a woman is exposed to conception, which basically is the period she is in a stable sexual union. A widespread form of such unions is formal marriage. In Latin America and the Caribbean, in parts of Africa and in the Nordic countries of Europe consensual unions have traditionally been common. In the second half of the twentieth century informal cohabitation is increasingly practiced throughout Europe and other developed countries. Occasional sexual encounters can also have an effect on fertility, but this tends to be small. The average age of union formation varies considerably. Traditionally it has been relatively early with the exception of most European countries. There the mean age of marriage was often about 25, whereas elsewhere it was frequently around 17. In recent decades of rapid social change the mean age of marriage has been increasing almost everywhere, including the developing countries. The considerable postponement of marriage taking place in many developing countries is an important factor in lowering fertility. Trends in the formation of consensual unions may have been similar, however, data to confirm this assumption worldwide are lacking.

Proportions of women who never marry differ among populations. Such women were rare in traditional societies, whereas in the developed countries their proportions have at times been 10 percent or more.

Widowhood is an important factor modifying fertility as long as mortality is high and women bear children late in their reproductive period. With declining mortality only small proportions of women

lose their spouses before the end of the childbearing period.

In recent decades divorce rates have been increasing, mainly in many developed countries. The impact of this trend on fertility has been insignificant as remarriage is frequent.

2.2 The Onset of Sterility

The average age of menopause is around 50. Effective sterility starts several years earlier due to irregular ovulation among many aging women and the higher risk of spontaneous intrauterine death with increasing age. A certain increased sterility of spouses has an additional effect of lowering the age of couple sterility. An average age of last birth around 40 in societies not practicing contraception confirms the assumption that combined sterility sets in when women are in their lower forties.

2.3 Postpartum Infecundability

Breast-feeding is nature's contraceptive. The mother provides the infant with food and protection and the process of suckling acts as a contraceptive for the mother. In some known contemporary hunter-gatherer societies, infants are constantly with their mothers and can suckle up to several times an hour. Here births tend to be spaced four years apart without the use of any contraception. Suckling on the nipple produces nervous impulses which pass from the breast to the brain thus inhibiting the release of hormones responsible for ovulation from the pituitary gland (see Potts and Short 1999 and *Fertility Control: Prevalence and Consequences of Breastfeeding*). In the absence of breast-feeding menses return shortly after birth, on average after 1.5–2 months. In developed countries periods of exclusive breast-feeding have been reduced to only a few months. In developing countries profound changes are also underway, but the postpartum infecundability period may still be considerable. The longest average duration documented by a national survey was in Bangladesh in 1975, namely 18.6 months.

2.4 Fecundability

The monthly probability of conception, fecundability, is at most 0.15–0.25 among newlyweds and declines with increasing age and duration of marriage. It is this low because fertilization can occur only during about two days around the time of ovulation in the middle of the menstrual cycle. In addition some cycles are anovulatory and a substantial proportion of fertilized eggs fail to implant or are aborted spontaneously during the first two weeks after fertilization. Typical

delays to conception range from five to ten months. It is assumed that there are no significant differences between populations in this respect. The overall effect of involuntary sterility tends to be minor and significant variations between populations are the exception.

2.5 Contraception

Contraception is arguably the main factor modifying fertility in contemporary societies. There have been exceptions in eastern Europe where for some periods of time the practice of induced abortion was extremely high and thus more important in inhibiting fertility. In a few developing countries the use of contraception is still low, but practically everywhere it has been growing. During the second half of the twentieth century the unprecedented increase in the use of contraceptives and the availability of new modern more effective contraceptives have led to the perception that a 'contraceptive revolution' has taken place (see Westoff and Ryder 1977).

The effectiveness of contraception is measured as the percentage reduction in fecundability. For instance, a contraceptive with a 95 percent effectiveness used by women with a 0.2 fecundability will yield a 0.01 monthly probability of conception. Actual contraceptive effectiveness of a method depends primarily on the extent to which it is used as prescribed. Improper use results in contraceptive failure. The effectiveness of modern methods, such as oral contraceptives and intrauterine devices, is more than 95 percent in developed countries and that of conventional methods, such as the condom, diaphragm, or spermicides, about 90 percent. Depending on the knowledge and motivation of the average user, those levels tend to be lower in developing countries. In a number of countries female and/or male sterilization is popular and is 100 percent effective. The mix of methods used differs very much from one country to another.

2.6 Spontaneous Intrauterine Mortality

While there is no doubt that many conceptions are spontaneously aborted, it has proven difficult to estimate the magnitude of this factor. Not only retrospectively but even for women in prospective studies it is very difficult to accurately determine and report intrauterine deaths in the early stages of pregnancy. Carefully designed studies have provided estimates of 20 percent of conceptions not resulting in live births with half of these occurring before the third month of pregnancy. It is known that the rate of intrauterine mortality is the lowest in the middle of the reproductive period. Differences between societies appear to be insignificant, except with regard to stillbirths, defined as intrauterine deaths after the twenty-eighth week of pregnancy. These occur more frequently in some poor countries, 4 percent of conceptions compared with 1 percent in developed countries, presumably due to health and environmental factors.

2.7 Induced Abortions

Induced abortions have been practiced in every society and usually play a meaningful role in some stage of a country's fertility transition. The incidence depends mainly on motivation to restrict childbearing, the availability and use of contraception, and to some extent on the legal status of induced abortion (see *Demographic Techniques: Family Reconstitution*). The practice of induced abortion was exceptionally widespread in some countries of eastern Europe in the second half of the twentieth century where they were legalized earlier than elsewhere and where access to contraceptives, especially modern ones, was restricted. Around 1990 in Romania, for instance, three-quarters of all pregnancies were intentionally terminated. In contrast, the incidence was low in west European countries with easy access to modern contraceptives.

Table 1
Rating of proximate determinants with respect to sensitivity and variability among populations

Proximate determinants	Sensitivity of fertility to determinant	Variability among populations	Overall rating
1. The beginning and discontinuation of stable sexual unions, marriages, and consensual unions	+++	+++	+++
2. The onset of sterility	++	+	+
3. Postpartum infecundability	++	+++	+++
4. Fecundability	++	++	++
5. Contraception	+++	+++	+++
6. Spontaneous intrauterine mortality	+	+	+
7. Induced abortion	++	+++	+++

Source: Bongaarts (1982)

+++ = high; ++ = medium; + = low or absent.

During the second half of the twentieth century the number of countries with liberal induced abortion legislation was increasing steadily. By the end of the twentieth century over 60 percent of the world's population was living in such countries. In contrast, 25 percent resided in countries where induced abortions were generally prohibited. Almost all of the latter are developing countries, where induced abortions are performed frequently by unqualified personnel under unsanitary conditions and thus women's health is often seriously impaired.

3. The Relative Importance of Proximate Determinants

The joint effect of all the proximate variables determines fertility levels and trends. For the analysis of fertility differentials and trends only those determinants are meaningful whose effect differs in time and space. If large variations of a determinant have only a small effect on fertility trends and differentials, or if a determinant hardly varies among populations, its analysis will not yield useful scientific or policy relevant insights. Thus the selection of the analytically important factors depends on how sensitive fertility is with respect to a particular determinant and the variability of the determinant among populations. In Table 1 the proximate determinants are rated according to these criteria.

Abundant evidence confirms that fertility is the least sensitive to the risk of spontaneous intrauterine mortality and the most sensitive to time spent in sexual unions and contraception. Variability is the lowest for onset of sterility and the risk of spontaneous intrauterine mortality. Altogether, for the analysis of fertility levels and trends, the four most important proximate determinants are the time spent in stable sexual unions, postpartum infecundability, contraception, and induced abortion.

See also: Adolescent Behavior; Demographic; Family Theory; Economics of Childbearing; Fertility and Culture; Anthropological Insights; Fertility Control: Overview; Fertility Control: Prevalence and Consequences of Breastfeeding; Fertility: Institutional and Political Approaches; Fertility of Single and Cohabiting Women; Fertility: Political and Political-Economic Perspectives; Teenage Fertility

Bibliography

- Bongaarts J 1978 Framework for analyzing proximate determinants of fertility. *Population and Development Review*. **4** (1): 105-32
- Bongaarts J 1982 The fertility-inhibiting effects of the intermediate fertility variables. *Studies in Family Planning*. **13** (6-7): 179-89

- Bongaarts J, Potter R G 1983 *Fertility, Biology, and Behavior: An Analysis of the Proximate Determinants*. Academic Press, New York and London
- Davis K, Blake J 1956 Social structure and fertility: An analytic framework. *Economic Development and Cultural Change*. **4** (3): 211-235
- Leridon H 1977 *Human Fertility: The Basic Components*. University of Chicago Press, Chicago, IL
- Potts M, Short R 1999 *Ever Since Adam and Eve: The Evolution of Human Sexuality*. Cambridge University Press, Cambridge, UK
- Westoff C F, Ryder N B 1977 *The Contraceptive Revolution*. Princeton University Press, Princeton, NJ

T. Frejka

Fertility Theory: Caldwell's Theory of Intergenerational Wealth Flows

1. Introduction

John Caldwell's wealth flows theory proposes a direct link between family structure and fertility (Caldwell 1976b). According to the theory, there are only two major forms of family structure, differing principally in the direction of wealth flows among generations. In 'primitive' and 'traditional' societies, net wealth flows are primarily upward from younger to older generations, and individual interests are subjugated to corporate interests. In developed nations, family structure is organized in terms of downward wealth flows where parents are expected to provide for children's economic well-being.

The theory proposes that fertility decisions in all societies are economically rational responses to familial wealth flows. In societies with net upward wealth flows, the economically rational decision is to have as many surviving children as possible (within the constraints imposed by biology), because each additional child adds positively to a parent's wealth, security in old age, and social and political well-being. In societies with net downward wealth flows, the economically rational decision is to have no children or the minimum number allowed by a psychological disposition that derives pleasure from children and parenting. The worldwide transition from high to low fertility is the result of a change in family structures from upward to downward wealth flows.

This change in family structure was due to the spread of new values that placed a premium on individual satisfaction and achievement (Caldwell 1980). Those values emanated from the educated, middle-class in the west and are now being exported to the developing world through mass formal education. Implicit in the educational materials and expectations of schools is the individualistic value system that

produces downward wealth flows. Labor markets make the adoption of these individualistic values, which are adversative to the family or group production characteristic of pretransition societies, economically feasible. The transition from 'traditional' to 'modern' family structure occurs when a critical mass of individuals adopt the new values, and respond with low fertility. The attainment of mass education in a country should therefore precipitate and predict the fertility transition.

The most serious challenge to the wealth flows theory has come from evolutionary biologists. The argument that prior to modernization, upward wealth flows characterized human family structures is inherently antithetical to theory in evolutionary biology (Turke 1989). Biologists expect net wealth flows to be downward in all organisms, including humans. This research generated by the evolutionary biological critique of wealth flows theory has stimulated the development of a new theoretical framework for understanding human demographic patterns (see *Fertility Theory: Embodied-capital Theory of Life History Evolution*).

2. Historical Context

Wealth flows theory is not unique in proposing a link between costs of children and fertility. Notestein (1945, see *Fertility Transition: Economic Explanations*) proposed that the cost of children and their economic value are major determinants of fertility. Becker (1960) also developed formal models of the demand for children based on the household production function, focusing on the tradeoff between quantity and quality of children. Leibenstein (1957), also building on Notestein, argued that two forms of utility derived from children in pretransition (agricultural) societies were the proceeds of child labor and old age security. Evidence regarding the value of children's labor in agricultural societies was offered by Boserup (1965), Clark (1967), and Mamdani (1972). During the 1970s Cain (1978), Caldwell (1976b), Mueller (1976), and Nag et al. (1978) all developed theories concerning the economic contribution of children and the importance of child labor in peasant societies and how these factors impact family size and structure.

Caldwell's 'wealth flows' theory differs from the others by expanding the definition of intergenerational transfers across the life course and by directly linking changing value systems regarding intergenerational transfers of wealth to fertility transition. Defining wealth as 'all the money, goods, services, and guarantees that one person provides to another' (Caldwell 1982, p. 333), the theory directly incorporates effects on status and political position as well as on material wealth. Together, those material and nonmaterial transfers across the life course determine the net flow of wealth between generations and also affect societal

norms regarding the emotional value of children. As with economic models, Caldwell's theory recognizes technological impacts on wealth flows, such as the changing economic value of education. While there is unique pleasure derived from having children, pleasure will be derived from fewer children as their economic value decreases or the costs of educating them increase (Caldwell 1982, p. 338). However, unlike the more strictly economic models (Becker 1981), value systems regarding support in old age and the relationship between the individual and the larger family are the primary determinants of intrafamilial wealth flows in Caldwell's view.

3. The Economic Contribution of Minor Children

Wealth flows theory has been tested in several ways. Some tests are designed to determine if changes in the economic roles of children are associated with changes in fertility. Several researchers have found evidence of such effects in sub-Saharan Africa. In addition to Caldwell's (1976a) original microlevel study in southwest Nigeria, DeLancey (1990) and Makinwa-Adebusoye (1994) both found through meta-analysis of sub-Saharan African national level survey data that parents perceived upward wealth flows in high fertility contexts.

Similar results have been reported in several Asian countries with national level data. Chang (1990) reports that the adoption of socialist economic principles in China may reduce the economic contribution of children and result in fertility decline, but Moore (1998) suggests recent economic policy initiatives encouraging family economic independence may promote higher fertility. Some studies, using micro-level data, do not find evidence of upward intergenerational wealth transfers in high fertility contexts in India, (Das Gupta 1994 (in Punjab), Vlassoff 1982).

Willis (1982) has criticized Caldwell's hypothesis strongly on theoretical grounds, arguing that members of an altruistic older generation will see offspring success as a substitute for their own, and will therefore be willing to invest in offspring human capital at a higher rate than would be predicted if parents were acting solely in their own self-interests.

In general, the tests discussed above are weak because they do not measure the absolute flow of wealth but rather indicators of changes in the economic contributions of children. Strong tests of the theory require that children provide *net* economic benefits (i.e., amount provided minus amount received) to parents in high fertility societies and net economic costs in low fertility societies. Studies that have actually measured the net flow of wealth in high fertility societies do not support a strict interpretation of the theory. On the basis of data on net production and consumption through the life course, Kaplan (1994) has shown that in three lowland South Amer-

ican indigenous populations engaged in traditional economic pursuits of foraging and swidden horticulture, offspring do not become net producers until their twenties and never compensate their parents in terms of economic investment. Mueller (1976) in a study of 27 peasant agricultural societies found that while children were able to contribute to the household economy at a far higher rate than those in Kaplan's study, they still do not become net producers until their late teens and never repay the initial parental investment. In a review of all published studies, Lee (1996) concludes there are no data showing a net wealth flow from children to parents for any society—ironically, net upward wealth flows between generations are found in *low* fertility societies when public transfers, such as Medicare and social security, are included (Lee 1996).

4. *Old Age Security*

According to wealth flows theory, a second period in the life course for parents to receive economic benefit from their offspring is during old age. Those benefits motivate high fertility in traditional societies where adult children are expected to support their aging parents. In modern society, the ethic of individualism and public supports for the aged remove this motivation. Empirical evidence bearing on this hypothesis is also mixed. Hugo (1997), using national level data in Indonesia, found high fertility to be associated positively with perceptions regarding the likelihood of receiving old age support. Using historical data from England and Wales, Schellekens (1993) also found that old age security was correlated with high fertility. Cain has argued that children provide risk insurance for the elderly in south Asia (Cain 1981). Using micro-level data from Bangladesh, he also suggests that women desire high fertility to receive support from sons in old age as insurance against the loss of support from husbands in this patriarchal society. In contrast to Caldwell, however, Cain also argues that the breakdown of extended kin networks makes women even more dependent on sons for security. Vlassoff and Vlassoff (1980), using microlevel data from India, found no evidence that old age security motivated high fertility levels. (See Dharmalingam (1994) for a review of the debate between Cain and the Vlassoffs.)

Several studies using national level data have found that there are high expectations or levels of old age support in peasant agricultural societies in sub-Saharan Africa (De Lancey 1990), Mexico (regional level data) (Nugent and Gillaspay 1983), Malaysia (Lillard and Willis 1997), and Taiwan and the Philippines (Lee et al. 1994). Many studies have found extremely low levels of support or expectations of support of the elderly by adult children in the United States and other industrialized countries (Eggebeen 1992). None of these studies in the developing or

developed world were explicitly concerned with the relationship of fertility to old age security.

Critiques of wealth flows theory concerned with the economic contribution to the household of minor children are equally appropriate with regards to a net upward flow of wealth realized as support of the elderly. The evidence in support of wealth flows theory is based upon associations of high fertility with expectations of support. There is no solid evidence about whether high fertility actually increases the likelihood of support at the individual level, nor about the net flow of goods and services between older people and their descendent kin in high fertility societies. In fact, there is some evidence that older men and women make contributions to household economics in peasant agricultural societies through the skills and knowledge they bring to crisis situations. Researchers working with hunter-gatherers and forager-horticulturalists also observe that elderly individuals are often net producers and engage in other forms of work to support members of younger generations (Hawkes and Jones 1997).

5. *Mass Education*

Education plays a prominent role in many theories of fertility transition. In the wealth flows theory, education affects fertility through three routes, all affecting the net costs of rearing children. First, education subverts traditional communally-oriented value systems and trains children to pursue individual goals at the expense of supporting their parents (see also Cochrane 1979); second, it reduces children's ability to contribute to the household economy through diversion of time and loss of traditional skills and knowledge (Bock 2001, Caldwell 1980); and third, it increases the direct and indirect costs of children by increasing their demand for consumer goods and by establishing childhood as a period for receiving investment rather than contributing to the household economy (Caldwell 1980).

While there is a great deal of empirical evidence regarding the relationship between education and fertility, less is known about the relative contributions of these alternative causal routes, and particularly about education's impact on fertility through changing values regarding familial wealth flows. The empirical evidence regarding the impact of education on values is mixed. A meta-analysis of national level data from 37 sub-Saharan African countries found that primary school enrollment for both males and females in 1960 and 1980 had no significant effects on total fertility rate 15 and 30 years later. Secondary school enrollment for males and females in 1960 had a weak negative effect on total fertility rate, while only male secondary school enrollment in 1980 had a strong negative effect on actual and estimated total fertility rates 10 to 25 years in the future (Adamchak and Ntseane 1992).

Several studies show that children's school attendance affects parental fertility, (see, for example, Axinn (1993)). However, the causal process underlying this association is unclear, especially since parents who actively seek formal education for their children may differ from parents who do not with respect to characteristics that are themselves determinants of lower fertility. Lloyd et al. (2000) also found that children's school attendance predicts the timing of fertility transitions in a cross-national sample of African countries. Many studies report a negative impact of women's education and fertility, but again several alternative causal hypotheses are consistent with the data. At present, there is no strong evidence for or against the strict interpretation of fertility transition as a cultural watershed generated by mass education.

6. Summary of Wealth Flows Theory

Consistent with the logic of wealth flows theory, there is a great deal of suggestive evidence that changes in the costs of rearing children and in parental perceptions of those costs are associated with fertility transition. People in many high fertility societies also report higher expectations regarding economic support during old age than do people in low fertility societies. Also consistent with the theory is that education and fertility are negatively related in virtually every study measuring their association.

There are two major deficiencies in the theory, however. The first is that the data do not support a strict interpretation of the wealth flows hypothesis. There is no direct quantitative evidence of net-upward wealth flows in any traditional high-fertility society. Virtually all studies supporting the theory have used national-level, cross-sectional data, and have relied on proxy measures and informant reports. Actual measures of wealth flows in micro-level, longitudinal, ethnographic studies, using behavioral observation and measurement of intrahousehold productivity and exchange, all show *downward* wealth flows in traditional high fertility societies. Caldwell recognized that adequate tests of theory would require longitudinal, microlevel, ethnographically oriented research due to the subtleties and complexity of intra-familial wealth flows (Caldwell 1982).

The second deficiency is that the theoretical foundations for the determination of familial wealth flows are not well specified. This is evident in the treatment of education. In the theory, education is at once a vehicle of values transmission, a response to micro- and macroeconomic forces, and a direct cost to parents. The theory does not specify why certain value systems governing wealth flows arise, nor the relationships between exogenous economic factors, human psychology, and cultural evolution.

7. Conclusion

According to wealth flows theory, cultural transmission of new family values is the principal driving force in fertility transition. However, without a theory which can link ecological conditions to cultural features and their distribution, the concept of culture risks being vague, both cause and effect, the unexplained explanator. Culture is many things: a body of accumulated information, a set of customs and rituals, group beliefs, and a mindset about the organization of the social and physical environments. Each of these components may be related to fertility; but a generalized culture concept cannot be successfully used to explain patterning and variation in fertility without risking causal circularity (Bock 1999).

The wealth flows theory of fertility transition is a major contribution to demography. The theory is general in that it explains both high and low fertility regimes. A second strength is that it addresses both exogenous variables and the psychological and social processes that translated those exogenous variables into individual behavior, emphasizing both economic and social factors in understanding fertility transitions. The theory helped to lay the groundwork for theories of the family to become a major perspective within the field of demography, and broadened the scope of demography by directing attention to culture. It also inspired and influenced a large number of empirical studies. However, in the strict sense of predicting two fertility regimes to be associated with upward and downward wealth flows, empirical evidence does not support the theory. In the weak sense of predicting a negative relationship between net wealth flows to children and fertility, there is a good deal of empirical support.

Bibliography

- Adamchak D J, Ntseane P G 1992 Gender, education, and fertility: A cross-national analysis of sub-Saharan African nations. *Sociological Spectrum* **12**: 167–82
- Axinn W G 1993 The effects of children's schooling on fertility limitation. *Population Studies – Journal of Demography* **47**: 481–93
- Becker G S 1960 An economic analysis of fertility. In: National Bureau of Economic Research (ed.) *Demographic and Economic Change in Developed Countries*. Princeton University Press, Princeton, NJ, pp. 209–31
- Becker G S 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA
- Bock J 1999 Evolutionary approaches to population: Implications for research and policy. *Population and Environment* **21**: 193–222
- Bock J 2001 Learning, life history, and productivity: Children's lives in the Okavango Delta, Botswana. *Human Nature*, in press
- Boserup E 1965 *The Conditions of Agricultural Growth*. Aldine, New York

- Cain M T 1978 The household life cycle and economic mobility in rural Bangladesh. *Population and Development Review* 4: 421–38
- Cain M 1981 Risk and insurance: Perspectives on fertility and agrarian change in India and Bangladesh. *Population and Development Review* 7: 435–74
- Caldwell J C 1976a Fertility and the household economy in Nigeria. *Journal of Comparative Family Studies* 7: 193–253
- Caldwell J C 1976b Toward a restatement of demographic transition theory. *Population and Development Review* 2: 321–66
- Caldwell J C 1980 Mass education as a determinant of the timing of fertility decline. *Population and Development Review* 6: 225–55
- Caldwell J C 1982 *Theory of Fertility Decline*. Academic Press, London
- Chang K-S 1990 Socialist institutions and family wealth flows reversal: An assessment of post-revolutionary Chinese rural fertility. *Journal of Family History* 15: 179–200
- Clark C 1967 *Population Growth and Land Use*. St. Martin's Press, New York
- Cochrane S H 1979 *Fertility and Education: What Do We Really Know?* Johns Hopkins University Press, Baltimore, MD
- Das Gupta M 1994 What motivates fertility decline? A case study from Punjab, India. In: Egero B, Hammarskjöld M (eds.) *Understanding Reproductive Change: Kenya, Tamil Nadu, Punjab, Costa Rica*. Lund University Press, Lund, Sweden, pp. 101–33
- DeLancey V 1990 Socioeconomic consequences of high fertility for the family. In: Acsadi G T F, Acsadi G J, Bulatao R A (eds.) *Population Growth and Reproduction in Sub Saharan Africa: Technical Analyses of Fertility and its Consequences*. The World Bank, Washington, DC, pp. 115–30
- Dharmalingam A 1994 Old age support: Expectations and experiences in a south Indian village. *Population Studies – Journal of Demography* 48: 5–19
- Edgerton D J 1992 Family structure and intergenerational exchanges. *Research on Aging* 14: 427–47
- Hawkes K, Jones N G B 1997 Hadza women's time allocation; offspring provisioning; and the evolution of long postmenopausal life spans. *Current Anthropology* 38: 551–77
- Hugo G 1997 Intergenerational wealth flows and the elderly in Indonesia. In: Jones G W, Douglas R M, Caldwell J C, D'Souza R M (eds.) *The Continuing Demographic Transition*. Oxford University Press, New York, pp. 111–33
- Kaplan H 1994 Evolutionary and wealth flows theories of fertility: Empirical tests and new models. *Population and Development Review* 20: 753–91
- Lee R 1996 A cross-cultural perspective on intergenerational transfers and the economic life cycle. *Seminar on intergenerational economic relations and demographic change: papers*. International Union for the Scientific Study of Population [IUSSP], Committee on Economic Demography, Liege, Belgium, pp. 1–25
- Lee Y J, Parish W L, Willis R J 1994 Sons, daughters, and intergenerational support in Taiwan. *American Journal of Sociology* 99: 1010–41
- Liebenstein H M 1957 *Economic Backwardness and Economic Growth*. Wiley, New York
- Lillard L A, Willis R J 1997 Motives for intergenerational transfers: Evidence from Malaysia. *Demography* 34: 115–34
- Lloyd C B, Kaufman C E, Hewett P 1999 *The Spread of Primary Schooling in Sub-Saharan Africa: Implications for Fertility Change*. Population Council, New York
- Makinwa-Adebusoye P 1994 Changes in the costs and benefits of children to their parents. In: Locoh T, Hertrich V (eds.) *The Onset of Fertility Transition in Sub Saharan Africa*. International Union for the Scientific Study of Population [IUSSP], Liege, Belgium, pp. 175–192
- Mamdani M 1972 *The Myth of Population Control: Family, Caste, and Class in an Indian Village*. Monthly Review Press, New York
- Moore T W 1998 Fertility in China 1982–1990: gender equality as a complement to wealth flows theory. *Population Research and Policy Review* 17: 197–222
- Mueller E 1976 The economic value of children in peasant agriculture. In: Ridker R (ed.) *Population Development: A Search for Selective Interventions*. John Hopkins University Press, Baltimore, MD, pp. 98–153
- Nag M, White B N F, Peet R C 1978 An anthropological approach to the study of the economic value of children in Java and Nepal. *Current Anthropology* 19: 293–306
- Notestein F 1945 Population: The long view. In: Schultz T (ed.) *Food for the World*. University of Chicago Press, Chicago, pp. 36–57
- Nugent J B, Gillaspie R T 1983 Old age pensions and fertility in rural areas of less developed countries: Some evidence from Mexico. *Economic Development and Cultural Change* 31: 809–29
- Schellekens J 1993 Wages, secondary workers, and fertility: A working-class perspective of the fertility transition in England and Wales. *Journal of Family History* 18: 1–17
- Turke P W 1989 Evolution and the Demand for Children. *Population and Development Review* 15: 61–90
- Vlassoff M 1982 Economic utility of children and fertility in rural India. *Population Studies – Journal of Demography* 36: 45–59
- Vlassoff M, Vlassoff C 1980 Old age security and the utility of children in rural India. *Population Studies – Journal of Demography* 34: 487–99
- Willis R J 1982 The direction of intergenerational transfers and demographic transition: The Caldwell hypothesis reexamined. *Population and Development Review* 8: 207–34

H. S. Kaplan and J. Bock

Fertility Theory: Embodied-capital Theory of Life History Evolution

The goal of this article is to provide an introduction to the embodied-capital theory of life history evolution and the empirical research upon which it is based. The embodied-capital theory generalizes existing life-history theory in biology by treating the processes of growth, development, and maintenance as investments in stocks of somatic or embodied-capital (see Kaplan (1996, 1997) and Kaplan et al. (2000) for more complete treatments). In a physical sense, embodied capital is organized somatic tissue—muscles, digestive organs, brains, etc. In a functional sense, embodied capital includes strength, immune function, coordination, skill, knowledge, and other abilities. Since such stocks tend to depreciate in response to metabolic

activity (e.g., the production of free radicals and mutations associated with mitosis) and aging, allocations to maintenance can also be seen as investments in embodied capital. The general theory is introduced and then applied to human hunting and gathering societies, which characterized most of our species' evolutionary history. This framework is then used to develop models of modern fertility behavior as the result of an interaction between historically novel conditions and the evolved psychological, cultural, and physiological mechanisms governing human parental investment and fertility.

1. The Embodied-capital Theory of Life History Evolution

Hominids lived as hunter-gatherers for the vast majority of their evolutionary history (more than 2 million years). Agriculture originated only 10,000 years ago and has been practiced by the majority of the world's population for just two or three millennia, a relatively brief period of time for selection to act. The shift to a fully mercantile, highly urbanized economy with competitive wage-labor markets and a vast array of consumer goods since the mid-nineteenth century is completely novel. Since natural selection is a historical process, humans should be well adapted to the foraging lifestyle, but there is no reason to expect that they will respond adaptively (in the sense of maximizing reproductive fitness) to today's novel circumstances (see Kaplan et al. (1995a, 1995b) for evidence that modern fertility behavior does not maximize fitness). Therefore, an evolutionary explanation of human fertility and parental investment has three requirements: it must be consistent with our general understanding of evolution by natural selection; it must explain the behavior of humans living under traditional hunting and gathering conditions characteristic of our evolutionary history with adaptive models; and it must be able to predict the pattern of modern responses to novel conditions. Our approach to these requirements has been to follow a four-step procedure (Kaplan 1996, Kaplan and Lancaster 2000). The first step is to develop a model of natural selection on fertility and parental investment for organisms in general. The second step is to apply that model to the specific selection pressures acting on human hunter-gatherers. The third step is develop a model of the proximate physiological, psychological and cultural mechanisms produced by selection and how they would interact to determine fertility and parental investment behavior. The fourth step is to generate predictive models of how those mechanisms would respond to conditions currently prevailing in the world. The theory is both ecological and psychological, integrating socioeconomic influences with a theory of mind to identify the processes by which salient features of the social and physical environment

relate to people's reproductive and parenting behavior across diverse social, cultural, economic, and ecological contexts.

2. The Action of Natural Selection on Fertility and Parental Investment

Natural selection acts on variability in the traits of individual organisms within populations. Traits (and the genes which code for those traits) whose *average* effects on the individuals possessing those traits act to maximize their long-term production of descendents through time increase in frequency relative to other traits. (Technically, selection acts on the 'inclusive fitness' of genes coding for traits. Inclusive fitness includes effects both on the reproductive success of the individual bearing the gene and on other individuals, related by common descent, who also bear the gene. For example, selection on genes affecting alarm calls in response to predators depends both on their effects on the reproductive fitness of the caller (who may risk a greater threat of predation) and on relatives bearing those genes (whose lives may be saved by the call).) Fertility is the most direct contributor to an organism's fitness (i.e., the number of descendents it produces). In fact, all other fitness components, such as mortality, only affect fitness through their effects on fertility (for example, mortality rates affect fitness by affecting the probability of living to the next reproductive event). All else constant, any increase in fertility increases an organism's fitness. However, there are two trade-offs affecting natural selection on fertility.

The first is the trade-off between present and future reproduction. By growing, an organism can increase its energy capture rates in the future and thus increase its future fertility. For this reason, organisms typically have a juvenile phase in which fertility is zero until they reach a size at which some allocation to reproduction increases fitness more than growth. Similarly, among organisms that engage in repeated bouts of reproduction (humans included), some energy during the reproductive phase is diverted away from reproduction and allocated to maintenance so that it can live to reproduce again. Natural selection on age of first reproduction and on the adult reproductive rate is expected to maximize total allocations of energy to reproduction over the life course, and it depends on ecological factors, such as mortality rates (Charnov 1993) and the effects of body size on both energy capture (Charnov 1993) and vulnerability to predation (Werner 1986).

The second trade-off is between quantity and quality of offspring, where quality is a function of parental investment in offspring and reflects its ability to survive and reproduce. In general, natural selection on offspring number and investment per offspring is expected to maximize the number of offspring that survive to reproduce themselves during an organism's lifetime

(Smith and Fretwell 1974). Ecological factors, such as food supply, disease and predation rates, also affect the optimal fertility rates and optimal expenditures on offspring.

Virtually all complex organisms exhibit flexibility in both age at first reproduction and fertility rates. Natural selection has resulted in physiological and psychological mechanisms by which individuals adjust fertility onset and fertility rates in relation to changing environmental conditions. For example, plants often exhibit many thousand-fold variations in fertility rates in relation to yearly and seasonal variations in rainfall and other environmental factors. Extensive research on many bird species has shown that this phenotypic plasticity tracks fitness fairly well (see Godfray et al. (1991) for a review). Birds under variable conditions adjust clutch sizes in ways that tend to maximize the number of surviving young produced during the life course. The pervasiveness of adaptive phenotypic plasticity in fertility among nonhumans suggests that humans are also likely to possess adaptations to adjust fertility onset and fertility rates in relation to changing environmental conditions.

3. The Embodied-capital Theory and its Application to Fertility and Parental Investment Among Traditional Humans

The embodied-capital theory treats the present–future reproduction trade-off in terms of optimal investments in own embodied capital vs. reproduction, and the quantity–quality trade-off in terms of investments in the embodied-capital of offspring. Models based on the theory allow us to address problems that have not been addressed with standard life history models. First, an exclusive focus on physical growth *per se* is an impoverished way of understanding human development. The large human brain, a stock of embodied capital, supports a great deal of learning and knowledge acquisition during both the juvenile and adult periods. Second, parental investment among humans affects not only survival to adulthood, but also the adult socioeconomic status of offspring. This is true not only in modern societies, but also among traditional hunter–gatherers. These models of investment in embodied capital have produced some fundamental results. Of central interest here, the models show that investments in embodied capital affecting adult income or energy capture co-evolve with investments affecting mortality and longevity. The longer the time spent growing and learning prior to reproducing, the more natural selection favors investments in staying alive to reap the benefits of those investments. Similarly, any investments that produce increased energy-capture rates later in life select for additional investments to reach those older ages. The converse is also true. Ecological features or investments that increase the probability of survival to older ages also produce

selection for greater investments in income-related embodied capital. These co-evolutionary effects have been particularly important in human life-history evolution.

The human adaptation is broad and flexible, in one sense, and very narrow and specialized, in another. It is broad in the sense that as hunter–gatherers, humans have existed successfully in virtually all of the world's major habitats. This has entailed eating a very wide variety of foods, both plant and animal, both within and among environments. It also has entailed a great deal of flexibility in the contributions of different age and sex classes of individuals. The relative contributions of men and women to food production vary from group to group, as do production profiles by age.

The human adaptation is narrow and specialized in that it is based on extremely high investments in brain tissue and learning. In every environment, human foragers consume the largest, highest-quality, and most difficult-to-acquire foods, using techniques that take years to learn (Kaplan et al. 2000). This adaptation has profound effects on fertility, mortality, and wealth flows. Compared with other primates and mammals, there are at least four distinctive characteristics of human life histories: (a) an exceptionally long lifespan, (b) an extended period of juvenile dependence, resulting in families with multiple dependent children of different ages, (c) support of reproduction by older postreproductive individuals, and (d) male support of reproduction through the provisioning of females and their offspring.

The co-evolution of brains, learning and life histories is evident in Fig. 1. This figure plots net expected cumulative productivity by age (cumulating the probability of being alive at each age times the net productivity at that age) for human hunter–gatherers and wild-living chimpanzees, averaging over all available data from different groups (see Kaplan et al. (2000) for details). The long human training period is evident when the troughs in the human and chimpanzee curves are compared. Humans are not net producers until about age 18–20 years, but are very productive as adults, whereas chimpanzees begin to produce as much as they consume by about age 5 years and are much less productive as adults. In fact, peak production rates for both human males and females do not occur until the mid-thirties for the most difficult-to-acquire resources. Humans also have an expected adult lifespan of about 40 years when they become net producers, whereas chimpanzees have an expected adult lifespan (beginning at about age 13) of 14 years. The dashed line is a hypothetical cross of human production profiles with chimpanzee survival rates. It shows that the human production profile would not be viable with chimpanzee survival rates, because expected lifetime net production would be negative. The human adaptation depends on about 20 years of parental support and a long life span to recoup the investments in learning.

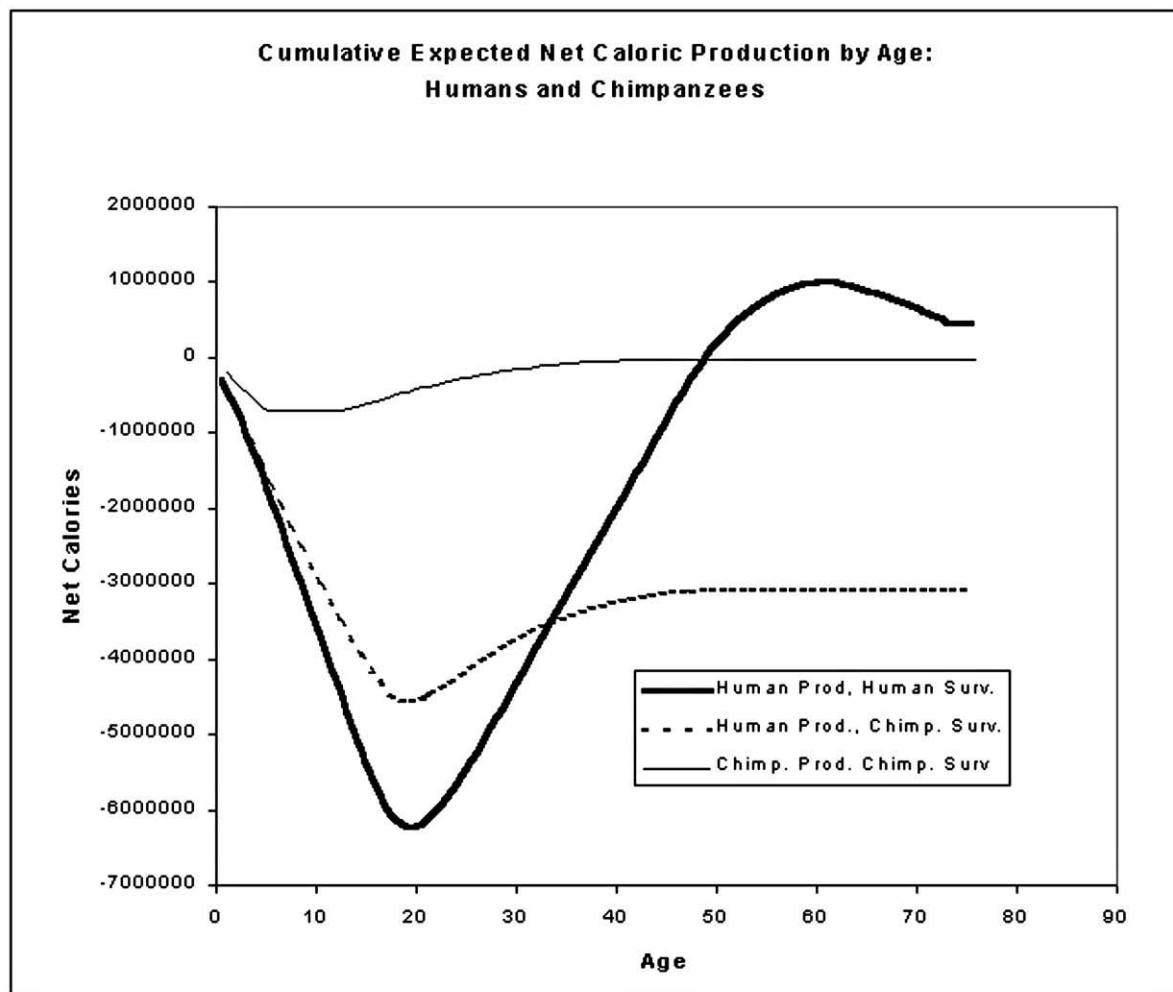


Figure 1
Cumulative expected net caloric production by age: humans and chimpanzees

The first major demographic transition separating our species from apes is a shift to low mortality, an increased lifespan, a long dependent development period, a large commitment to learning and intelligence, and an age profile of production shifted towards older ages with downward wealth flows across generations.

4. Ecology and the Proximate Mechanisms Governing Fertility and Parental Investment

The mortality and production profiles discussed above have direct implications for fertility regulation. Since human parents and grandparents provision children,

it is likely that natural selection would have produced mechanisms by which fertility could respond to the number of children parents could afford to raise at any given time. Figure 2 presents a conceptual model of the evolved mechanisms of fertility regulation in humans and its responsiveness to physical, biological and social conditions (*socioecology*). The figure depicts two classes of socioecological factors, one affecting survival and the other embodied capital, and three domains of evolved responses, psychology and culture, behavior, and reproductive physiology. The plain text and bold arrows refer to the traditional hunting and gathering lifeway, which we consider first. The italicized text and lighter arrows include additional factors and causal pathways existing under modern conditions.

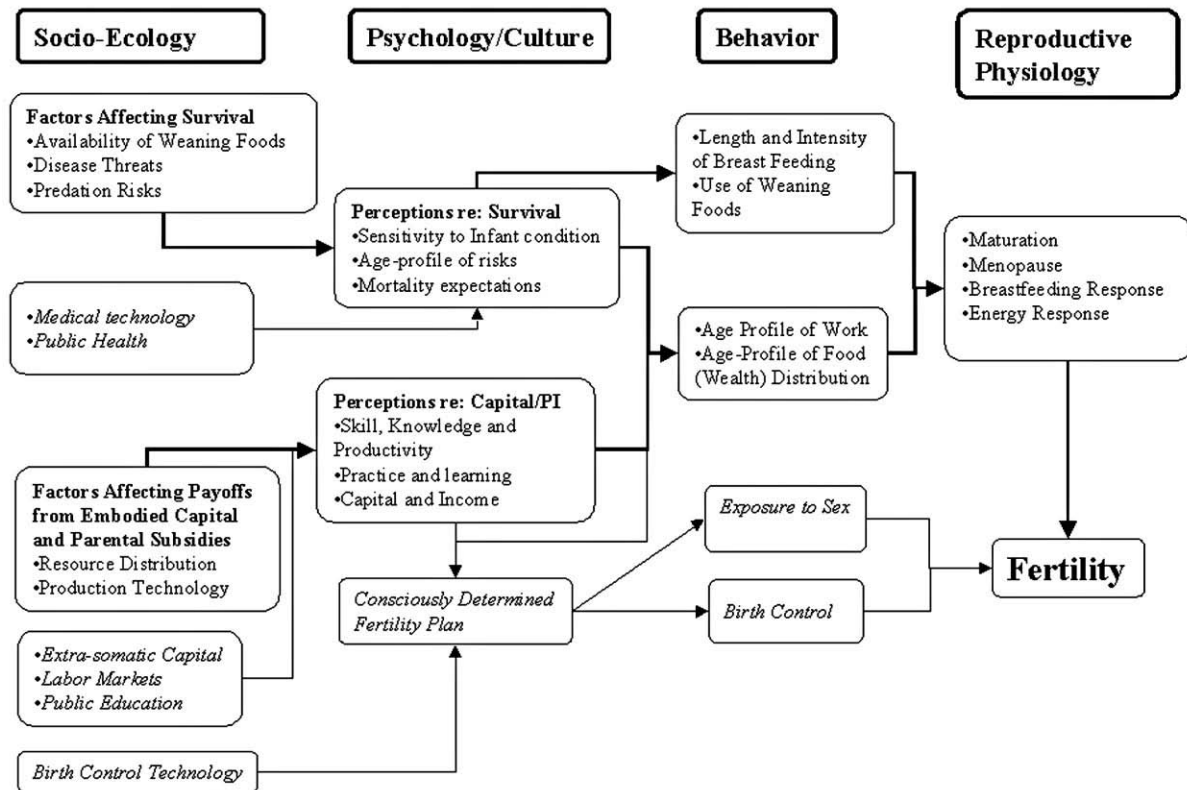


Figure 2
A mode of fertility determinants in traditional and modern societies

4.1 Traditional Foraging Societies

There are two main causal pathways regulating fertility that are both related to the energy constraints in provisioning offspring, one involving breastfeeding and the other involving the age profiles of work and rate of weaning. Natural selection on the timing and quantity effects (favoring earlier weaning) and quality effects on survival and growth of offspring (favoring later weaning). Socioecological conditions, such as the availability of easy-to-digest weaning foods and disease threats, are likely to impact the relationship between age at weaning and survival. Similarly, infant condition is likely to be individually variable and differentially affected by weaning. We expect that natural selection has acted on both infant and maternal psychology to make them responsive to those variable conditions. For example, sickly infants may be more demanding of the breast, and larger and healthier babies more interested in weaning foods. In the human case, maternal perceptions regarding breastfeeding are not only guided by her individual experience and sensitivity to her infant's condition,

but also by the cumulative experience of her group, as reflected in cultural conventions and ideas about proper breastfeeding practices. Maternal and infant psychology interact in determining the length and timing of breastfeeding. The evolved reproductive physiology translates this pattern of breastfeeding into differing degrees of fecundity.

The second causal pathway is due to the additional energy constraints involved in provisioning children. The age/sex profile of work and productivity determine the total energy available for consumption. Given some age profile of food distribution and of mortality, the fertility rate must be constrained so that energy consumed does not exceed energy produced. The longer children are provisioned and the greater proportion of their food needs subsidized by parents and grandparents, the lower the reproductive rate that can be supported with a given adult income. Natural selection should therefore act both on psychological abilities to assess the short- and long-term costs and benefits of different activity profiles and on fertility regulation so that energy flows are balanced. One fundamental trade-off with respect to activity profiles

is between present and future production (Kaplan 1996, Bock 2002). If children engage in easy-to-perform foraging activities, such as fruit collection, this will provide immediate caloric benefits and lower the parental subsidy required. However, time allocation to those activities detracts from time spent learning more difficult activities, such as hunting and extractive foraging. During the learning phase, those activities produce very little food, but generate high rates of production in the future. Another fundamental trade-off is between production and mortality risks, since foraging exposes people to predation, accidents, and getting lost. This suggests that parental psychology should be responsive to age-specific mortality rates and how they are affected by alternative activity profiles, and on their short- and long-term consequences for production/productivity.

There is a growing body of evidence supporting the view that people in foraging societies are sensitive to these trade-offs and adjust behavior accordingly. In a seminal series of papers, Blurton Jones and co-workers (Blurton Jones et al. 1994a, 1994b, Hawkes et al. 1995) show that Kung hunter-gatherer parents consciously assess foraging risks to children and that differences between Kung and Hadza children's productivity are due to features of the local ecology related to ease of resource extraction and dangers associated with productive tasks. Bock (1995, 2002), in a study of forager/agropastoralists in Botswana, shows that the age profiles of children's activity budgets reflect immediate skill and strength requirements as well as opportunities to learn. In all foraging societies for which data are available, people are aware of the many years of learning required to become a competent hunter and young men are provisioned while they learn those skills (Kaplan et al. 2000).

These time-allocation decisions affect fertility through maternal energetics. In addition to breast-feeding effects, fecundity responds to women's net energy flows. Seasonal changes on food intake (Bailey et al. 1992), net energy expenditure (Hurtado and Hill 1987) and weight (Bailey et al. 1992, Tracer 1996) all have been shown to predict fecundity as do individual differences among women (Hill and Hurtado 1996).

To summarize, in this model fertility is determined by the interaction of individual psychology and cumulative cultural knowledge, beliefs and practices, behavior, and reproductive physiology. Each of these domains co-evolved by natural selection to respond to ecological variability in mortality regimes and payoffs to parental subsidization of skill acquisition. The connection between psychology and fertility is indirect in that psychology adjusts parental investment (through productive behavior, wealth flows, and breastfeeding) and reproductive physiology translates parental investment decisions into fertility. The key to this system is that maximizing lifetime expected resource production through the optimal allocation of activities and wealth flows will tend also to maximize

fitness when all wealth is in the form of food and extra food translates into higher fertility.

4.2 Fertility Under Modernizing Conditions

In Fig. 2, the additional socioecological factors and causal pathways proposed to result in fertility transition are shown with italicized type and thin arrows. Our proposal is that this fertility regime results from the interaction of new socioecological conditions with the psychological processes governing parental investment and wealth flows, evolved in response to the hunting and gathering lifeway. The two critical new factors are the existence of extrasomatic capital and changes in the payoffs to embodied capital, especially in the form of education.

Physical capital, such as land and money, is stored and invested in children in the form of inheritance and parental support for education. Unlike the traditional case where all wealth is embodied, there is no guarantee that a wealth-maximizing psychology will maximize fitness, because much wealth may never be converted into offspring and remain stored in other physical forms.

Changes in the payoffs to embodied capital occur for two reasons. Changes in the technology of production to education-based labor markets have led to very high returns on parental investments in children's education. There are multiple inputs into the production of education-based embodied capital. Parental time inputs prepare children for school and affect their progress in school. Variation in the quality of those inputs, perhaps based on parental education, may predict both levels of parental investment and fertility. Teachers, educational materials, and the quality of student preparation all affect rates of learning (Card and Krueger 1992).

Changing medical technology and public health efforts greatly reduced mortality rates for all age groups. Increased survival rates during the period of parental investment increase the expected costs per child born, favoring further increases in offspring quality (Kaplan 1996). Increased survival during the adult period increases the expected years of return of educational investments, further increasing the incentive to invest in children's education. It will be especially productive to investigate the public and private dynamics involved in this process. Increased private interest in education may have inspired more public investment in education. The quality of public education, in turn, increased the rate of return on private investments in education. Similarly, increased private interest in health (perhaps inspired, in part, by increased investment in education) also stimulated increased public investment in health. This, in turn, increased payoffs to educational investments.

These factors, in combination with parental psychology, result in fertility being regulated by a consciously determined childbearing plan realized

through birth-control technology and/or controlled exposure to sex. In response to the increased payoffs to investments in education and the expected costs of those investments, parents determine the number of children they can afford to raise, given their wealth. The low mortality rates also allow parents to plan reproduction at the outset, because the number of children born accurately predicts the number of children that will be raised to adulthood. Reproductive physiology affects fertility in the present in that it sets the broad limits within which a childbearing plan can be realized, and may actually lower fertility when plans include significant delays to first reproduction (Kaplan et al. 2001).

A great deal of further research is necessary, however, before we can understand why these changes in socioecology have resulted in such low levels of fertility and high levels of parental investment and wealth consumption. It does not appear that people are simply maximizing family wealth because net wealth of families would be maximized by higher fertility than is currently observed. It is also clear that people are not maximizing personal consumption; in that case, they would have no children. One possible hypothesis is that the social dynamics of small groups in hunting and gathering economies resulted in greater fitness for those of higher social standing and selected for a psychology in which relative social position of self and offspring is valued highly. Such a psychology would also emerge if people judged the well-being of themselves and their children based upon the wealth and consumption of others. This psychology may have been fitness maximizing under traditional conditions. If relative, as opposed to absolute, wealth and social standing guide human decisions regarding wealth flows, parental investment, and fertility, it is possible that 'run-away' consumption and investment in children's education result from the interaction of this psychology and modern education-based labor markets and consumption possibilities.

5. Implications of Embodied-capital Theory for Understanding Current Fertility Transitions

Embodied-capital theory links several existing theories of demographic transition and, in so doing, provides a new perspective on each one. It rationalizes the shift from natural fertility to parity-specific fertility in terms of a changing quality/quantity trade-off, as a transition from high fertility/low parental investment to low fertility/high parental investment. It also shows why the shift to lower mortality rates can lead to lower fertility, but for reasons that are very different from what traditional demographic-transition theory proposes. According to the latter, lower mortality precedes and drives lower fertility, because fewer children are required to maintain the population equilibrium and cultural ideologies evolve to maintain population

equilibrium. According to the present view, mortality rates are but one important factor in the overall rate of return on investments in embodied capital.

This model can further serve to organize research on fertility transitions, with the fundamental expectation that exogenous factors affecting the rate of return on parental investment in the embodied capital of offspring will predict the timing of fertility decline and its distribution within populations. Understanding these dynamics may provide insights into the course of fertility reduction and its variation across countries and among individuals within countries. For example, most studies of predemographic transition societies show that there is a positive association between wealth and fertility (Borgerhoff Mulder 1992). There is some evidence for a negative relationship during the demographic transition. (Kaplan et al. 1995a, Vining 1986) and no relationship after the transition is complete (Kaplan et al. 1995a). It may be that only the wealthy can take advantage of high levels of investment in children's embodied capital when there are few schools and limited public expenditures on health and education. The process of demographic transition may progress through different sectors of the society as more and more people can take advantage of increased rates of return due to changes in the economy and changes in public investments. These same dynamics, in combination with international transfers of educational and health technology, may also explain the national differences in the timing of fertility reduction.

Understanding these dynamics also requires an understanding of how people process the mix of information derived from direct assessment, observation of others, and cultural guidelines. Theoretical work modeling the action of natural selection on information processing suggests that the relative reliance on direct experience and on cultural sources should vary according to environmental conditions (Bock 1999). Cultural learning should be a better guide for behavior when environments are relatively stable than when they are changing (Bock 2001). The rapid change in economics of production, public education and public health may lead people to reject existing cultural models, look for new models, and assess costs and benefits of alternative courses of action more consciously on the basis of direct experience.

International family planning programs present one set of new models, promoting lower fertility and greater investment in children's education. When those models are consistent with individual assessments of costs and benefits, the direction of change is clear. It is less clear, however, how people will respond to family-planning messages in the short and long run if the payoffs to investment in embodied capital are low, either because schools are poor or because employment opportunities for educated people are limited (as appears to be the case in many sub-Saharan African countries).

6. Conclusion

According to embodied-capital theory, fertility and parental investment behavior result from the interaction of exogenous conditions with an evolved response system shaped by natural selection. In addition, a distinctive feature of the learning-intensive nature of human psychology is that information about the environment and guidelines for behavior are obtained through cultural sources as well as through direct personal assessment (Bock 1999, 2001). Given the very limited capacities for cultural transmission displayed by nonhuman primates (Lefebvre 1995), it is evident that the capacity for culture itself evolved by natural selection. We still know very little about the mix of physiological, psychological, and cultural processes underlying decisions regarding wealth flows, educational investments and consumption patterns, or how those processes relate to fertility decisions. Understanding the complex interplay of exogenous conditions, cultural forces, and evolved response systems in the determination of family structure and fertility presents a formidable challenge, but one that promises great rewards.

See also: Fertility: Institutional and Political Approaches; Fertility: Political and Political-Economic Perspectives; Life Course: Sociological Aspects

Bibliography

- Bailey R C, Jenike M R, Ellison P T, Bentley G R, Harrigan A M, Peacock N R 1992 The ecology of birth seasonality among agriculturalists in Central Africa. *Journal of Biosocial Science* **24**: 393–412
- Blurton Jones N G, Hawkes K, Draper P 1994a Differences between Hadza and Kung children's work: Original affluence or practical reason. In: Burch E S (ed.) *Key Issues in Hunter-Gatherer Research*. Berg, Oxford, UK, pp. 189–215
- Blurton Jones N G, Hawkes K, Draper P 1994b Foraging returns of Kung adults and children: Why didn't Kung children forage? *Journal of Anthropological Research* **50**: 217–48
- Bock J A 1995 The determinants of variation in children's activities in a Southern African community. PhD thesis, University of New Mexico
- Bock J 1999 Evolutionary approaches to population: Implications for research and policy. *Population and Environment* **21**: 193–222
- Bock J 2001 Evolutionary theory and the search for a unified theory of fertility. *American Journal of Human Biology* (in press)
- Bock J 2002 Learning, life history and productivity: children's lives in the Okavango Delta, Botswana. *Human Nature* (in press)
- Borgerhoff Mulder M 1992 Reproductive decisions. In: Smith E A, Winterhalder B (eds.) *Evolutionary Ecology and Human Behavior*. Aldine, Hawthorne, NY, pp. 339–74
- Card D, Krueger A B 1992 Does school quality matter: Returns to education and the characteristics of public schools in the United States. *Journal of Political Economy* **100**: 1–40

- Charnov E L 1993 *Life History Invariants*. Oxford University Press, Oxford, UK
- Godfray H C J, Partridge L, Harvey P H 1991 Clutch size. *Annual Review of Ecology and Systematics* **22**: 409–29
- Hawkes K, O'Connell J F, Jones N G B 1995 Hadza children's foraging: Juvenile dependency, social arrangements, and mobility among hunter-gatherers. *Current Anthropology* **36**: 688–700
- Hill K, Hurtado A M 1996 *Ache Life History: The Ecology and Demography of a Foraging People*. Aldine, Hawthorne, NY
- Hurtado A M, Hill K R 1987 Early dry season subsistence ecology of the Cuiva foragers of Venezuela. *Human Ecology* **15**: 163–87
- Kaplan H S 1996 A theory of fertility and parental investment in traditional and modern human societies. *Yearbook of Physical Anthropology* **39**: 91–135
- Kaplan H S 1997 The evolution of the human life course. In: Wachter K, Finch C (eds.) *Between Zeus and Salmon: The Biodemography of Aging*. National Academy of Sciences, Washington, DC, pp. 175–211
- Kaplan H S, Lancaster J B, Anderson K G, Tucker W T 2001 An evolutionary approach to below replacement fertility. *American Journal of Human Biology* (in press)
- Kaplan H S, Lancaster J B 2000 Skills-based competitive labor markets, the demographic transition, and the interaction of fertility and parental human capital in the determination of child outcomes. In: Cronk L, Irons W, Chagnon N (eds.) *Human Behavior and Adaptation: An Anthropological Perspective*. Aldine, New York, pp. 283–322
- Kaplan H S, Lancaster J B, Bock J A, Johnson S E 1995a Does observed fertility maximize fitness among New Mexican men? A test of an optimality model and a new theory of parental investment in the embodied capital of offspring. *Human Nature* **6**: 325–60
- Kaplan H S, Lancaster J B, Bock J A, Johnson S E 1995b Fertility and fitness among Albuquerque men: A competitive labor market theory. In: Dunbar R I M (eds.) *Human Reproductive Decisions: Biological and Social Perspectives*. St. Martin's Press, New York, pp. 96–136
- Kaplan H S, Lancaster J B, Hill K R, Hurtado A M 2000 A theory of human life history evolution: Diet, intelligence, and longevity. *Evolutionary Anthropology* **9**: 156–85
- Lefebvre L 1995 Culturally-transmitted feeding-behavior in primates: Evidence for accelerating learning rates. *Primates* **36**: 227–39
- Smith C C, Fretwell S D 1974 The optimal balance between size and number of offspring. *American Naturalist* **108**: 499–506
- Tracer D P 1996 Lactation, nutrition, and postpartum amenorrhea in Lowland Papua, New Guinea. *Human Biology* **68**: 277–92
- Vining D R 1986 Social versus reproductive success: The central theoretical problem of human sociobiology. *Behavioral and Brain Sciences* **9**: 167–216
- Werner E E 1986 Amphibian metamorphosis: Growth-rate; predation risk; and the optimal size at transformation. *American Naturalist* **128**: 319–41

H. S. Kaplan and J. Bock

Fertility Transition: China

Between 1970 and 1980, in one decade, China's national fertility level was cut in half. This is his-

torically the most drastic fertility decline among all human populations of significant size, aside from wars, epidemics, and famines. As the largest population in the world, China's fertility transition has also had the greatest impact on the slowing down of the world's population growth. China's fertility decline alone reduced the world's population growth by at least 250 million individuals during the second half of the twentieth century. This entry reviews the process, its historical background, the roles of the government programs and other factors, and the consequences of this transition.

1. The Process of Fertility Transition

At the national level, China's fertility transition began in the early 1970s. In 1970, the total fertility rate was 5.8, close to the highest national level ever recorded for China. In the following decade, a most drastic reduction in fertility occurred. China's total fertility rate declined to 3.6 by 1975, and to 2.3 by 1980. The fertility level fluctuated between 2 and 3 throughout the 1980s. By the early 1990s, it reached and thereafter stayed at the replacement level of 2.1. China, in other words, completed its fertility transition in two decades, from 1970 to 1990.

In China's urban areas and among the more educated subgroups of the population, the fertility transition began much earlier, dating back to the early 1950s (Lavelly and Freedman 1990). Since 1957, urban Chinese fertility declined almost consistently, with only one exception, when fertility rebounded in 1963 after a plunge caused by the Great Leap Forward famine. By 1966, total urban fertility dropped to 3.1, and since 1974, it has never been higher than 2. In Shanghai, China's most cosmopolitan city, the total fertility rate dropped to below the replacement level (of 2.1 births per woman) for the first time in 1964, and to below 1 in 1975. By 1970, on the eve of China's first national family planning program, use of contraception and abortion among young women had already risen to 35 and 20 percent, respectively, in cities. Even in the countryside, over 15 percent of women at age 35 had already used a modern contraceptive method, and 7 percent had had at least one induced abortion.

In China's rural areas, where close to 80 percent of the population resided in the 1970s, fertility transition did not begin until much later—around 1970. In fact, during most years between 1950 and 1970, rural fertility showed signs of increase under the collective commune system. Fertility generally stayed at a level of above 6, with the exception of the Great Leap Famine years of 1959 to 1961, when rural fertility was nearly halved. Consistent fertility decline in rural China began after 1970. The total fertility rate dropped to 5 by 1973, to 4 by 1975, and to 3 by 1979. After 1980, the total fertility rate has fluctuated mostly between 2 and 3.

There were also regional variations in the fertility transition in rural China. In provinces with a more advanced economy and a stronger tradition of population control, such as the Lower Yangtze provinces of Jiangsu and Zhejiang, rural fertility declined earlier. In 1973, when the national level was 5, the total fertility rate was only 2.8 and 3.5 in these two provinces. By comparison, rural fertility was much above the national average in economically less developed provinces: it was 7.4 in Guizhou and 6.5 in Gansu (Coale and Chen 1987).

2. Historical Background

China's large population size along with a strong cultural preference for sons and a practice of early and universal marriage among Chinese women have often led many to believe that the Chinese did not attempt to control fertility before China's contemporary fertility transition. Malthus, for instance, stated that 'To account for this population, it will not be necessary to recur to the supposition of Montesquieu, that the climate of China is in any peculiar manner favourable to the production of children, and that the women are more prolific than in any other part of the world.' Among the causes he identified is 'The extraordinary encouragements that have been given to marriage, which have caused the immense produce of the country to be divided into very small shares, and have, consequently, rendered China more populous, in proportion to its means of subsistence, than perhaps any other country in the world' (Malthus 1971, pp. 101, 103).

Contrary to such earlier observations, many Chinese controlled their demographic outcomes in the past by relying on multiple mechanisms. The best-documented and most widely cited examples pertain to infanticide, especially female infanticide (Lee 1981, Lee et al. 1994, Lee and Campbell 1997). Chinese families killed some of their newborns in order to achieve their desired goals for the number and sex composition of their offspring. Poor harvests and economic difficulties also exacerbated the practice of infanticide.

Moreover, studies of Chinese historical population data also reveal that Chinese fertility within marriages was also significantly lower than their pretransition European counterparts (Barclay et al. 1976, Lee and Wang 1999). This lower marital fertility level was achieved by a much longer interval between marriage and first birth, longer interbirth intervals, and an earlier ending of childbearing. The same rationales that motivated infanticide also operated for fertility control. Chinese parents sped up or slowed down reproduction according to the number and sex composition of their existing children, and according to their economic familial circumstances (Wang et al. 1995, Lee and Campbell 1997, Zhao 1997). Fertility control in the past relied mainly on sexual abstinence,

prolonged breast-feeding, and, to a lesser extent, abortion and traditional contraceptives (Hsiung 1994, Lee and Wang 1999).

These behaviors suggest that controlling births is not a novel idea to the Chinese population. The modern fertility transition did not require new ideas for regulating population behavior, but rather a new environment in which fertility control could be seen as necessary and feasible. In contrast to traditional birth control that relied on both prenatal and postnatal methods, the modern fertility transition is mostly prenatal, facilitated by new contraceptive and abortive technologies. Historical traditions of demographic control not only lay the foundation for contemporary fertility transition in mainland China, but they also help to explain the similarities in the pace and patterns of fertility transitions observed in several populations. In spite of having different political and economic systems, mainland China, Taiwan, Hong Kong, and Singapore all exhibited highly similar patterns of fertility transition, partly due to a shared cultural tradition (Coale and Freedman 1993). The Chinese fertility transition process is therefore fundamentally a continuous process, extending a long tradition of demographic control.

3. Government Programs and Other Factors

The most prominent feature of China's fertility transition is the role played by government family planning programs. The most drastic fertility decline coincided with the launching of the most forceful government policies and programs of population control. Development of government policy can be traced to the early 1950s. Early concerns about rapid population growth and the need for a government family planning program emerged in the early 1950s, when China's 1953 census confirmed its large population size. Abortion was legalized and contraceptives began to be provided on a limited scale. Unlike later programs, which were motivated by government birth control goals, these early provisions were largely to meet an increasing demand among the population for birth control. These early efforts, however, soon became entangled in ideological debates about whether China needed to control its population growth and were interrupted by a series of political campaigns. As a result, it was not until the 1960s that the government seriously promoted birth control in urban and densely populated rural areas. Overall, government family planning programs at this early stage may have facilitated the early decline, but they were not the main driving force.

Government intervention in reproduction reached a new phase in the early 1970s when the policy of *wan* (later marriage), *xi* (longer birth intervals), and *shao* (fewer births) was formed. Although largely accepted by most people on a voluntary basis, the implemen-

tation of the policy became increasingly intensified and regularized. The government utilized its control over the media to promote family planning as a national policy, and mobilized a whole system of vertically controlled work and political organizations to implement the policy. Special government organizations in charge of family planning were also organized and expanded during this period. The official marriage age was stipulated to be 25 and 23 for urban males and females, and 22 and 20 for rural youths. Two children were considered the proper family size for an urban couple and three for most rural couples. Under this program, Chinese ages at marriage rose sharply and fertility declined swiftly. Nationally, the female mean age at first marriage rose from 21 in 1970 to 23 in 1979, and the fertility level was halved (Banister 1987).

In spite of this rapid fertility decline of the 1970s, the Chinese government was not content. Instead, motivated by a keen desire to increase quickly the per capita living standard within a short period of time, it launched an even more ambitious and much more controversial family planning program in 1979. Known as the one-child policy, it required each couple to have only one child, and was an emergency measure to further reduce China's population growth rate. Under the pressure of such an unprecedented and extreme policy, officials and especially family planning cadres resorted to all methods, including physical abuse in some locales, to meet the goals established by various levels of government. Women with two children, sometimes even only one, were required to have an intrauterine contraceptive device (IUD) inserted, and women with three or more births were required to undergo sterilization. Pregnancies not meeting the requirement of family planning had to be aborted. Family planning campaigns were also launched in the 1980s, resulting in sharp increases in the numbers of abortions and sterilization (Hardee-Cleaveland and Banister 1988).

This extreme policy met strong resistance from the Chinese population from the very beginning, especially from the vast population in the countryside. Whereas the earlier policies of family planning were by and large acceptable to Chinese peasants, this new extreme measure required them not only to forgo their traditional perceptions of family and kinship, but also to reduce or even remove their reliance on children for labor and support in old age. Such resistance led to a relaxation in implementation in the Chinese countryside and to changes in the policy. As a result, Chinese couples in cities (accounting for about 20 percent of the population in the late 1970s) obeyed the one-child policy, whereas most rural couples had two, and some more than two, children (Wang 1996, Feeney and Yuan 1994).

The draconian government policy may have masked the roles of many other factors underlying China's fertility transition process. In addition to the historical tradition of demographic regulation, many contem-

porary social and economic changes contributed to China's fertility transition, especially before the 1980s. These changes included a drastic decline in mortality, especially infant mortality, a rise in nonagricultural and nonfamilial employment, a rapid reduction in female illiteracy, and increased provision of social welfare, especially to the urban population (Peng 1989). A changing social and economic context thus produced rapid adaptation to changes in childbearing in the Chinese mainland, just as it did elsewhere in East Asia (Greenhalgh 1988). At both the regional and the individual level, fertility differentials associated with employment type, residence, education, and income existed from as early as the mid-1950s (Lavelly and Freedman 1990).

4. Consequences

In its speed and intensity, the Chinese fertility transition has produced several distinctive consequences. The first is a whole generation of single children; and a whole generation of parents with only one child. These families are almost all in urban China, where the one-child policy was effectively implemented. Children who grew up in these families have no siblings, and in the future their children will not have any uncles, aunts, or cousins, all of whom are important members of the traditional Chinese kinship network. Moreover, there will be a severe shortage in familial support in their old age. At least half of all parents of these single children will not be able to co-reside with their married children in their future old age, when emotional and functional support from children may be needed.

Second, rapid fertility decline also means an accelerated process of population aging. This is again especially true for China's urban population, in which the ratio of the elderly aged 60 and above to those at the working age of between 20 and 60 will reach an unprecedentedly high level, the highest in the world, around the year 2040. Nationally, the rate of population aging will also be high, but not as high as in urban areas due to a higher fertility rate in the vast rural areas. Rural to urban labor migration, therefore, can alleviate the shortage of labor in urban areas.

Lastly, China's rapid fertility transition has also led to an acutely unbalanced sex ratio between male and female births. During the 1980s, the sex ratio at birth rose steadily, from 107 boys to 100 girls in 1981 to 115 boys to 100 girls in 1991. The sex ratio among births of higher orders rose even more. Among second births, the recorded sex ratio rose from 107 in 1981 to 116 in 1985, and to 130 in 1993 (Gu and Roy 1995). Much of this increased sex ratio is due to the widespread use of sex-selective abortion. Whereas the increase in the sex ratio may be attributable in part to its government program, the fact that similar increases in sex ratio are also observed in several other East Asian populations suggests this may at least be a temporary by-product of the fertility transition process in these societies. In

China, such a lopsided sex ratio will have long-lasting effect on marriage patterns and on population growth in the long run.

See also: East Asian Studies: Gender; Family Size Preferences; Family Systems and the Preferred Sex of Children; Fertility Control: Overview; Fertility Transition: Cultural Explanations; Fertility Transition: East Asia; Fertility Transition: Economic Explanations; Fertility Transition: South Asia; Fertility Transition: Southeast Asia; Infant and Child Mortality in the Less Developed World; Land Rights and Gender

Bibliography

- Banister J 1987 *China's Changing Population*. Stanford University Press, Stanford, CA
- Barclay G W, Coale A J, Stoto M A, Trussell J 1976 A reassessment of the demography of traditional rural China. *Population Index* 42: 606–35
- Coale A J, Chen S 1987 *Basic Data on Fertility in the Provinces of China, 1942–1982*. East-West Population Institute, Honolulu, HI
- Coale A J, Freedman R 1993 Similarities in the fertility transition in China and three other East Asian populations. In: Leete R, Alam I (eds.) *The Revolution in Asian Fertility: Dimensions, Causes, and Implications*. Clarendon Press, Oxford, UK, pp. 208–38
- Feeney G, Yuan J 1994 Below replacement fertility in China? A close look at recent evidence. *Population Studies* 48: 381–94
- Greenhalgh S 1988 Fertility as mobility: Sinic transitions. *Population and Development Review* 14: 629–74
- Gu B, Roy K 1995 Sex ratio at birth in China, with reference to other Areas in East Asia: What we know. *Asia-Pacific Population Journal* 10: 17–42
- Hardee-Cleaveland K, Banister J 1988 Fertility policy and implementation in China, 1986–88. *Population and Development Review* 14: 245–86
- Hsiung P 1994 More or less: Cultural and medical factors behind marital fertility in late imperial China. Paper presented at the International Seminar on Abortion, Infanticide and Reproductive Culture in Asia, Kyoto, Japan
- Lavelly W, Freedman R 1990 The origins of Chinese fertility decline. *Demography* 27: 89–116
- Lee B 1981 Infanticide in China. In: Guisso R, Johannson S (eds.) *Women in China*. Philo Press, Youngstown, NJ, pp. 163–77
- Lee J, Campbell C 1997 *Fate and Fortune in Rural China: Social Organization and Population Behavior in Liaoning, 1774–1873*. Cambridge University Press, Cambridge, UK
- Lee J, Wang F 1999 *One Quarter of Humanity, Malthusian Mythology and Chinese Realities 1700–2000*. Harvard University Press, Cambridge, MA
- Malthus T R 1971 *An Essay on the Principle of Population, Seventh Edition*. Augustus M. Kelley, New York
- Peng X 1989 Major determinants of China's fertility transition. *The China Quarterly* March: 117: 1–37
- Wang F 1996 A decade of the one-child policy: Achievement and implications. In: Goldstein A, Wang F (eds.) *China: The Many Facets of Demographic Change*. Westview, Boulder, CO, pp. 97–120

- Wang F, Lee J, Campbell C 1995 Marital fertility control among the late imperial Chinese nobility: Implications for two types of preventive check. *Population Studies* **49**: 383–400
- Zhao Z 1997 Deliberate birth control under a high-fertility regime: Reproductive behavior in China before 1970. *Population and Development Review* **23**: 729–68

F. Wang

Fertility Transition: Cultural Explanations

1. Introduction

Demographers have used two approaches to culture in their attempts to understand fertility transitions, both borrowed from anthropology and, to a lesser extent, sociology. The first sees culture as a set of largely inflexible shared beliefs that constrain action, and thus as a barrier to fertility transitions. The second sees culture as shared interpretations that are subject to negotiation by actors; whether these hinder or facilitate fertility transitions is contingent. Demographers have made considerably more use of the former than the latter, although the more flexible and contingent view of culture is evident in some recent work. This article describes these two approaches to culture, giving examples of their uses in the demographic literature and their links to anthropology and sociology. Although the following begins with the introduction of culture in analyses of fertility transitions in Europe in the nineteenth century, the focus is on fertility transitions in third world countries. Much of the literature cited concerns Africa, but the approaches to culture in understanding fertility transitions are similar in other parts of the world.

2. The Princeton European Fertility Project

The interpretations of culture by scholars participating in the Princeton European Fertility Project set a persistent pattern. An influential paper summarized the project's findings with respect to culture: 'cultural setting influenced the onset and spread of fertility decline independently of socioeconomic conditions. Proximate areas with similar socioeconomic conditions entered the transition at different times, whereas areas differing in the level of socioeconomic development but with similar cultures entered the transition at similar times' (Knodel and van de Walle 1986, p. 412; see also Coale 1973). By cultural setting, the project meant region, taken to be a proxy for ethnicity and language (which were considered synonymous); other potential markers of regional cultures, such as gender, were not explored (Watkins 1993). Regional cultures were assumed to be exogenous and inflexible. Culture was also assumed to hinder rational action. Thus, the Princeton Project's model of

culture was one that has its roots deep in the European Enlightenment: individuals are assumed to be fundamentally rational, but their vision could be clouded by superstitions and myths.

3. Culture as Barrier

As demographers began studies in the third world, interest in culture increased. This interest had several sources. First, researchers saw third world couples as profoundly different from Western couples—even more different than, say, the differences among regions in Europe. They emphasized aspects of culture that appeared quite exotic from a Western perspective—polygamy, strong son preference, long post-partum abstinence—and were seen to be associated with ethnicity (sometimes a specific ethnicity, sometimes a broader category of 'African' or 'Asian' culture). Second, those promoting family planning programs intended to alter the reproductive patterns of third world couples interpreted their initial failures as due to the 'cultural beliefs' of those who should want fewer children and who should use family planning but who were resistant to reproductive change.

The literature on third world fertility declines is replete with views of culture as barriers to change. These views that are expressed by analysts from third world countries as well as from the West, suggest that they inhabit the same cultural space. The solutions both propose are also similar: education to dispel ignorance and 'myths and rumors,' and cultural change. The latter is either presented as the inevitable accompaniment of modernization, or as something that can be promoted through appropriate policies.

The definition of culture implicit in much of the literature on cultural barriers draws on the view of culture that dominated cultural and social anthropology of the 1920s to the 1940s (Hammel 1990). Because of their far-flung empire, British social anthropologists were particularly influential. For the British social anthropologists who worked in Africa, this was structural functionalism, which emphasized the integrated social wholeness, indigenous morality, and the logical coherence of societies; particular attention was paid to kinship as a central aspect of culture (S. F. Moore 1994, p. 23). Culture was viewed as external to the individual, a set of beliefs shared by a local group, with little attention to differences within a culture or to change. Moreover, it was 'traditional' societies that had culture, whereas 'modern' societies did not.

Despite the views of anthropologist that there was a logical coherence to culture or that reproductive behavior might be the outcome of motivations that were far from reproduction (Kreager 1985), demographers focused on customs or practices thought to be closely related to reproduction, particularly those believed to support high fertility such as polygamy and social control by clan or lineage elders, but also local

practices that reduced fertility such as extended postpartum abstinence (Lesthaeghe 1989, Caldwell and Caldwell 1987). Again, other aspects of local cultures were ignored. The analytic models of demographers sometimes included ethnicity or region, but no other indicators of local culture, and much of the attention was focused on aspects of modernization that were expected by sociologists and economists to diminish the hold of culture.

4. Theoretical Shifts in Anthropology and Sociology

Between the 1940s and the 1960s, the interpretation of culture in anthropology—and in sociology—changed substantially. Anthropologists working in third world countries began to focus on the break up of local cultures associated with modernization, especially migration to urban areas (S. F. Moore 1994), and even ethnicity—the stalwart of demographers' views of culture—has come to be seen as fluid and socially constructed (Ekeh 1990). Beginning roughly in the 1970s, there was an even more significant reorientation of theory towards social actors and their strategies for everyday life. A particularly influential figure on this reorientation is the interpretative school associated with Clifford Geertz, who turned attention to meaning in terms of the actors rather than the models of analysts. Geertz proposed attempting to understand the perceived world of actors and their motivations: 'the trick is to figure out what the devil they think they are up to' (Geertz 1983).

Geertz's insights were extended and modified, particularly in the direction of recognizing more within-culture diversity and strategic agency. For example, although the cultural anthropologists of an earlier era, and demographers, saw a close correspondence between beliefs and behavior, in the new interpretations of culture, beliefs and norms were diverse and often contradictory, and which ones were chosen to guide (or perhaps rationalize) behavior was negotiated (Holy and Stuchlik 1983). Pierre Bourdieu (1977) and others emphasized the extent to which individuals know the rules of the game, and use them strategically.

In this approach, culture remains supraindividual, as it was for the structural-functionalists, but it is clearly endogenous, altered, and adapted to specific circumstances. Although local actors may invoke 'our culture since time immemorial' and they themselves may speak of the customs that appear to outside analysts to be exotic, culture is flexible, and concerns rules for daily living. Anthropologists and sociologists began to speak of culture as a 'spice rack' or a 'tool kit' from which various items may be drawn as needed (Greenhalgh 1988, Swidler 1986). Thus, it is argued, analysts need to pay attention to the process by which norms and behavior are evaluated, i.e., what people can expect regarding appropriate behavior, how they

assess motivation, behavior, and the social worth of others (H. Moore 1986, p. 40).

In this interpretive view, then, culture changes through communication. An important part of this communication takes place in 'microcultures,' where 'shared meanings [are] directly tied to specific, likewise shared, experiences of people, settings and events' (Hannerz 1992, p. 77, Watkins 2000). Some of it occurs through gossip, a mechanism by which people not only interpret the behavior of others but also discover other people's interpretations; they can thus learn and change cultural rules at a distance. Culture may provide ideal rules to apply to particular circumstances, but the contingencies of life themselves restructure the rules, even change them over time. Thus, in gossip people may justify different conclusions in the same way: for example, in a community in Malawi, one respondent, when asked how many children she wanted, replied 'Only a few, I am too poor to have more,' whereas another said 'Many, I am poor so I need many children to help me.' Much of this communication is local, but it may also be global, following the 'cultural flows' that appear to be increasingly linking societies and are collectively interpreted not only in villages but in global agencies and international meetings and among functionaries of national agencies such as the Ministry of Health (Appadurai 1990, Meyer et al. 1997, Bongaarts and Watkins 1996, Renne 1996).

5. What Role Do Interpretive Views of Culture Play in Demography?

The interpretive views of culture have been a basis of vigorous critiques of demographers' interpretations of culture (Greenhalgh 1995, Kerzer 1995, Kertzer and Fricke 1997). Although many demographers have continued to draw on 'fossilized' views of culture drawn from an older era of anthropology (Hammel 1990), the newer views of culture have made some inroads, primarily through anthropologists with an interest in reproduction (many of the younger ones of whom received training in demography through a program established by the Mellon Foundation). Three articles illustrate the sorts of insights that derive from analyses informed by the newer views of culture.

Alaka Basu and Sajeda Amin (2000) emphasize the role of culture in the region of Greater Bengal, but interpret region very differently from the Princeton Project model. Their region is characterized by a common language and a high degree of identity with that language. But the region crosses the border between Bangladesh and West Bengal in India, such that the Bengali language is shared by groups that do not share political or religious boundaries. Nonetheless, they say, cross-border interaction has long been sustained and intense among the elites, as well as between urban and rural areas. More significantly—

and in stark contrast to the common view among demographers that the culture shared through the common language is a barrier to fertility change—they argue that in this specific context culture has been characterized for several centuries by its openness to innovation, evident in receptivity to Western contraception as well as to social reform in other areas of social life.

Tom Fricke's (1997) analysis of fertility control among the Tamang of Nepal shows that the same cultural value can lead to precisely opposite reproductive action when the frame shifts. Between his fieldwork in 1981–2 in the village of Timling and a subsequent visit in 1987, he found that some men had undergone a vasectomy. These men were not characterized by more education than others, nor by more contact with the world outside their village. In listening to their explanations of why they had made such a move, he found that they referred constantly to an underlying cultural value: that of reciprocity and exchange, which 'provides Timling actors with an interpretive model for their own actions and the actions of others' (p. 262). The ethos of reciprocity and exchange is a core cultural value among the Tamang and has associated secondary values, including having many children in order to reproduce broader exchange and social networks and providing their children with a share of inheritance.

Over time, these secondary values had come in conflict. Thus, the men justified their decision to undergo a vasectomy as necessary to protect the inheritance shares of the children they had already borne. The conflict was provoked by changes in the material context and individual experience, but it was the actors' commitment to preserving a core cultural value that led them to vasectomy.

An even more radical step is taken by Caroline Bledsoe, Fatoumata Banja, and Allan Hill (1998), who contrast the cultural model of high fertility that population scientists use to guide their analyses with the cultural model of reproduction that rural Gambians use to guide their behavior. The former is familiar to demographers: age and number of children lead a woman to change her fertility behavior, those who use contraceptives intend to limit live births, and early contraceptive users are expected to be educated and relatively autonomous from family pressure. Some users of contraception in the Gambia conform to this Western model of reproduction. For others, however, their behavior is inconsistent with this model: they are motivated to use Western contraception because they want to *protect* the woman's reproductive potential so that she may bear more children, either in the current marriage or, should this marriage fail and another be necessary, in a future marriage.

Moreover, reproductive potential is measured not in terms of age, as in Western models, but in terms of 'body resource expenditure,' the loss of 'muscles,'

strength, and blood that accompany childbearing. Indeed, the authors found that the highest rates of use of Western contraceptives to be among women who had experienced reproductive mishaps—abortion or miscarriage, or child death—and who wanted to 'rest' in order that they might bear more children. Fertility is highly important in this society, but it is less the number of children that matter to the women than marriage.

There are other examples of the way that a more interpretive and activist view of culture has informed demographic analyses of fertility transitions, not only in the work of these authors but in the work of others (e.g. Szreter 1993, Townsend 2000, Kaler 2000).

6. *Conclusions*

Both the older and the newer views of culture can be useful in understanding fertility transitions. Culture may at times be a barrier to fertility change, but it may also facilitate fertility change, and culture itself may change as people alter old reproductive models or adopt new ones. The critics of the use of culture by demographers are fundamentally correct in arguing that the perception that culture is a barrier to rational action—a view that has been adopted by those writing on the spread of HIV/AIDS—is based on a view of third world people as implausibly passive. Indeed, the transformation in reproductive behavior in much of the third world between 1960 and 1990 suggests that new cultural models of reproduction can supersede old ones, or old models can lead to different behavior in new circumstances.

What are the implications of this newer view of culture for demographic analysis? Most demographers will undoubtedly continue to do the kind of statistical analyses they have always done, and are not likely to abandon these for the methods of cultural anthropologists or sociologists of culture. Thus, they will continue to require labels for aspects of culture, labels that can be attached to individuals or groups. The emphasis of the new approaches on communication as a source of cultural flexibility and change suggests that linguistic labels may still be useful, supplemented by attempts to open the 'black box' of culture in order to understand better differences in perceptions and behaviors across (and within) groups. Considering other ways of categorizing individuals and groups may also be fruitful for statistical analysis.

The newer view of culture suggests, however, that what these categories mean may change as the perceptions and motivations of actors change, as well as how that change might occur and the implications for demographic behavior. Most importantly, the recent empirical work guided by the newer understanding of culture suggests that demographic analysis would benefit were our analytic models closer to the models that actors who produce fertility transitions use to guide their behavior.

When demographers have taken culture seriously, it has typically been local cultures in third world countries. A few, however, have suggested new directions. One is the interpretation of what might be called governmental and bureaucratic cultures, both national and global. The second is that demographers examine their own culture, one that has a particular model of reproduction and views culture as a barrier to reproductive change. The newer approaches to culture as explanations for fertility change suggest that the current interpretations of culture in anthropology and sociology may change, and that the culture of demographers may also be flexible.

Bibliography

- Appadurai A 1990 Difference and disjunction in the global cultural economy. *Public Culture* 2: 1–24
- Basu A M, Amin S A 2000 Conditioning factors for fertility decline in Bengal: History, language identity and an openness to innovations. *Population and Development Review* 26: 761–94
- Bledsoe C, Banja F, Hill A G 1998 Reproductive mishaps and Western contraception: An African challenge to fertility theory. *Population and Development Review* 24: 15–57
- Bongaarts J, Watkins S C 1996 Social interactions and contemporary fertility transitions. *Population and Development Review* 22: 639–82
- Bourdieu P 1977 *Outline of a Theory of Practice*. Cambridge University Press, Cambridge, UK
- Caldwell J C, Caldwell P 1987 The cultural context of high fertility in sub-Saharan Africa. *Population and Development Review* 13: 409–37
- Coale A J 1973 The demographic transition reconsidered. In: *International Population Conference, Liege, 1973*. International Union for the Scientific Study of Populations, Liege, Vol. 1
- Ekeh P P 1990 Social anthropology and two contrasting uses of tribalism in Africa. *Comparative Studies in Society and History* 32: 660–700
- Fricke T 1997 Culture theory and demographic process: Toward a thicker demography. In: Kertzer D, Fricke T (eds.) *Anthropological Demography: Toward a New Synthesis*. University of Chicago Press, Chicago
- Geertz C 1983 *Local Knowledge: Further Essays in Interpretive Anthropology*. Basic Books, New York
- Greenhalgh S 1995 Anthropology theorizes reproduction: Integrating practice, political economic, and feminist perspectives. In: Greenhalgh S (ed.) *Situating Fertility*. Cambridge University Press, Cambridge, UK
- Greenhalgh S 1988 Fertility as mobility: Sinic transitions. *Population and Development Review* 14: 629–74
- Hammel E A 1990 A theory of culture for demography. *Population and Development Review* 16: 455–86
- Hannerz U 1992 *Cultural Complexity: Studies in the Social Organization of Meaning*. Columbia University Press, New York
- Holy L, Stuchlik M 1983 *Actions, Norms and Representations*. Cambridge University Press, Cambridge, UK
- Kaler A 2000 Who has told you to do this thing?: Toward a feminist interpretation of contraceptive diffusion in Rhodesia, 1970–1980. *Signs: Journal of Women in Culture and Society* 25: 677–708
- Kerzer D I 1995 Political-economic and cultural explanations of demographic behavior. In: Greenhalgh S (ed.) *Situating Fertility*. Cambridge University Press, Cambridge, UK
- Kerzer D I 1997 The proper role of culture in demographic explanation. In: Jones G W, Douglas R M, Caldwell J C, D'Souza R M (eds.) *The Continuing Demographic Transition*. Clarendon Press, Oxford, UK
- Kertzer D I, Fricke T 1997 Toward an anthropological demography. In: Kertzer D, Fricke T (eds.) *Anthropological Demography: Toward a New Synthesis*. University of Chicago Press, Chicago
- Knodel J, van de Walle E 1986 Lessons from the past: Policy implications of historical fertility studies. In: Coale A J, Watkins S C (eds.) *The Decline of Fertility in Europe*. Princeton University Press, Princeton, NJ
- Kreager P 1985 Demographic regimes as cultural systems. In: David C, Schofield R (eds.) *The State of Population Theory*. Basil Blackwell, Oxford, UK
- Lesthaeghe R 1989 Production and reproduction in sub-Saharan Africa: An overview of organizing principles. In: Lesthaeghe R J (ed.) *Reproduction and Social Organization in sub-Saharan Africa*. University of California Press, Berkeley, CA
- Meyer J W, Boli J, Thomas G M, Ramirez F 1997 World society and the nation-state. *American Journal of Sociology* 1: 144–81
- Moore H L 1986 *Space, Text and Gender: An Anthropological Study of the Marakwet of Kenya*. Cambridge University Press, Cambridge, UK
- Moore S F 1994 *Anthropology and Africa: Changing Perspectives on a Changing Scene*. University of Virginia Press, Charlottesville, VA and London
- Renne E P 1996 Perceptions of population policy, development, and family planning programs in Northern Nigeria. *Studies in Family Planning* 27: 127–36
- Swidler A 1986 Culture in action: Symbols and strategies. *American Sociological Review* 51: 273–86
- Szreter S 1993 The idea of demographic transition and the study of fertility change: A critical intellectual history. *Population and Development Review* 19: 659–701
- Townsend N 1999 Male fertility as a lifetime of relationships: Contextualizing men's biological reproduction in Botswana. In: Bledsoe C, Guyer J, Lerner S (eds.) *Fertility and the Male Life Cycle in the Era of Fertility Decline*. Oxford University Press, New York
- Watkins S C 1993 If all we knew about women was what we read in demography, what would we know? *Demography* 30: 551–77
- Watkins S C 2000 Local and foreign models of reproduction in Nyanza Province, Kenya. *Population and Development Review* 26: 725–59

S. Cotts Watkins

Fertility Transition: East Asia

East Asia experienced extremely rapid declines in fertility after World War II. In the last decades of the twentieth century, fertility declined further to below replacement levels in some countries of the region, notably in Japan, the Republic of Korea (called 'Korea' hereafter for linguistic simplicity), and the

Republic of China (called 'Taiwan' hereafter for the same reason). Focusing on these three countries, this entry looks at the patterns and factors associated with the postwar fertility transition. After an overview of trends in fertility levels, the entry discusses changes in the two major demographic factors that influence fertility: the timing of women's marriage and fertility within marriage. The article then discusses socio-economic factors that influence fertility including women's education and employment.

1. Fertility Trends

Although there were differences in timing and magnitude, Japan, Korea, and Taiwan all experienced a dramatic downturn in fertility after World War II (see Fig. 1). In Japan the total fertility rate (TFR) declined from 4.5 children per woman in 1947 to 2.0 in 1957. After this rapid decline, fertility stabilized at about the replacement level (a TFR of 2.0 to 2.1 children per woman) until it started to decline again in the mid-1970s, reaching an all-time low of 1.3 children per woman in 1999. High fertility persisted in Korea until rapid economic development began in the early 1960s. Then the TFR dropped precipitously from 5.9 children per woman in 1963 to 1.6 in 1987. This remarkable decline was followed by a modest upturn in the early 1990s and subsequent stabilization at a level of 1.6 to 1.8 per woman. Taiwan's fertility transition was equally dramatic. Starting in the early 1950s, the TFR declined steadily from 7.4 children per woman in 1951 to 1.9 in 1985. Since the mid-1980s, TFR has stabilized at 1.7 to 1.8 children per woman.

In summary, Japan experienced two fertility transitions, the first a rapid decline from historically high to replacement levels of childbearing, which occurred

shortly after World War II, and the second an ongoing decline from replacement to very low levels of fertility. Korea and Taiwan both experienced one continuous fertility transition from very high to below replacement levels.

2. Demographic Factors in Fertility Decline

In all three of these countries the fertility transition is related to two demographic factors, namely delayed marriage among women and declining fertility within marriage. Because out-of-wedlock childbearing is minimal, these two factors determine the overall levels of fertility. Demographic analysis indicates that roughly 90 percent of the decline in Japan's TFR between 1950 and 1975 was related to a decline in marital fertility, and the remaining 10 percent was related to delayed marriage among women (Tsuya and Mason 1995). Similarly, approximately 80 percent of the TFR decline in Korea from 1960 to 1990 was due to declining fertility within marriage, and the remaining 20 percent was due to delayed marriage. In Taiwan, around 75 percent of the decline in fertility between 1960 and 1985 was due to a decline in marital fertility, and the remaining 25 percent was due to delayed marriage. By contrast, the fertility decline in Japan after 1975 was due solely to delayed marriage among women, while marital fertility showed moderate increases. In a corresponding manner, around 40 percent of the fertility decline from 1980 to 1990 can be attributed to delayed marriage in Korea, and around 85 percent in Taiwan.

2.1 Decline in Marital Fertility

The dramatic fertility declines in these three East Asian countries were caused primarily by a reduction in fertility among married women, who gained control of their family size through contraception and induced abortion. The rate of contraceptive use among married women under the age of 50 increased steadily in Japan from around 20 percent in 1950 to 62 percent in 1979 (Kobayashi 1990). After 1979, the prevalence of contraception stabilized. With the expansion of government family-planning programs, contraceptive practice spread even more rapidly in Korea and Taiwan. In Korea contraceptive use among married women aged 15 to 44 increased from 20 percent in 1966 to 70 percent in 1985 (Palmore and Bulatao 1989). In Taiwan contraceptive use among married women aged 20 to 39 was 23 percent in 1965; among married women aged 15 to 49 it was 78 percent in 1985 (Freedman et al. 1994).

Induced abortion is a traditional means of fertility control long practiced in East Asia. Along with the rapid increase in contraceptive use, rates of induced abortion also increased in the three countries. In Japan the proportion of married women of repro-

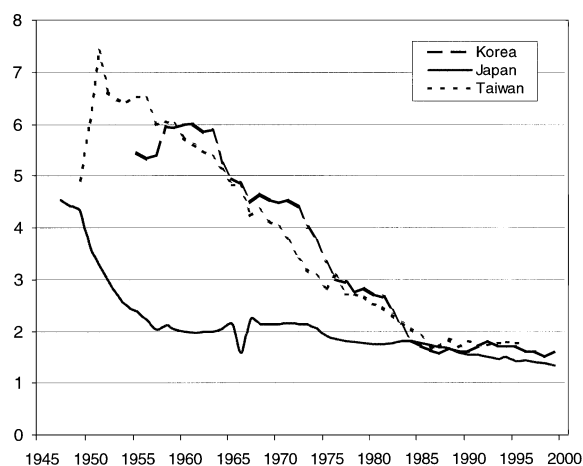


Figure 1
Total fertility rates in Japan, Korea, and Taiwan after World War II

ductive age reporting having experienced an abortion increased from 15 percent in 1952 to 41 percent in 1961 (Muramatsu 1990). Since the early 1960s, this proportion has remained in the range of 30–40 percent. In Korea the proportion of married women of reproductive age reporting having experienced an abortion increased from around 10 percent in the mid-1960s to nearly 40 percent in the late 1970s (Cho et al. 1982). Although the rate declined somewhat in the 1980s, abortion remains an important option for terminating unwanted pregnancies. The reported use of abortion among married women of reproductive age also rose steadily in Taiwan from the 1960s to the 1980s (Freedman et al. 1994).

2.2 *Delayed Marriage*

As we have noted already, being older at the time of marriage has also played an important role in the fertility transitions. The postponement of marriage for women in the three countries is remarkable in light of the East Asian tradition of universal marriage and the cultural expectation that women should marry by their mid-twenties. Between 1975 and 1995 the proportion single rose from 21 percent to 48 percent among women aged 25 to 29, and from 8 percent to 20 percent among women aged 30 to 34 in Japan. Korean women are also marrying later, although almost all marry by their mid-30s. Between 1955 and 1995, the proportion never married increased from 21 percent to 83 percent among women aged 20 to 24, and from 3 percent to 29 percent among women aged 25 to 29. Taiwan has experienced an equally dramatic delay in women's first marriage. From 1956 to 1995, the proportion single rose from 29 percent to 79 percent among women aged 20 to 24, and from 5 percent to 40 percent among women aged 25 to 29. Since the mid-1970s, the percentage single has also risen among women in their early thirties—from 3 percent in 1975 to 16 percent in 1995.

3. *Socioeconomic Factors in Fertility Decline*

3.1 *Women's Educational Attainment*

Dramatic socioeconomic changes after World War II brought about important changes in women's status; educational attainment and employment grew, and these changes have also affected fertility. Women's education is a key factor affecting fertility through the timing of marriage, marital fertility, employment, and earnings. In Japan most women were already receiving primary education before World War II; the most notable postwar change has been the steady, rapid increase in the proportion of women with higher education. The proportion of female high-school graduates advancing to higher education increased from 6 percent in 1960 to 48 percent in 1995. The proportion of women aged 25 to 29 who had higher

education increased from 10 percent in 1970 to 43 percent in 1990, reaching a par with men in the same age group.

In contrast with Japan, the educational attainment of Korean women was very low in the early postwar years and a sizable proportion of older women had no formal education. This situation has changed dramatically. Only 3 percent of women born in 1936–40 had higher education, compared with 41 percent of women born in 1971–5. The corresponding proportion for men increased from 16 percent to 44 percent; thus the once vast gender gap has almost disappeared. Taiwan has also made impressive improvements in women's education. Educational attainment among Taiwanese women was very low in the early postwar years, with a sizable proportion of older women illiterate. Only 20 percent of women aged 20 to 24 had a high-school education in 1974, increasing to 34 percent in 1995. Since the mid-1970s, about 11 percent of women in this age group have experienced higher education.

3.2 *Economic Changes*

Japan, Korea, and Taiwan all experienced extremely rapid economic growth after World War II. Since the mid-1950s, Japan's per capita GNP in constant prices has quadrupled every ten to fifteen years, so that Japan is now the second largest economy in the world after the USA. Between the mid-1960s and the mid-1980s, Korea's per capita GNP also quadrupled every 10 to 15 years. Over the same period, Taiwan's per capita GNP doubled every 10 years.

In the course of rapid economic growth, Japan, Korea, and Taiwan experienced rapid urbanization. In Japan, the proportion of the population living in urban areas increased from 37 percent in 1950 to 63 percent in 1960 and 78 percent in 1995. Korea has experienced even more rapid urbanization, with the urban population rising from 21 percent in 1950 to 81 percent in 1995. In Taiwan the proportion of the population living in cities of 50,000 or more increased from 27 percent in 1952 to 73 percent in 1988.

Along with rapid economic growth and urbanization, these countries experienced dramatic changes in the structure of their economies. In Japan the proportion of the workers employed in primary industries fell from 49 percent in 1950 to 6 percent in 1995. In Taiwan it dropped from 56 percent in 1952 to 11 percent in 1995. In Korea the proportion of workers engaged in agriculture fell from 66 percent in 1960 to 12 percent in 1995.

3.3 *Women's Employment*

With economies shifting rapidly from primary to secondary and tertiary industries, female employment also underwent notable changes. In Japan, although the overall female labor force participation rate has

remained roughly 50 percent over the past 40 years, the type of female employment has changed dramatically. The proportion of employed women who were family workers declined steadily from 43 percent in 1960 to 12 percent in 1995, while the proportion of employed women in paid employment increased from 41 percent to 78 percent. Although Japanese women are still more likely to work before and after rather than during their peak childbearing and child-rearing years, they are increasingly likely also to work between the ages of 25 and 34.

In Korea, female labor force participation increased from one-quarter in 1960 to about one-half in 1995, but women in their peak childbearing and child-rearing years are still much less likely to work than younger or older women. Whereas labor force participation among women aged 20 to 24 increased considerably in the last decades of the twentieth century, participation rates for women aged 25 to 34 have remained stable. This suggests that while employment of young (mostly single) women has increased, paid employment among married women has remained limited. In Taiwan, overall female labor force participation rose from 19 percent to 45 percent between 1956 and 1995, and rates for women at peak childbearing and child-rearing ages increased especially dramatically—from 17 percent to 65 percent among women aged 25 to 29, and from 15 percent to 58 percent among women aged 30 to 34.

See also: East Asian Studies: Economics; East Asian Studies: Gender; East Asian Studies: Society; Family Size Preferences; Family Systems and the Preferred Sex of Children; Fertility: Institutional and Political Approaches; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations; Fertility Transition: South Asia; Fertility Transition: Southeast Asia; Land Rights and Gender

Bibliography

- Cho L, Arnold F, Kwon T H 1982 *The Determinants of Fertility in the Republic of Korea*. National Academy Press, Washington, DC
- Freedman R, Chang M C, Sun T H, Weinstein M 1994 The fertility transition in Taiwan. In: Thornton A, Lin H-S (eds.) *Social Change and the Family in Taiwan*. The University of Chicago Press, Chicago, pp. 264–304
- Kobayashi K 1990 General remarks on contraception. In: Population Problems Research Council (ed.) *Summary of Twentieth National Survey on Family Planning*. Mainichi Shimibun, Tokyo, pp. 39–52
- Muramatsu M 1990 Induced abortion. In: Population Problems Research Council (ed.) *Summary of Twentieth National Survey on Family Planning*. Mainichi Shimibun, Tokyo, pp. 83–97
- Palmore J A, Bulatao R A 1989 The contraceptive method mix: An overview. In: Bulatao R, Palmore J A, Ward S (eds.) *Choosing A Contraceptive: Method Choice in Asia and the United States*. Westview, Boulder, CO, pp. 3–24
- Tsuya N O, Mason K O 1995 Changing gender roles and below-replacement fertility in Japan. In: Mason K O, Jensen A (eds.)

Gender and Family Change in Industrialized Countries. Clarendon Press, Oxford, UK, pp. 137–67

N. O. Tsuya

Fertility Transition: Economic Explanations

1. Approaches of Economics and Demography to the Fertility Transition

With the establishment of modern economic growth, fertility has tended to decline, first across high income countries starting largely in the last quarter of the nineteenth century, and then occurring even more rapidly across most low income countries in the last quarter of the twentieth century (Kuznets 1966, Schultz 1981). Economic models of the allocation of time and wealth of families among production and consumption activities suggests several hypotheses linking changes in the economic environment in which people live to their reproductive demands and fertility (Becker 1960, Mincer 1963). Empirical counterparts for the price, income, and environmental constraints on people's fertility choices have been used to explain aspects of fertility and other related outcomes of the family, strengthening the view that this household demand model is a source of insights into the determinants of the historic fertility transition (Schultz 1974, Becker 1981).

This economic approach to fertility determinants has been criticized by those who believe that fertility is not a rational response to changing economic conditions (Cleland and Wilson 1987). The proposed alternative to the economic approach, called 'ideational change' emphasizes social interactions, but has not yet evolved to a form that is empirically distinguishable. However, research is trying to quantify the role of groups external to the family such as peers and neighborhoods, and to interactions between family members as postulated in cooperative and non-cooperative game strategies within families. But because these groups are generally self selected, it is complex to specify the factors that affect the formation of the group and do not also affect the behavior of individuals who enter such groups (Heckman 1979, Manski 1995). There is active debate on how to apply the economic framework and to exploit statistically cross sectional and time series variation in fertility to improve our understanding of the determinants of fertility (e.g., Schultz 1997).

There is a fundamental methodological divide between those who want to describe a 'structural model' that involves interactions between choices made at the household level, including those choices spanning an adult's lifetime, such as between a woman's fertility, marriage, and labor force history, and those who

investigate a 'reduced form model' where only the fixed constraints on household choice are considered as potential explanatory variables—market prices, income sources, technologies, and the environment. It is widely recognized that the associations between lifetime choices represent more (or less) than a simple causal effect of one choice on the other. These covariations between family choices embody possible effects transmitted in both directions, effects of unobserved factors which impact both outcomes, and heterogeneity due to unobserved differences, perhaps in individual preferences (Schultz 1981). Strong identification restrictions are generally required to estimate from nonexperimental data the critical parameters to such 'structural models,' and therefore the empirical conclusions drawn from them are often controversial, because they depend on a host of untested and debatable modeling choices (Manski 1995).

The alternative approach is to estimate fertility as a function of explanatory variables that are not family choice variables or are not likely to be impacted by unobserved factors that could also affect fertility. By delimiting the potential list of explanatory variables, these 'reduced form' relationships represent a summation of the direct effect of the environmentally fixed constraints on the couple's fertility, plus the possible indirect effect of the same constraints on fertility that operate through the modification of other lifetime choice variables, such as the woman's age at marriage, extent of labor force activity outside of her home, contraceptive and sexual behavior, etc. The limitation of reduced form estimates is that they do not necessarily clarify the mechanism by which the constraints impact fertility, for example through the timing of marriage, or the adoption within marriage of various so-called 'natural' or 'unnatural' forms of birth control. The strength of reduced form estimates is that they claim to provide an unbiased approximation for the total effect of the constraint on fertility, which is often a central concern of the social scientist, forecaster, or policy maker.

Some historical demographers assign special importance to the moment when married women begin to restrict their birth rates in response to their prior number of births, or surviving children (Coale and Watkins 1986). Although this may be a well-defined and interesting threshold that predicts the onset of the long term decline in fertility in certain European populations, other populations may adopt other mechanisms to reach their changing fertility goals, such as a change in median age at marriage (i.e., Malthus' preventive check), an increase in celibacy or the proportion who never marry in a birth cohort, an increase in the spacing of births within marriage, as appears to be evident in Africa today (Caldwell et al. 1992, Shapiro and Tabashe 2000), or a change in non-marital fertility. Even in the region of Northern Europe, standard measures of the emergence of parity-specific fertility control within marriages are not

precisely defined by existing data (Guinnane et al. 1994). Thus, economists generally consider comprehensive measures of cohort fertility, and do not focus on one component, such as marital fertility.

2. Stages of the Demographic Transition

The 'demographic transition' refers to a sequence of three periods. In the first period, fertility and mortality rates are high and mortality tends to be highly variable, with population growth fluctuating widely about a moderate long run trend of growth or decline. Short run responses in fertility (and mortality) can be attributed to cycles in weather or harvests or possibly to other exogenous variables using time series statistical methods, such as vector-auto-regression.

In the second period, age-specific mortality rates decrease gradually, raising life expectancy at birth from 30–35 years in pre-industrial Europe, to 70–75 today in the high income countries, and from 25–30 years in the low income countries in the 1920s, to levels which today range from about 45–73, excluding countries ravaged by war or the AIDS epidemic. This second period is one of accelerating population growth, which some observers such as Malthus attributed to improved food supplies, and more specifically to nutrition, often initiated by technical progress in agriculture, followed by improved industrial productivity, and advances in transportation and communication (Fogel 1999). A free press capable of publicizing food shortfalls in the twentieth century may have further alleviated excess mortality in the wake of periodic famines. Improvements in private and public health technologies are assigned an important role in reducing mortality, but not much before the start of the twentieth century.

In the third period, the secular decline in fertility begins, after which the number of births per woman falls by more than half, from five or six to about two, in the high income countries, and six to eight to about two or three in most low income countries. When historically high fertility levels have declined by more than 10 percent, they are not expected to rise again on a sustained basis (National Research Council 2000). Medium-term swings in fertility are nonetheless distinguished from time series, occurring in response to business cycles (e.g., the post-World War II baby boom), wars, and economic shocks, as during the transition from centrally planned to market oriented economies in the 1990s in Eastern Europe and Russia, or recently due to reversals in sub-Saharan Africa.

3. Economic Determinants of Fertility

Notestein (1945) described the early dimensions of the demographic transition, and speculated on the economic pressures which he expected to reduce fertility in

the low income world: rising cost of children in urban life, individualistic promotion of child health, and education. Economic models of a lifetime fertility choice emphasize different facets of this decision process, but the trade-off between quantity of children and their quality (i.e., education and health) has become a central feature (Becker 1960, Mincer 1963, Schultz 1981). Because women's time in child-care is a large fraction of the modern opportunity cost of fertility, her allocation of time is treated as jointly determined with fertility. Many cross-sectional and a few time-series studies have proceeded to relate statistically these fixed constraints to fertility outcomes.

A couple's lifetime wealth (endowments of human capital and nonhuman capital) is expected to increase with modern economic growth, and to encourage parents to demand more children, if the relative costs and functions of children did not change (Becker 1960, Schultz 1981). However, for reasons discussed below, the opportunity cost of children increased along with the cost of other goods and activities that complement the rearing of children, whereas the private price of some substitutes for children, such as old-age pensions and healthcare, have been socialized and subsidized by governments. Marked improvements in the technology of birth control were developed in the 1950s, translating into the rapid diffusion of the pill (oral steroid contraceptive) and the IUD (inter-uterine device) to most regions of the world. The monetary and psychic cost and inconvenience (independent of coitus) of avoiding too many births was reduced, and should thereby have reduced the number of births. Family planning programs are not, however, a necessary condition for the occurrence of a fertility transition, because birth rates declined in high income countries without the benefit of new control technologies, or of family planning extension activities designed to foster the adoption of birth control. Finally, improvements in child health technology increased the proportion of births that survived to adulthood. These health developments put downward pressure on birth rates, under two assumptions: parent reproductive goals are defined primarily in terms of their number of surviving children, and these demands for survivors are highly price inelastic (Schultz 1981).

The actual expenditures of a couple on each of their children is not a suitable indicator of the 'market price' of a child on which a couple would base their demand for children. Expenditures per child are also a family choice, which Becker called the demand for 'child quality.' The market-determined prices of the inputs that households combine to rear children are the relevant prices that condition fertility. There is no consensus on which inputs to children are most important and how their unit prices are measured.

If one accepts Becker's (1960) hypothesis that quality and quantity of children are viewed by parents as substitutes, and much of the variation in child quality is embodied in the education parents give their

child, then the net price of schooling would affect negatively the demand for quality and affect positively the demand for quantity of children. There is growing evidence in the twentieth century that wage returns to schooling increased, and this can be viewed as a decline in the price of schooling, encouraging parents and societies to invest more resources in the education of children. If parents view schooling per child as a substitute for numbers of children, as is now often empirically confirmed (e.g., Rosenzweig and Wolpin 1980), the increasing wage return to schooling among workers should encourage parents to substitute child schooling for number of children. The increase in wage returns to schooling can be attributed to a technological bias favoring skilled labor, or to a complementarity in production between the schooling of workers and the accumulating stock of technical knowledge available in the global economy.

Becker (1981) also assumed that the income effect on the demand for child quality was a larger positive magnitude than the income effect on the demand for child quantity. This might also have encouraged parents to substitute child quality for child quantity as their incomes increased from one generation to the next. Here was a second possible explanation for the anomalous fall in fertility in a period when *per capita* income was increasing.

4. The Increasing Value of Parents' Time as an Opportunity Cost of Children

In the twentieth century, the fraction of national income paid out to labor has risen, despite the increase in capital employed per worker and the decrease in hours worked/year per worker (Kuznets 1966, Fogel 1999). Wage rates have grown relative to *per capita* income, implying that consumption activities that use more time of the consumer have become relatively more expensive over time. Children are thought to be a more 'time-intensive commodity' than the average of all consumption commodities, and thus the opportunity costs for household's to produce and consume a child have risen as parent wages have risen, offsetting some part of the increased demand for children that might have been expected in a period when *per capita* income is rising (Mincer 1963, Becker 1965).

The household production-consumption model drew attention to gender specialization in social production, and to the 'female time-intensity' of children, thereby suggesting the woman's value of time could be a key component in the price of children. In environments where the opportunity cost of female (i.e., mother's or child-care giver's) time is higher, fertility is expected to be lower, other things equal (Schultz 1981). The market wage of women might thus be an 'opportunity price' variable, and it has notably increased in many countries in the twentieth century at a time when fertility has fallen. But just as 'parental

expenditures per child' does not represent a fixed market-determined price of a child, the woman's market wage does not represent a fixed price of a woman's time input to child rearing. First, the wage a woman receives in the labor force is affected by her prior job experience, which accumulates at different rates depending on her preferences among types of work and the allocation of her time. Consequently, a woman's wage is an endogenous choice variable closely linked to her lifetime allocation of time, which in turn is intimately related to her fertility decisions (Mincer 1963).

A second problem with the woman's wage variable is that it is observed only for a self-selected fraction of women who work for a wage in the survey period, and this fraction has tended to rise over time in many populations, and to differ by socioeconomic strata within a population. It is possible that women who work for wages are those who receive an unusually high market wage offer, suggesting that the observed average wage of women overstates the value of time of an average woman. Conversely, women who are unusually productive in nonwage work, and thus are less likely to participate in the wage labor force, may still have an above average market-wage offer, in which case the observed average wage of women would understate the value of time of an average woman. Sample selection bias could distort the wage as a measure of the value of women's time, and thus bias estimates of the woman's wage effect on fertility, probably overstating the causal connection in time series (Heckman 1979, Schultz 1981).

5. Determinants of the Value of Women's Time and the Cost of Children

The variation in the opportunity price of children that arises from the variation in women's opportunity value of their time remains a key economic determinant of fertility, and econometric procedures are required to estimate the importance of this connection for the fertility transition. In other words, because the observed wage of women cannot be treated as a fixed determinant of fertility, it cannot be included in a reduced form equation for fertility. How then, does one approximate in a more structural model the effect of employment and productive opportunities for women on the opportunity cost of a woman's time, which in turn affects fertility?

One approach to this problem is to employ the woman's education as a proxy for her wage; schooling tends to be the best variable available for predicting in the cross-section women's wages, just as it is for men's wages. For most women their human capital investment in schooling is completed before childbearing starts, and thus the schooling decision can be viewed as predetermined with respect to fertility. If the schooling decision is made by the child's parents, it

may also be argued that the woman's schooling is unaffected by her adult preferences between labor market and fertility choices, and then schooling may be a satisfactory 'instrumental variable' for predicting the adult lifetime wage opportunities for women, which constrain the choice of fertility.

A second approach is to treat the structure of wages in the local labor market as fixed and thus independent of the woman's preferences. This appears to be a reasonable strategy, proceeding from the individual wage to the community averaged wage, and thus eliminating the idiosyncratic individual variation in preferences between working and fertility. Analyses of fertility can then include all women based on the regional wage, and not be restricted to a selected sample of women who are currently reported working for a wage. But another problem could still bias the estimated effect of women's regional wage opportunities on regional or individual fertility. Because an increase in the labor supplied by women in a region will tend to depress the market-clearing wage women receive locally, any omitted regional variables that contribute to greater female labor force participation are expected to also be correlated with lower wages, and hence be correlated with higher fertility. Consequently, these regional omitted labor supply variables would introduce a positive bias in the estimated effect of woman's regional wage on fertility, masking the theoretically expected negative impact of aggregate demand-induced variation in women's wages on fertility.

One estimation strategy to deal with this regional omitted variable problem is to specify a regional variable that shifts only the aggregate regional economy's demand for female labor (relative to the demand for male labor), and thereby affects the local wage available to women (relative to men). Variables of this form which shift the aggregate demand schedule for women's labor are candidates for being an 'instrumental variable' that can account for demand-induced variation in women wages, and can thus be used to obtain an unbiased estimate of the effect of women's wages on fertility. One such demand variable is the regional location of natural resources or the composition of industries, which influence the gender composition of employment opportunities and wages.

Coal mines, for example, employ mostly men, and communities dependent on coal mines for their livelihood have relatively low wages for women. Historically, these coal communities in the UK and the United States reported relatively high fertility levels, other things equal (Haines 1979). In the Northern provinces of Sweden where the price and quantity of timber increased at the end of the nineteenth century due to exports, men's wages increased relative to women's wages, and fertility remained high in these regions well into the twentieth century. By contrast, fertility in the Southern provinces of Sweden declined after the 1860s, as women's wages increased relative to

those of men, in response to higher output prices for exports of livestock, dairy, and food manufactured products, which employed relatively more women (Schultz 1985). Natural resource exports strengthen government finances by providing an inelastic source of taxation, notably in the case of oil exports, with a tendency to sustain customary social institutions that have been associated with relatively low levels of female education, low female labor force participation, and high levels of fertility, holding constant for other factors (Schultz 1994).

Labor economists have noted since the 1960s that in regions where the share of employment in certain industries, for example the services, is relatively large, female labor force participation rates tend to be higher. The importance of part-time, less-skilled, and physically less-demanding jobs may thus be associated with a larger share of women in the labor force, and a rise in the economic status of women relative to men. This form of 'female bias' in the industrial composition of local employment is found to predict increased women's wages relative to men's and lower fertility.

Changes in fertility over time can then be explained by changes in factors external to the household sector which govern the expansion of employment opportunities in industries that hire a disproportionate share of women (or men). The price of the industry's output when it is determined in international markets is one such factor. Terms of trade indexes linked to gender-specific employment and wages opportunities can predict women's wages relative to men's and provide a basis for estimating how demand-driven variation in women's employment opportunities influence fertility. However, these changes in relative prices and employment prospects must be perceived by couples as sufficiently long term in nature that they feel the need to adapt accordingly their reproductive goals.

6. Evidence from European Fertility Transitions

The decline in the price of food grains in Europe in the nineteenth century led to such a profound reallocation of labor and allied resources. With the opening of the Erie canal in 1825 and the reduction in ocean transportation costs, exports from the USA, Russia, and other world grain exporters depressed grain markets across Europe. Although restrictions of trade attempted to protect domestic agriculture, European farmers gradually adapted to their changing comparative advantage and left agriculture for urban jobs or emigrated, whereas those who remained in agriculture reallocated their effort from grain to livestock production and to the expansion of food and dairy processing industries. Because grain production traditionally employed more male labor, and livestock and dairy production used more female labor, these changing output prices boosted women's wages relative to men's. This reallocation of resources occurred more rapidly where economies were more open to

international trade in agricultural commodities. This transformation of composition of European agriculture or its consequences on the productive roles of women and men is not quantitatively studied in the Princeton European Fertility Project, although these changes could have influenced the diffusion of the fertility transition (Coale and Watkins 1986).

In Sweden, as noted earlier, these agricultural relative price changes were closely associated with the improvement in women's wages relative to men's, and with declines in fertility, by province over time, starting after 1860. The level of real wages for men was related to younger childbearing, but not to a change in total fertility rates. About one-third of the 25 percent decline in total fertility rates for Sweden in the period 1860 to 1910 can be attributed to these externally driven increases of women's wage opportunities, whereas the rest of the decline in Swedish fertility in this half century is associated with urbanization and the increase in child survival, under the assumption that these are exogenous developments (Schultz 1985). A similar demographic pattern unfolded in Denmark, which was also relatively open to international agricultural trade, while its demographic transition was only slightly less rapid than in Sweden.

The German fertility transitions has also been analyzed from 1880 to 1910. When regional persistent or fixed-effects are introduced in the analysis of times series of cross sections, the evidence of strong partial effects of industrialization and urbanization emerge (Richards 1977). Smaller regional units were examined in Prussia from 1875 to 1910, with the same regional fixed-effect specification for fertility (Galloway et al. 1994). However, the female labor force participation rates are used to explain fertility in a reduced form specification, as though the labor force participation of women was a fixed constraint on fertility decisions. Not surprisingly, this choice variable reflecting women's time allocation accounts for the largest fraction of the fertility decline, and if it is jointly determined by women with their fertility, as seems likely, the reported estimates are subject to simultaneous equation bias and are difficult to interpret as evidence for any causal model.

7. Advancement in Women's Schooling

Another widespread development in the world, which has contributed to increasing the cost of children relative to most other forms of consumption, is the advance in women's schooling. In the USA, by the start of the twentieth century, the gender gap in schooling was modest, except for college education. But this gap remained substantial in most countries in Southern and Eastern Europe until after the World War I. Moreover, women's schooling attainment began to gain relative to that of men in low income countries only in the last few decades. This progress has been particularly sluggish in South Asia where the

fertility transition has also been most gradual. Given the empirical tendency for wages to increase for women by at least the same percentage as they do for men when they obtain an additional year of schooling, the closure in the gender gap in schooling implies that women's wages are likely to continue to increase in most parts of the low income world for some time, and indeed increase at a more rapid rate than will the wages of men (Schultz 1995). The factors that have triggered these differential trends in the schooling of women and men have not been thoroughly studied, but the expansion of women's schooling closely parallels the decline in fertility across high and low income countries since 1960. Indeed, changes in women's schooling accounts for the largest share of the changes in total fertility rates observed in low income countries since 1960, and in Africa since 1980 (Schultz 1994). Even in an environment where economic progress has been minimal since the 1960s—Kinshasa City in Congo—the increasing rate of female education is the key factor explaining different levels of fertility across the population, and accounting for the decline in fertility in the 1980s and 1990s (Shapiro and Tambashe 2000).

However, educated women could have lower fertility for many reasons, in addition to the greater opportunity cost of their time in child rearing. The social and intellectual advantages that educated women enjoy help them in deciphering, adopting, and using effectively, new and old forms of birth control, and thereby avoiding unwanted births.

Female schooling may also affect fertility through its impact on the risks of child mortality. Censuses and surveys after the World War II have documented for dozens of countries that child mortality rates tend to be 5 to 10 percent lower for mothers who have completed an additional year of schooling (Schultz 1981, 1997). However, the proportionate decline in fertility associated with women's education is larger than the proportionate increase in their child survival rates. Therefore, better educated women have smaller numbers of surviving children, which implies they contribute less to the overall rate of population growth than their less educated peers (Schultz 1994).

Education of women is also closely related to their labor force participation outside of their family, holding constant for their husband's income or education (Schultz 1981). Household models of production and consumption generally assume that family labor supply choices are jointly determined, and women's education is expected to affect not only lifetime fertility, but also the timing of marriage and the matching of the productive characteristics between husband and wife in the marriage 'market.' Within such a framework of family coordinated decision making, it is clearly difficult to assess how one of these jointly determined life-cycle choices made by women affects another of their choices, as stressed earlier. To decompose the variation in fertility into that which is

due to the timing of marriage, marital fertility rates, and non-marital fertility rates is a suggestive arithmetic exercise, but one that can misguide statistical analysis because the three components are likely to be stochastically interdependent. Consequently, when analyses of fertility differentials during the demographic transition focus only on marital fertility differentials, correlations can be causally understood only if marriage behavior and fertility demands are behaviorally independent of each other, a working assumption that few social scientists would adopt (Coale and Watkins 1986, Cleland and Wilson 1987).

8. Conclusions

Economic explanations for the fertility transition recognize the catalytic contribution of increasing returns to schooling, especially for women, which has encouraged women's education relative to men's, and fostered the rise in women's wages. These attributes of development are not equally evident in all countries. Declines in fertility have therefore proceeded quite unevenly, first across the high income countries, and more recently across the low income countries, but following a pattern of economic developments that has been widely quantified in household surveys and suggestively followed in cross-country comparisons. At an aggregate level, states have simultaneously legislated socialized support systems to provide for the consumption requirements of the elderly, and for the school and health investment requirements of the young, which may have also on balance, reduced the private demands of parents for children.

See also: Children, Value of; Fertility Control: Overview; Fertility: Institutional and Political Approaches; Fertility: Political and Political-Economic Perspectives; Fertility: Proximate Determinants; Fertility Theory: Caldwell's Theory of Intergenerational Wealth Flows; Fertility Theory: Embodied-capital Theory of Life History Evolution; Fertility Transition: Cultural Explanations; Gender and Feminist Studies in Economics; Land Rights and Gender

Bibliography

- Becker G S 1960 An economic analysis of fertility. *Demographic and Economic Change in Developed Countries*. Princeton University Press, Princeton, NJ
- Becker G S 1965 A theory of the allocation of time. *Economic Journal* 70: 493–517
- Becker G S 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA
- Caldwell J C, Orubuloye I O, Caldwell P 1992 Fertility decline in Africa: A new type of transition. *Population and Development Review* 18(2): 211–42
- Cleland J, Wilson C 1987 Demand theories of the fertility transition: An iconoclastic view. *Population Studies* 41(1): 5–30

- Coale A J, Watkins S C (eds.) 1986 *The Decline of Fertility in Europe*. Princeton University Press, Princeton, NJ
- Fogel R W 1999 Catching up with the economy. *American Economic Review* **89**(1): 1–21
- Galloway P R, Hammel E A, Lee R D 1994 Fertility decline in Russia 1875 to 1910: Pooled cross section time series. *Population Studies* **48**(1): 135–58
- Guinnane T W, Okun B S, Trussell J 1994 What do we know about the timing of the European fertility transition. *Demography* **31**(1): 1–20
- Haines M R 1979 *Fertility and Occupation: Population Patterns in Industrialization*. Academic Press, New York
- Heckman J J 1979 Sample selection bias as a specification error. *Econometrica*, **47**(1): 153–61
- Kuznets S 1966 *Modern Economic Growth: Rate Structure and Spread*. Yale University Press, New Haven, CT
- Manski C F 1995 *Identification Problems in the Social Sciences*. Harvard University Press, Cambridge, MA
- Mincer J 1963 Opportunity costs and income effects. In: Christ C, Friedman M, Goodman L, Griliches Z, Harberger A, Liviatan N, Mincer J, Mudlak Y, Nerlove M, Patinkin D, Telser L, Theil H (eds.) *Measurement in Economics*. Stanford University Press, Stanford, CA
- National Research Council 2000 *Beyond Eight Billion*. National Academy Press, Washington, DC
- Notestein F W 1945 Population—the long view. In: Schultz T W (ed.) *Food for the World*. University of Chicago Press, Chicago
- Richards T 1977 Fertility decline in Germany: An econometric appraisal. *Population Studies* **31**(3): 537–53
- Rosenzweig M R, Wolpin K I 1980 Testing the quantity-quality fertility model. *Econometrica* **48**: 227–40
- Schultz T P 1981 *Economics of Population*. Addison-Wesley, Reading, MA
- Schultz T P 1985 Changing world prices, women's wages and the fertility transition: Sweden, 1860–1910. *Journal of Political Economy* **93**(6): 1126–54
- Schultz T P 1994 Human capital, family planning and their effects on population control. *American Economic Review* **83**: 255–60
- Schultz T P 1995 *Investment in Women's Human Capital*. University of Chicago Press, Chicago
- Schultz T P 1997 The demand for children in low-income countries. In: Rosenzweig M R, Stark O (eds.) *Handbook of Population and Family Economics*. North Holland, Amsterdam, The Netherlands, Vol. 1A, Chap. 8
- Schultz T W (ed.) 1974 *Economics of the Family*. University of Chicago Press, Chicago
- Shapiro D, Tombashe B O 2000 *Women's Education, Employment and Fertility: Kinshasa in Transition*. Book manuscript, Pennsylvania State University, PA

T. P. Schultz

Fertility Transition: Latin America and the Caribbean

In the second half of the twentieth century Latin America experienced a remarkable demographic transformation, with fertility declining from a regional

average of more than six children per woman in 1950 to less than three children per woman by the year 2000. In this article we describe the nature of the decline, assess the role played by the proximate determinants of fertility, and discuss the ultimate causes and consequences of this profound transformation.

1. Fertility Levels and Trends

CELADE (1999) has published estimates and projections of the total fertility rate (TFR) for 20 countries in Latin America and the Caribbean for the period 1950–2000. These estimates update Chackiel and Schkolnik's (1996) excellent overview of the transition. The estimates are based on population censuses and vital statistics analyzed using direct and indirect techniques, and are generally consistent with alternative estimates based on survey data, such as the World Fertility Survey (WFS) and the Demographic and Health Surveys (DHS). Figure 1 plots the TFRs for a subset of 10 countries that represent the types of transitions observed in the region.

Back in 1950 most countries in Latin America and the Caribbean had TFRs in excess of six children per woman. Notable exceptions (the last four solid lines in Fig. 1) were Argentina and Uruguay, with rates around three; and Cuba and Chile, which hovered around four and five, respectively, but dropped below three in the mid- to late 1970s. Cuba then went on to have the lowest fertility in the region, 1.6 children per woman by the year 2000. A second group of 10 countries experienced declines from much higher fertility to cross the four-child level in the 1970s and 1980s. These include Panama, Brazil, El Salvador, and the Dominican Republic (the dashed lines in Fig. 1), as well as Colombia, Costa Rica, Ecuador, Mexico, Peru, and Venezuela (not shown, but their trajectories are bracketed at both ends by the countries shown). By the end of the twentieth century only six countries still had TFRs above three, but they all had experienced substantial fertility declines for 20 years or longer. This third group includes Guatemala and Paraguay (the top two solid lines in Fig. 1, representing the highest and lowest TFRs in this group in 2000), as well as Bolivia, Haiti, Honduras, and Nicaragua (not shown).

While there are substantial differences among countries, age-specific fertility has fallen more rapidly at older than at younger ages, as one would expect of a transition driven primarily by a desire to limit fertility rather than space births (Chackiel and Schkolnik 1996). This notion is confirmed by birth-interval analyses that show a substantial decline in parity progression ratios (the proportion of women proceeding from one parity to the next), particularly at high parities, with little if any change in average birth intervals (Juárez and Llera 1996).

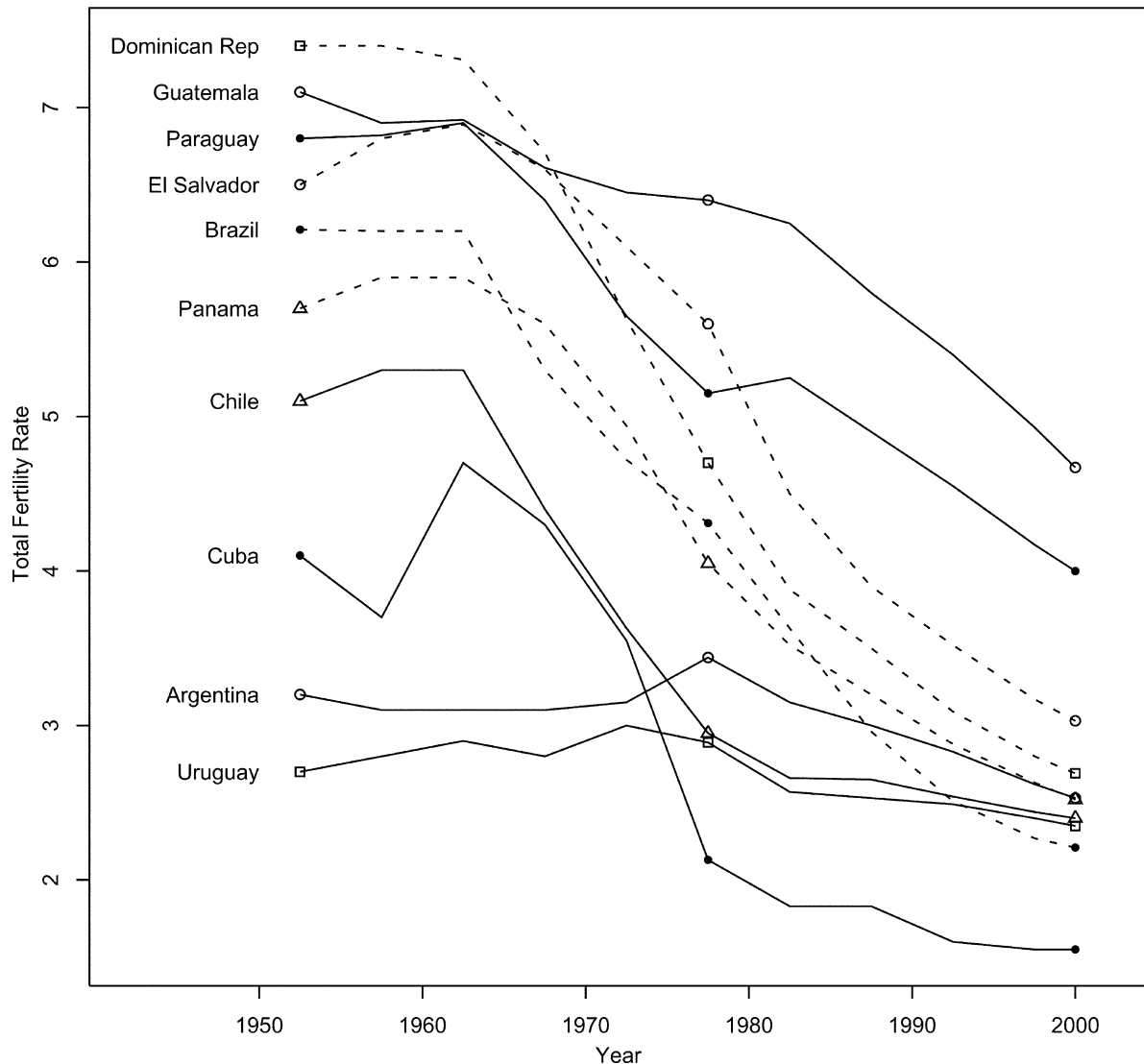


Figure 1
Total fertility rates, Latin America and the Caribbean, 1950–2000

The national averages conceal substantial variation across social groups in each country. Not surprisingly, fertility in Latin America has been substantially lower in urban than in rural areas; it decreases with increasing education of women, and with increasing status in occupation of the husband as one moves from agricultural workers through manual laborers to white-collar strata. The fertility decline started in the higher strata, but it soon spread to other groups, eventually reaching rural areas and women with no education married to agricultural workers. As a result, differentials increased at the start of the transition but

became narrower as fertility levels declined (Rodríguez 1996, Chackiel and Schkolnik 1996).

2. *The Proximate Determinants*

Following Davis and Blake's classic framework, changes in fertility can be traced to changes in intermediate variables affecting exposure to the risk of pregnancy, to conception, and to gestation. Most analyses focus on three key factors: marriage, contraception, and postpartum infecundity. A fourth important factor, abortion, is often omitted for lack of

reliable data. Moreno and Singh (1996), in a thorough analysis of these variables using WFS and DHS data, apply Bongaarts' model as well as an alternative model by Moreno to show that the bulk of the fertility decline in Latin America can be attributed to increases in contraceptive use, with changes in nuptiality and breastfeeding making only modest contributions.

Rosero-Bixby (1996) has examined the role of nuptiality in more detail. Using census and survey data he finds a slight increase in the time spent in union between 1950 and 1960 (consistent with isolated fertility increases in this period) followed by a flat trend. The proportion of consensual unions exceeds 50 percent in several countries, particularly in Central America and the Caribbean, and has increased over time. More remarkably, he finds no systematic trend in age at marriage. These results suggest that nuptiality trends could hardly contribute to the reduction in fertility in the region after 1960, and a more formal analysis shows only modest effects in selected countries.

Weinberger (1996) analyzed the role of contraception. She estimates a prevalence rate for the region during 1990 of 56 percent of couples where the wife is in the reproductive ages, with country-specific rates ranging from 7 percent in Haiti to around 70 percent in Costa Rica. The earliest national-level surveys measuring contraceptive practice go back to the 1970s and document a rapid and continuing spread of contraceptive use that is consistent with the observed decline in fertility. A distinctive characteristic of contraceptive use in the region is the high prevalence of female sterilization (20 percent of couples) and oral contraception (16 percent), which together account for about two-thirds of all contraceptive use. Other methods used by at least 5 percent of couples are periodic abstinence or rhythm, and intra-uterine devices. In contrast, male sterilization is rare (around 1 percent of users). In general, changes in the method mix over time have tended towards increased effectiveness, a trend largely due to increased reliance on female sterilization.

The role of induced abortion is hard to determine because of a dearth of reliable data. Frejka and Atkin (1996) argue convincingly that induced abortion is used widely in the region, but can provide only indirect evidence in most cases. Exceptions include Cuba, where abortion has been available on request since 1960, and Chile, where abortion is illegal but reasonable estimates are available from health statistics and surveys; both countries had total abortion rates in excess of two in the 1960s, coinciding with periods of steep fertility decline. (Interestingly, the early surveys in Chile helped motivate the medical profession to provide family planning services as an alternative to abortion.) Singh and Sedgh (1997) use indirect estimates based on hospitalizations for the treatment of complications related to induced abortions to show that, in addition to helping determine fertility levels,

abortion has played a significant role in the fertility decline in Brazil (where fertility in the 1990s would have been about 13 percent higher had abortion not increased from its level in the late 1970s), and more moderate roles in Colombia and Mexico.

Finally, postpartum infecundity appears to be the least important of the proximate determinants in terms of its role in the fertility transition in the region. The duration of breastfeeding in Latin America and the Caribbean is relatively short (with the exception of Haiti and Peru) and has not changed much over time.

3. Ultimate Causes and Consequences of the Decline

Traditional explanations of fertility decline have focused on socioeconomic changes that reduce the demand for children, but more recently there has been a resurgence of interest in theories that emphasize ideational changes and the diffusion of new ideas and forms of behavior. Several features of the Latin American transition suggest that the process represents more an innovation than adaptation to changing economic conditions.

Starting in the 1960s, Latin America underwent a process of modernization that resulted in remarkable improvements in educational levels, a shift away from agricultural work, greater participation of women in the labor force, and a doubling of per capita income between 1955 and 1980. Impressive as these achievements are, the changing socioeconomic composition of the population explains only a small part of the overall decline in fertility, which was driven largely by changes within social strata (Rodríguez 1996). Moreover, the economic situation changed drastically in the 1980s, as the region experienced a profound crisis (the 'lost decade') that eroded standards of living and reduced upward mobility; yet had little if any perceptible effect on the pace of the decline.

More direct evidence on the demand for children comes from data on fertility preferences. Starting in the 1960s, a number of surveys found preferences for small families in high-fertility regimes where women were not using contraception—the so-called 'KAP gap'—that evolved later into more sophisticated measures of unmet need for contraception and decompositions of fertility into its wanted and unwanted components (Westoff and Moreno 1996.) In countries like Costa Rica the transition seems to have occurred, at least initially, without a corresponding reduction in ideal family size (Rosero-Bixby 1999). Once the decline is established, however, preferences may well be reduced further. In Peru, for example, the TFR in 1977 to 1978 was 5.3 and its wanted component was only three. Ten years later the TFR had declined to four, but by then wanted fertility had further declined to two children per woman (Bongaarts and Lightbourne

1996). Still, back in the 1960s demand was already low and there was ample room for fertility reduction by better implementing existing preferences.

The gap persisted because the means to reduce fertility were not widely available at the time, but the situation was soon to change. Family planning programs were introduced in Latin America starting in the 1960s, largely as the result of private initiatives organized with help from the International Planned Parenthood Federation (IPPF) and, in some cases, the United States Agency for International Development (USAID). A leading private program is Profamilia, which was established in Colombia in 1965 and was a pioneer in the development of distribution networks and the use of radio to promote family planning. In the 1970s a number of governments started to provide services as well: a leader in this area was Mexico, where government clinics opened in 1974. Today most countries have a mixture of private- and public-sector programs, the latter often integrated with maternal and child health services. While estimates of the demographic impact of these programs vary, there is little doubt that they facilitated the transition by increasing the acceptability and availability of contraception (Mundigo 1996, Bongaarts et al 1990).

Once modern contraceptive methods became generally available, all that remained was for women to adopt them. It is here that diffusion or social contagion enters the picture, with friends and relatives playing a key role in terms of information flow and demonstration effects, at least in the early stages of the transition. A multivariate analysis of the spacing and limiting components of fertility shows that once the transition starts in a social stratum, it follows a broad but well-defined path that is consistent with a basic diffusion process (Rodríguez 1996). In a detailed case study of Costa Rica, quantitative spatial analysis combined with qualitative focus group research provides strong evidence of diffusion effects (Rosero-Bixby 1999). There is also a growing body of evidence that the mass media, particularly radio and television, can play a significant role. Perhaps the best documented example comes from northeast Brazil, where the fertility decline tracks the spread of television and the extremely popular *telenovelas*, nightly soap operas that often deal with life-styles, family values, sexuality, and reproductive behavior, and which can have strong if unintended influences on attitudes and values (Faria and Potter 1999).

Assessment of the consequences of the fertility transition is difficult because these changes occurred in times of economic crisis and turmoil (Potter 1996). There is, however, clear evidence that the decline resulted in lower infant mortality rates (Taucher 1996). Perhaps of more enduring consequence, the status of women has changed as a result of improved education and reduced fertility. In the words of a *New York Times* editorial on the subject, 'For all the machismo in Latin America, women there have gained sub-

stantially more rights and better treatment than women in the rest of the developing world.'

See also: Economic Development and Women; Family Size Preferences; Family Systems and the Preferred Sex of Children; Fertility: Institutional and Political Approaches; Fertility Transition: Cultural Explanations; Land Rights and Gender; Latin American Studies: Education; Latin American Studies: Gender; Latin American Studies: Religion; Poverty and Gender in Developing Nations; Reproductive Rights in Developing Nations; Rural Industrialization in Developing Nations and Gender

Bibliography

- Bongaarts J, Lightbourne R 1996 Wanted fertility in Latin America: Trends and differentials in seven countries. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK
- Bongaarts J, Mauldin W P, Phillips J F 1990 The demographic impact of family planning programs. *Studies in Family Planning* 21: 299–310
- Centro Latinoamericano de Demografía (CELADE) 1999 Latin America: Urban and rural population projections 1970–2025. *Boletín Demográfico/Demographic Bulletin*, No. 63. CELADE, Santiago, Chile
- Chackiel J, Schkolnik S 1996 Latin America: Overview of the fertility transition, 1950–1990. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK
- Faria V E, Potter J E 1999 Television, telenovelas and fertility change in northeast Brazil. In: Leete R (ed.) *Dynamics of Values in Fertility Change*. Oxford University Press, Oxford, UK
- Frejka T, Atkin L C 1996 The role of induced abortion in the fertility transition of Latin America. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK
- Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK
- Juárez F, Llera S 1996 The process of family formation during the fertility transition. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK
- Leete R (ed.) 1999 *Dynamics of Values in Fertility Change*. Oxford University Press, Oxford, UK
- Moreno L, Singh S 1996 Fertility decline and changes in the proximate determinants in the Latin America and Caribbean regions. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK
- Mundigo A I 1996 The role of family planning programmes in the fertility transition of Latin America. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK
- New York Times 1999 *Latin America's Birth Surprise*. Editorial, June 13, New York Times, New York

- Potter J E 1996 The social consequences of rapid fertility decline during a period of economic crisis. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK
- Rodríguez G 1996 The spacing and limiting components of the fertility transition in Latin America. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK
- Rosero-Bixby L 1996 Nuptiality trends and fertility transition in Latin America. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK
- Rosero-Bixby L 1999 Interaction, diffusion, and fertility transition in Costa Rica: Quantitative and qualitative evidence. In: Leete R (ed.) *Dynamics of Values in Fertility Change*. Oxford University Press, Oxford, UK
- Singh S, Sedgh G 1997 The relationship of abortion to trends in contraception and fertility in Brazil, Colombia and Mexico. *International Family Planning Perspectives* 23: 4–14
- Taucher E 1996 The impact of fertility decline on levels of infant mortality. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK
- Weinberger M B 1996 Changes in the mix of contraceptive methods during fertility decline: Latin America and the Caribbean. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK
- Westoff C F, Moreno L 1996 Reproductive intentions and fertility in Latin America. In: Guzmán J M, Singh S, Rodríguez G, Pantelides E A (eds.) *The Fertility Transition in Latin America*. Clarendon Press, Oxford, UK

G. Rodríguez

Fertility Transition: Middle East and North Africa

1. Introduction

Countries in the Middle East and North Africa (MENA) region provide a complex and rich experience on fertility transition. During the second half of the twentieth century, fertility trends took more

divergent courses than ever before. Several forces influenced fertility in conflicting directions. On the one side the forces of modernization, development, modern family planning technology and women's education, among others, have worked toward rapid fertility decline. On the other side, political, religious, cultural and institutional forces have supported early marriage and high procreation. The balance of these two sets of conflicting forces has, over time, influenced the onset and magnitude of fertility transition in the region.

2. Fertility Transition During the Period 1950–70

Throughout this period fertility was high in a group of 22 predominantly Arab and Muslim MENA countries (henceforth these countries will be referred to by the acronym AMENA). The total fertility rates (TFR) were five or more live births per woman (see Table 1 and Fig. 1) and the crude birth rates (CBR) were above 40 births per 1,000 persons. Some of these countries experienced a slight fertility increase. For instance, fertility increased from 5.74 to 6.36 in Lebanon and from 6.93 to 7.25 in Tunisia. Similar increases occurred in Egypt, Bahrain, Jordan and Libya. An odd country among this group is the former Soviet Union of Azerbaijan where fertility was relatively lower (TFR = 5.49; see Fig. 1).

A second group of four MENA countries experienced low and declining fertility ($3 \leq \text{TFR} \leq 4.9$; see Table 1 and Fig. 1). Between 1950 and 1970, fertility declined, albeit slowly, from 4.49 to 3.45 and from 3.0 to 2.61 in the Soviet republics of Armenia and Georgia, respectively. Cyprus experienced a similar slow decline from 3.71 to 2.79 during the same period. Also, fertility declined from 4.16 to 3.79 in Israel, which represents a special case because of its large immigrant population from Europe, the Arab States and Africa (Okun 1997, Goldscheider 1992). Henceforth, this group of countries will be referred to by the acronym (NAMENA), as a separate group in the MENA region.

Table 1
Distribution of MENA countries by categories of total fertility rates (TFR) 1950–2000

TFR	1950–55	1960–65	1970–75	1980–85	1990–95	1995–2000
≤ 7.0	11	15	9	6	2	2
6.0–6.9	9	6	9	4	2	0
5.0–5.9	2	1	2	5	3	3
4.0–4.9	2	1	2	4	6	3
3.0–3.9	2	2	2	3	7	7
2.2–2.9	0	1	2	3	4	7
≥ 2.1	0	0	0	0	2	4
Total	26	26	26	26	26	26

Source: UN Population Division. World Population 1950–2050 The 1998 Revision (Medium Variant)

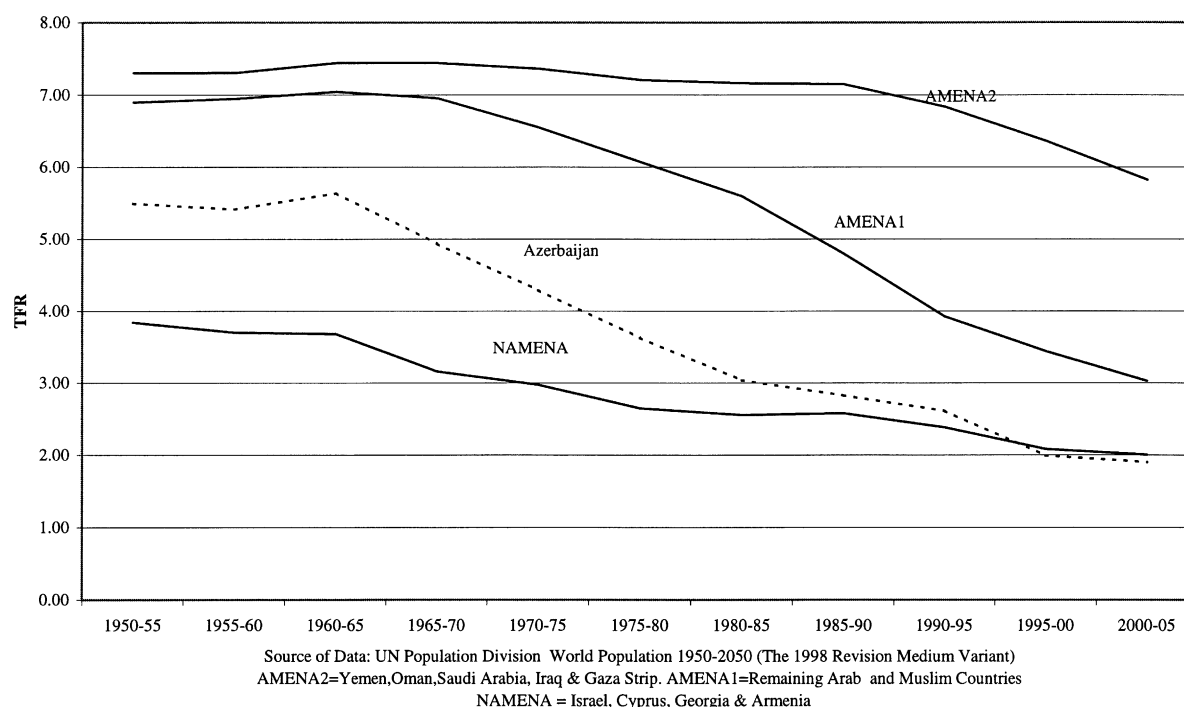


Figure 1
Average total fertility rates in the MENA countries

Fertility trends in the above mentioned two groups of MENA countries (AMENA and NAMENA) reflect considerable differences in reproductive behavior during this period. High fertility in the AMENA countries was due to political, religious, cultural, and institutional factors, and to the supremacy of patriarchal and extended family norms (Yousif et al. 1996, Courbage 1999). Moreover, public policies were against family planning interventions to regulate fertility (Obermeyer 1992). By contrast, culture and public policies in the NAMENA countries were in support of small family size, women's education and the use of modern contraceptive methods (Okun 1997).

3. Fertility Transition During the Period 1970–2000

The onset and magnitude of the fertility transition during this period varied considerably (see Table 1 and Fig. 1). The slow fertility decline mentioned before in Cyprus, Georgia, and Armenia continued unabated and reached replacement level ($TFR = 2.1$) during this period. However, the traditional high fertility in the AMENA countries gradually declined. Fertility continued to decline in Azerbaijan and it reached below replacement level by the year 2000 (see Fig. 1).

It also declined to 2.2–2.9 live births per woman in six countries (Lebanon, Tunisia, Turkey, Iran, Kuwait, and Bahrain) and to 3.0–4.9 in ten others (Jordan, Sudan, Syria, Algeria, Libya, Qatar, United Arab Emirates, Egypt, Morocco, and Western Sahara; see Fig. 1). Fertility remained high ($TFR \geq 5.0$) in only five AMENA countries (Yemen, Saudi Arabia, Oman, Iraq, and the Gaza Strip).

Tunisia, a pioneer country in the AMENA group, provides a good example of a sharp fertility decline from 6.21 in 1970 to 2.55 in 2000—a drop of 3.66 children in 30 years. The total fertility rate declined in Egypt from 5.6 in 1976 to 4.1 in 1991 and to 3.9 in 1992; a drop of 30 percent in 17 years. Kuwait provides a third example with a TFR decline from 5.9 in 1980 to 2.89 in 2000; a drop of three children in 20 years. Fertility declined in the Maghrib countries too, by 2.8 children in Algeria (1977–92) and 2.9 children in Morocco (1971–92).

By contrast, fertility remained high in Iraq, Saudi Arabia, Oman, and the Gaza Strip where the TFR ranged between 5.3 to 8 children per woman (see Fig. 1) and the CBR stayed above 40 throughout the second half of the twentieth century. A slight fertility decline finally occurred in a couple of these countries during the 1990s. Fertility declined from 6.8 to 5.8 in Saudi Arabia, and from 6.2 to 5.2 in Iraq. By contrast, fertility in Yemen and the Gaza Strip remained at a

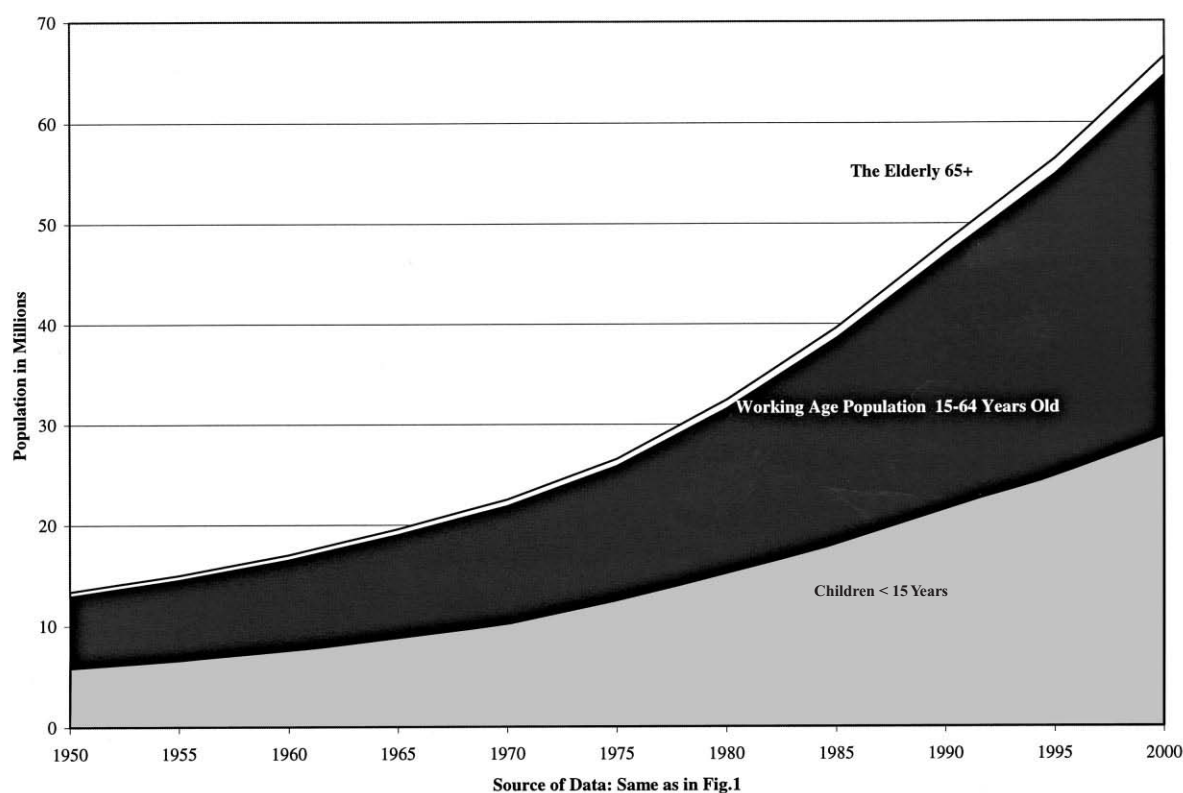


Figure 2
Population distribution by age in the AMENA 2 countries

high level of seven or more children per woman throughout the 1950–2000 period.

TFR declined to 3.6 in urban areas and it remained high at a level of 6.3 children in the rural areas of Egypt.

4. Urban vs. Rural Fertility

Place of residence (urban–rural) has a pervasive influence on fertility transition in the AMENA countries. It encapsulates many differences in opportunities available to women. Urban women have better access to the media and information technology, education, health, modern jobs, and family planning methods than their rural counterparts, most of whom are still under the grip of traditional beliefs and values (Yousif et al. 1996). The fertility decline is considerably higher and its onset earlier in urban than in rural areas. For instance the total fertility rate had dropped to below replacement level in Beirut by 1999, while in North Lebanon it averaged four live births (Kulczycki and Saxena 1999). In some AMENA countries a fertility decline occurred in the urban areas only. In Sudan, for example, fertility has declined to 3.7 in Khartoum and has remained high at above six children per woman in rural areas (Yousif et al. 1996). The

5. Women's Education

Women's educational attainment has contributed significantly to the fertility transition in the MENA countries. However, countries have differed from each other with regard to the level of education (or years of schooling) at which fertility declined. In several countries fertility declined in response to a few-years increase in women's education (Yousif et al. 1996). The rapid spread of women's education during the period 1970–2000 reduced the influence of traditional values, attitudes and norms and strengthened women's social status within households and communities. It facilitated their employment in modern activities outside the home as well. Consequently, the mean age at marriage for girls has increased to 25 or more years in several AMENA countries. Moreover, generations of educated girls have acquired modern aspirations and family desires.

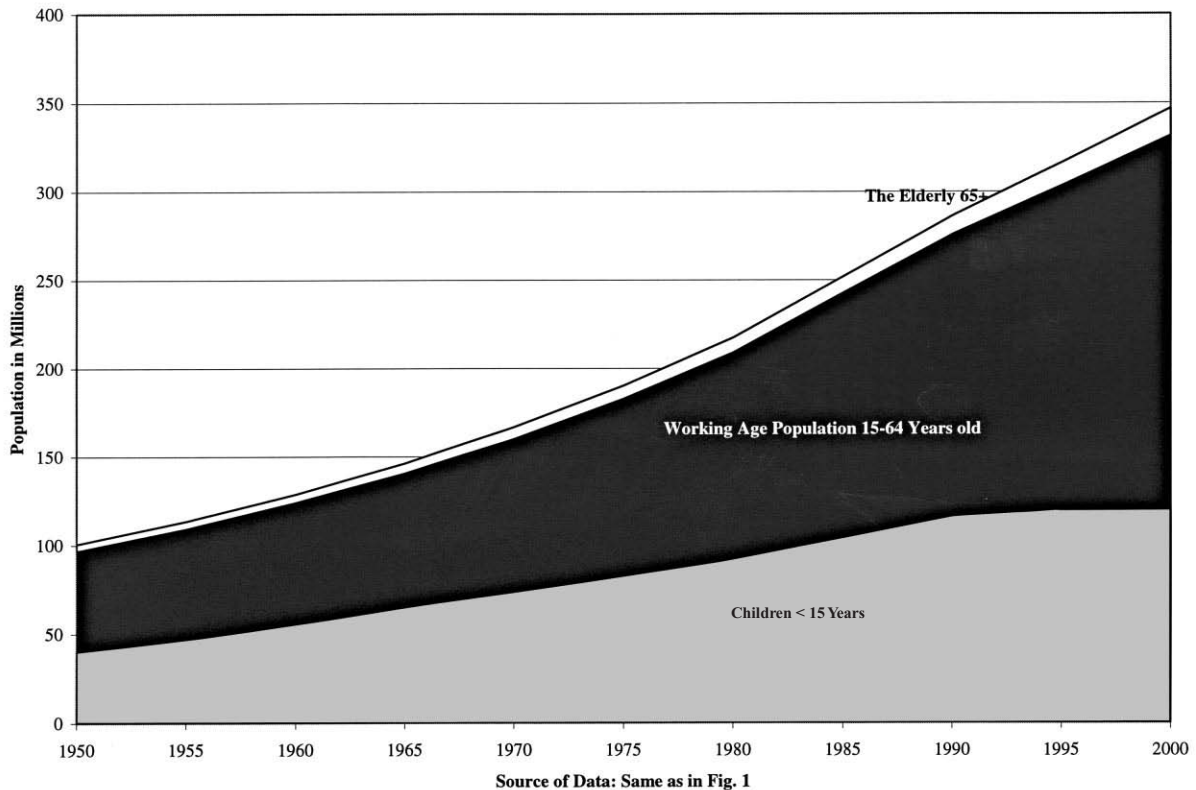


Figure 3
Population distribution by age in the AMENA 1 countries

6. Impact of the Fertility Transition on Age Composition

The impact of past high fertility, which often is associated with a high population momentum, on the age composition of the population is universal in the AMENA countries. This is reflected in increasing absolute and relative numbers of children below the age of 15 and in a precipitous increase in the working age population (15 to 64 years of age). The high population momentum is so pervasive that the impact of the recent fertility decline in these countries has not yet taken its full swing (see Figs. 2 and 3).

However, the impact is clear in the AMENA countries that have experienced fertility decline earlier than the others (see Fig. 3). In Azerbaijan and Tunisia, for example, the percentage of children below age 15 is declining, and it reached 30.3 percent and 28.6 percent, respectively, in 2000. In Lebanon and Kuwait this proportion is also declining (to 33.7 percent and 32.7 percent by 2000). In countries like Yemen, the Gaza Strip, Saudi Arabia, Iraq, Jordan, and Syria, children still make up more than 40 percent of the total population. However, in all AMENA countries, the

percentage of the population at working age is increasing, which implies a high demand for social services, particularly family planning, health, and education, as well as jobs.

By contrast, continuous fertility decline in the NAMENA has led to a relatively stable absolute and relative number of children below age 15 (see Fig. 4). The working age population is growing rapidly and its proportion of the total population exceeded 62 percent in all of these countries. The relative and absolute numbers of elderly have increased too and it has reached 12.6 percent in Georgia and 9.8 percent in Israel. This process will continue in the future, with profound implications for health and old age security schemes in these countries. These changes reflect an older age structure compared with the AMENA countries.

7. Impacts of Public Policies on the Fertility Transition

Public policies in the MENA countries provided the ideological and operational context for the fertility

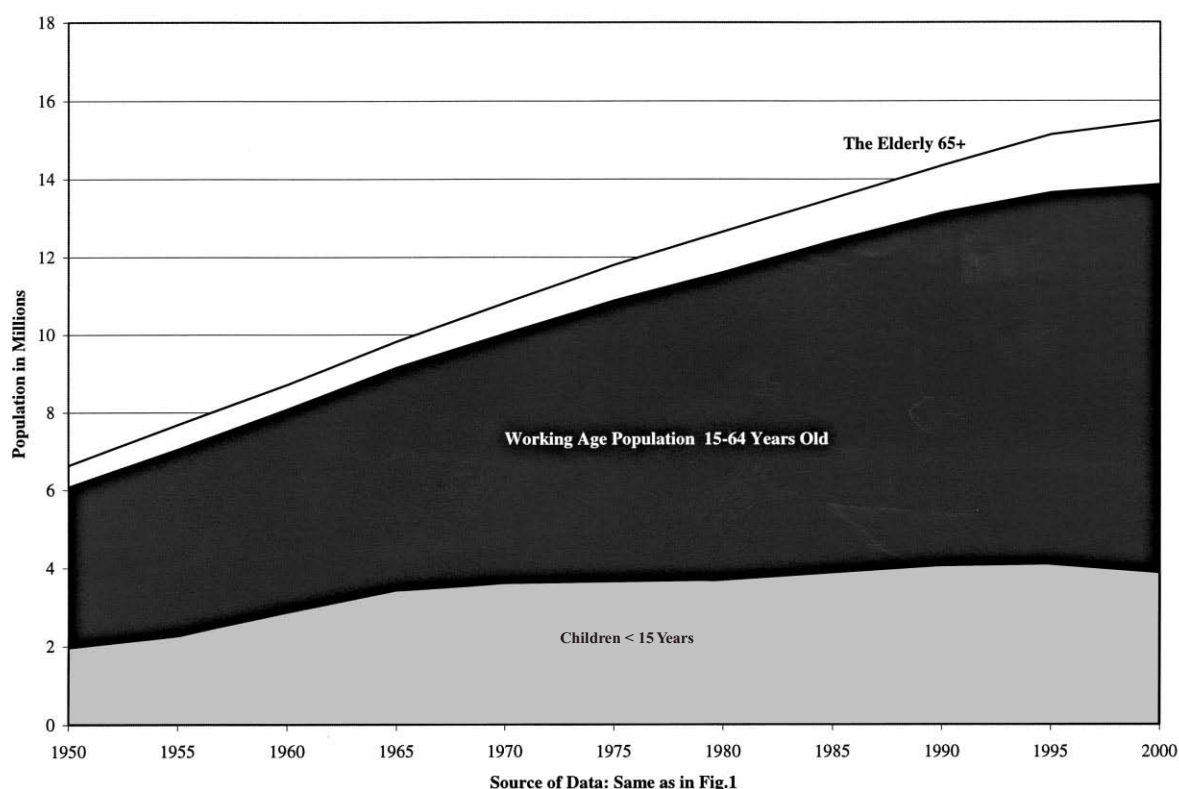


Figure 4
Population distribution by age in the NAMENA countries

transition. During the period 1950–70, population issues in the AMENA countries were seen as long-term problems of secondary importance, neither pressing nor urgent (Yousif et al. 1996). The reproductive role of women was regarded more important than their education and participation in modern employment. This public policy stand was evident during the International Population Conference in Bucharest in 1974. Delegates from the AMENA countries invariably argued in support of development, as opposed to family planning, as the best method for resolving population problems. Ten years later, by 1984, the political stand of these governments had changed, and in 1994 several AMENA countries reached an advanced stage on issues of population policy and development. Also, government support to population policies and programs gained importance, albeit at different times, in the AMENA countries.

Public policy in some countries changed direction several times. In Egypt, for instance, the government was against the use of modern contraceptive methods during the 1960s and 1970s. Starting in 1980, the government has changed its stand to support family

planning to regulate fertility and address population issues in the country. Its main challenge is in the rural areas, particularly in Upper Egypt, where fertility is highest and the potential for decline greatest.

Sudan and Iran provide other examples of countries with contradictory or vacillating public policies on population issues. The Sudan family planning program, which was progressing well during the 1970s and 1980s, has recently come to a standstill due to lack of government support. In Iran, the government's opposition to family planning after the Islamic revolution in 1979 has now changed completely to the opposite (Courbage 1999). In the Gulf States public policies on population issues and family planning are in their infancy.

8. Conclusions

Fertility transition mirrors divergences in culture and development of the MENA countries. Family values and traditions, social institutions, norms, ethnic divisions, religious beliefs, and all the various forces of

modernization and development are reflected in the fertility trends of the populations in this region. The downward trend in fertility in many countries and in urban areas is irreversible in view of ongoing social development, improving educational attainment, and increasing use of contraceptive methods. The potential for further fertility decline in the future is high, particularly in rural areas and in the AMENA countries where fertility is currently high.

See also: Fertility and Culture: Anthropological Insights; Fertility Control: Overview; Fertility: Institutional and Political Approaches; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations; Islam and Gender; Islam: Middle East; Middle East and North Africa: Sociocultural Aspects; Near Middle East/North African Studies: Economics; Near Middle East/North African Studies: Gender; Near Middle East/North African Studies: Religion; Near Middle East/North African Studies: Society and History

Bibliography

- Courbage Y 1999 Economic and political issues of fertility transition in the Arab world—answers and open questions. *Population and Environment* 20: 353–80
- Goldscheider C 1992 *The Embeddedness of the Arab–Jewish conflict in the State of Israel: Demographic and Sociological Perspectives*. PSTC Reprint Series, No. 92–02. Brown University, Population Studies and Training Center Providence, Rhode Island
- Kulczycki A, Saxena P C 1999 New evidence on fertility transition through wartime in Lebanon. *Genus* LV: 131–52
- Obermeyer C M 1992 Islam, women, and politics: The demography of Arab countries. *Population and Development Review* 18: 33–60, 205–7
- Okun B S 1997 Innovation and adaptation in fertility transition: Jewish immigrants to Israel from Muslim North Africa and the Middle East. *Population Studies* 51: 317–35
- Omran A R, Roudi F 1993 The Middle East population puzzle. *Population Bulletin* 48
- Yousif H M, Goujon A, Lutz W 1996 *Future Population and Education Trends in the Countries of North Africa, RR-96-11*. International Institute for Applied Systems Analysis, Laxenburg, Austria

H. M. Yousif

Fertility Transition: South Asia

South Asia is defined here as the countries of former British India (India, Pakistan, and Bangladesh) plus Nepal and Sri Lanka. Their combined population was estimated to be 1.34 billion in the year 2000, equivalent to 22 percent of the global total. Fertility trends in this

region are thus a crucially important determinant of future global population size. In this article the course of fertility over the past 50 years will be described, attempts to explain changes in reproductive behavior will be reviewed, and prospects for future fertility decline assessed.

1. Data Sources

Though only Sri Lanka possesses a vital registration system with high coverage of births, there is nevertheless a relative abundance of demographic data sources in South Asia, including decennial censuses and large-scale single-round demographic surveys. In terms of regular demographic monitoring, the outstanding achievement of the region is India's Sample Registration System (SRS). The SRS has yielded annual demographic estimates for most states in India since the early 1970s.

The collection of accurate demographic data in South Asia is extremely difficult and all national datasets are flawed to some extent. This has generated considerable controversy over the timing and magnitude of reproductive change in South Asia. Genuine uncertainty still exists over the details, particularly in Pakistan, where the quality of data is especially poor, but the broad contours of demographic change are well established.

2. Fertility Trends

The historic level of fertility in South Asia was probably similar to that of Europe and China, namely an average of 5.5 to 6.0 live births per woman, with fluctuations in the range of 4.5 to 6.5. Natality rose during the 1950s and 1960s in some parts of the region, a reflection of declines in widowhood and changes in breastfeeding patterns and sexual behavior within marriage. Fertility peaked at around seven births per woman in Bangladesh and Pakistan in the 1970s but remained about one birth lower in the other three countries.

Sri Lanka was the forerunner of fertility transition (see Fig. 1). The onset of decline dates back to the 1950s, and began largely in response to a rising age at marriage (Fernando 1975). The first appreciable fall in national fertility in India occurred in the 1970s, in Pakistan and Bangladesh in the 1980s, and in Nepal in the early 1990s. Comparing the highest post-1950 fertility level with estimated levels between 1995 and 2000, childbearing has fallen most steeply in Sri Lanka (63 percent), moderately in India and Bangladesh (about 50 percent), and least in Pakistan and Nepal (about 28 percent). Undoubtedly the most remarkable development has been the steep fall in fertility in Bangladesh. Between 1980 and 1995 fertility fell from 6.5 to 3.2 births per woman. This is among the most

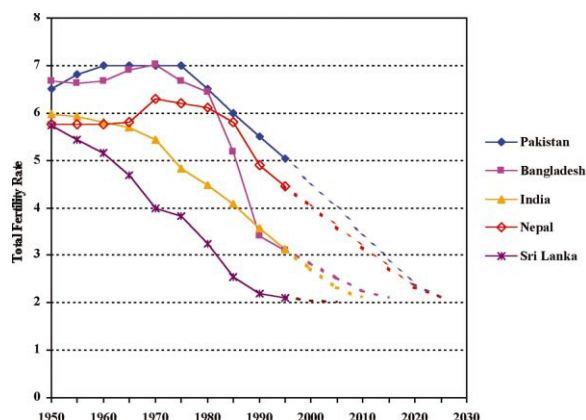


Figure 1
Past and projected trends in total fertility rates in South Asian countries
Source: United Nations 1999 *World Population Prospects: The 1998 Revision*. Population Division, New York

abrupt and pronounced changes in reproduction ever recorded in a developing nation.

These national trends mask geographical, ethnic, and socioeconomic variations in the timing of the onset and speed of reproductive change. In Bangladesh, fertility is lowest in the Rajshahi Division (in the north), characterized by relative poverty and landlessness, and highest in the Chittagong Division, generally considered to be the most conservative part of the country. In Nepal, fertility transition is more advanced in the Hill Zone than in the Terai (adjacent to India) or in the high-altitude Mountain Zone. The highest fertility in Pakistan is recorded in the least developed province, Baluchistan.

The diversity of reproductive trends in India is much greater (see Fig. 2). In the four northern states of Bihar, Madhya Pradesh, Rajasthan, and Uttar Pradesh, which comprise nearly 40 percent of the total population, demographic modernization has lagged far behind the rest of the country. Women's literacy, age at marriage, and life expectancy are all low; fertility has declined moderately but the level of childbearing was still above four births per woman in the mid-

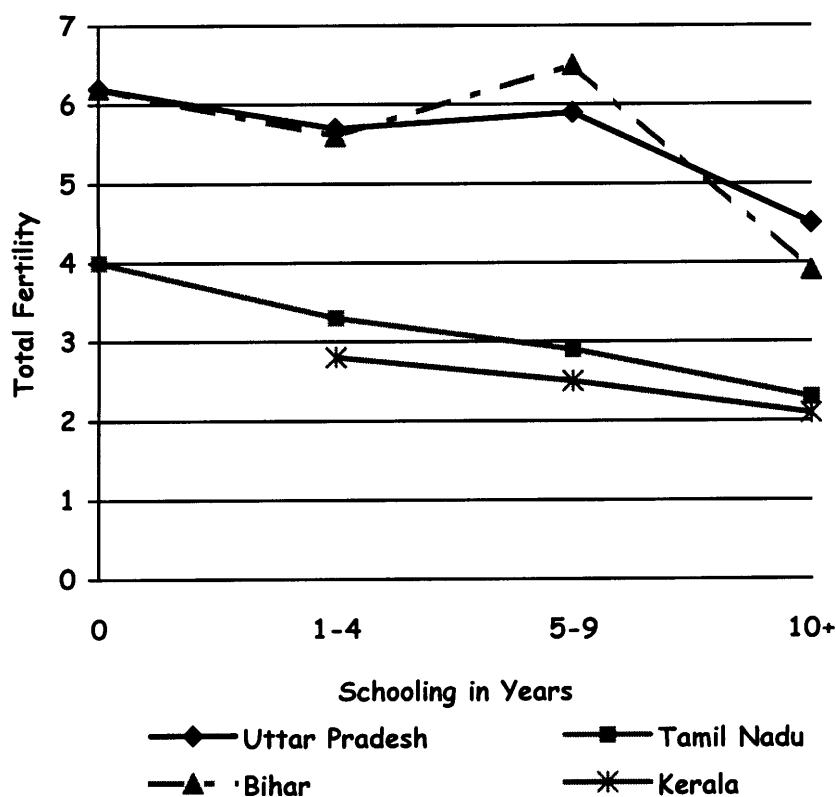


Figure 2
Rural total fertility rates by educational level of women, selected states of India, 1984
Note: Kerala estimate for 0 years of schooling not shown, because too few cases exist
Source: Office of the Registrar-General 1990 *Fertility Differentials in India*. Vital Statistics Division, New Delhi

1990s. By contrast, in the southern states of Kerala and Tamil Nadu fertility transition began early, probably in the late 1950s (Zachariah and Rajan 1997). According to the SRS the level of childbearing had already fallen to about four births per woman by the early 1970s. Kerala reached replacement level fertility (i.e., the level required to maintain stability of population size in the long term) of a little above two births by 1988, eventually stabilizing at 1.8 births per woman. Replacement level was reached in Tamil Nadu in 1993.

3. Determinants of Transition

3.1 Direct Causes

South Asian transitions to lower fertility have been driven by two direct causes: a rise in marriage ages, thus reducing teenage childbearing; and increased family size limitation by contraception and abortion. There is little evidence of postponement of live births following marriage, or of changes in birth spacing. Contraception and abortion have been used to limit family size but not to delay first births, nor for birth spacing. Reproduction is increasingly compressed within the age span of 20 to 30 years.

Throughout most of South Asia, early and universal marriage for women—with a wide gap of five to eight years in the ages of brides and grooms—has been the historic norm. Apart from Sri Lanka and parts of India (e.g., Kerala), the singulate mean age at marriage for women at the start of the twentieth century was about 13 years. (The singulate mean age is calculated from census or survey data on proportions single by current age: for details see Hajnal 1953 and also Ní Bhrolcháin.) Marriage has remained universal but women's age at marriage has steadily increased. By the 1980s, the singulate mean ages were: Bangladesh 16.6, India 18.3, Nepal 18.9, Pakistan 20.7, Sri Lanka 24.4.

Marriage postponement probably accounts for 20 to 30 percent of the decline in total fertility in Pakistan and Sri Lanka but rather less in the other three countries. However, delayed marriage is not consciously motivated by a desire for smaller family sizes. The fundamental cause is a change in the marriage market. In a rapidly growing population, each age group is larger than the preceding older one and, in societies where females traditionally marry older men, this creates a surplus of marriageable women. It thus becomes more difficult and expensive for parents to find suitable bridegrooms for their daughters. Increasing secondary schooling for girls accentuates this upward pressure on marriage ages.

The major direct cause of the South Asia fertility transition is increased use of contraception and abortion. Abortion is believed to be commonly practiced in all five countries but its precise contribution to fertility decline is impossible to gauge with any

confidence because in Nepal, Sri Lanka, and Pakistan it is illegal, and in Bangladesh and India most terminations are unrecorded.

In contrast, trends in contraceptive practice have been monitored reliably. A very pronounced link exists between reported contraception and fertility. In the 1990s, contraceptive prevalence (i.e., the percentage of married women in the reproductive age span who report current use of any contraceptive method) was highest in Sri Lanka (66 percent), and lowest in Pakistan (24 percent). Highly effective modern methods are dominant. In India and Nepal there is a heavy reliance on sterilization, mainly of women. In 1992 and 1993, sterilization accounted for 76 percent of contraceptive protection in India, and the median age of sterilization of women was 27 years. In contrast, the great surge in contraceptive practice in Bangladesh in the late 1980s and 1990s was dominated by increased use of hormonal methods.

3.2 Socioeconomic, Cultural, and Policy Influences

During the past 50 years, a fertility transition has spread across South Asia despite the variation in material, social, and cultural circumstances. Presumably some common underlying factor, or interrelated cluster of factors, accounts for the onset of transition over the same relatively short timescale. Such factors might include preceding mortality declines that have increased numbers of surviving children, or monetization of economies and growing aspirations for educating children and for material goods, all of which increase the cost of childbearing. The variations in the timing and speed of transition demand more context-specific explanations.

Population scientists disagree about the causes of change (see *Fertility Transition: Cultural Explanations* and *Fertility Transition: Economic Explanations*), and rather few widely-accepted generalizations can be made. One of the few is that the fertility transition in the region is not primarily a response to rising living standards. Bangladesh remains one of the poorest countries in the world and the southern states of India have income levels that are below the national average. Indeed, some commentators have proposed that reproductive change has been driven by poverty rather than prosperity, though the micro-level evidence does not support this view: poorer families have more children than richer families.

Progressive social policies, resulting in universal schooling, high life expectancy, relative equality of incomes, and state welfare provision, offer a compelling explanation for the early transitions in Sri Lanka and Kerala. But this thesis applies less obviously to Tamil Nadu, where literacy and life expectancy are not exceptionally high, and is implausible in the case of Bangladesh.

Gender inequality has been a major explanatory theme (see *Reproductive Rights in Developing Nations*).

In an influential paper, Dyson and Moore (1983) interpreted the north-south gradient in Indian fertility in terms of the kinship structure. North India, Pakistan, and the Indo-Aryan majority in Nepal are part of a wider West Asian kinship system dominated by patrilineage. Its key characteristics include early exogamous marriage for women, a joint family structure, and an emphasis on purdah, which result in very limited autonomy for young wives and considerable pressure to produce sons.

In contrast, southern India and Sri Lanka are part of a South and East Asian kinship system, where endogenous marriage and nuclear family are more prevalent, women's lives are less rigidly controlled, and preference for sons over daughters is less pronounced. According to Dyson and Moore this system, with its greater autonomy for women, is more amenable to demographic innovation, including the adoption of modern birth control. Bhat (1996) concluded that these elements of social structure and kinship account for some 30 to 50 percent of the north-south fertility difference in India. Possibly even more important is exposure to mass media, a result interpreted by Bhat as testimony to the powerful influence of Western ideas on reproductive behavior.

Considerable stress has also been laid on the role of education, particularly for women, as a facilitator of fertility transition. At the individual level the reproductive behavior of unschooled women and those who have received primary schooling differs little, whereas those with secondary schooling invariably have fewer children. Yet the contextual influence of schooling may be more important than the impact of individual attainment. Women in Kerala and Tamil Nadu have lower fertility than those in Bihar and Uttar Pradesh, regardless of their length of schooling (see Fig. 2). While it is widely accepted that high adult literacy has been conducive to reproductive change, it is also clearly not a precondition. In 1991, when Bangladesh was in the midst of rapid fertility transition, only 23 percent of adult women and 36 percent of men were literate.

The contribution of state family planning policies to the fertility transition is also controversial (see *Family Planning Programs: Development and Outcomes*). All five countries have population control policies with explicit demographic goals. All have offered financial incentives (or compensation payments) to individuals who agree to sterilization, and all have been criticized for coercive elements. In Sri Lanka, India, and Nepal, vigorous government promotion of family planning has probably accelerated fertility decline (see, for example, Caldwell et al. 1982). In Bangladesh it may have been even more important. A well-known experiment in the Matlab district demonstrated that the appropriate provision of services could reduce fertility (Phillips et al. 1988) and it has been claimed that, while a preceding mortality decline was the underlying cause of the fertility transition, the family planning program

was largely responsible for the timing and speed of change (Cleland et al. 1994). In Pakistan, the program has been largely ineffective because of weak political support.

4. Prospects for the Future

India's fertility is projected to reach replacement level by 2015, Bangladesh by 2020, and Nepal and Pakistan by 2030 (see Fig. 1). The prospects for the continuation of past downward trends are good. Recent surveys indicate that at least 20 percent of births are unwanted in Pakistan, Nepal, and the high-fertility northern states of India. However, a desire for two sons is still prevalent in these parts of South Asia. Acting on this preference implies a fertility rate of over four births per woman unless sex-selective abortion becomes widespread. Similar strong preferences for sons exist in East Asia, but the desire for two sons has subsided into a desire for one (see *Fertility Transition: China*; *Fertility Transition: East Asia*). The most likely prognosis is that the remaining high-fertility parts of South Asia will follow this example.

See also: Family Size Preferences; Fertility: Institutional and Political Approaches; Fertility Transition: Cultural Explanations; Fertility Transition: East Asia; Fertility Transition: Economic Explanations; Reproductive Rights in Developing Nations

Bibliography

- Bhat P N M 1996 Contours of fertility decline in India: A district level study based on the 1991 Census. In: Srinivasan K (ed.) *Population Policy and Reproductive Health*. Hindustan Publishing Corporation (India), New Delhi, India, pp. 97-177
- Bose A, Premi A K (eds.) 1992 *Population Transition in South Asia*. B. R. Publishing Corporation, New Delhi, India
- Cain M T 1983 Fertility as an adjustment to risk. *Population and Development Review* 9: 688-702
- Caldwell J C, Reddy P H, Caldwell P 1982 The causes of demographic change in rural south India. *Population and Development Review* 8: 689-727
- Cleland J, Phillips J F, Amin S, Kamal G M 1994 *The Determinants of Reproductive Change in Bangladesh: Success in a Challenging Environment*. World Bank Regional and Sectional Studies, Washington, DC
- Dyson T, Moore M 1983 On kinship structure, female autonomy, and demographic behavior in India. *Population and Development Review* 9: 35-60
- Fernando D F S 1975 Changing nuptiality patterns in Sri Lanka 1901-1971. *Population Studies* 29: 179-90
- Hajnal J 1953 Age at marriage and proportion marrying. *Population Studies* 7: 110-36
- Jeffery R, Basu A M (eds.) 1996 *Girls Schooling, Women's Autonomy and Fertility Change in South Asia*. Sage, New Delhi, India
- Leete R, Alam I (eds.) 1993 *The Revolution in Asian Fertility*. Clarendon Press, Oxford, UK

- Phillips J F, Simmons R, Koenig M A, Chakaborty J 1988 The determinants of reproductive change in a traditional society: Evidence from Matlab, Bangladesh. *Studies in Family Planning* **19**: 313–34
- Vlassoff M, Vlassoff C 1980 Old age security and the utility of children in rural India. *Population Studies* **34**: 487–99
- Zachariah K C, Rajan S I (eds.) 1997 *Kerala's Demographic Transition: Determinants and Consequences*. Sage, New Delhi, India

J. Cleland

Fertility Transition: Southeast Asia

The population of Southeast Asia was only one third the size of Europe's in 1950 (182 million compared to 547 million in Europe), but by 2050, the population of Southeast Asia is projected to be 25 percent larger

than Europe's (786 million vs. 628 million, United Nations 1999, pp. 442–3). This dramatic reversal is the product of a lag, of about a century, in the demographic transitions of the two regions. Fertility began to decline in several northwestern Europe countries during the last quarter of the nineteenth century and spread to all of Europe over the first half of the twentieth century. Fertility in Southeast Asia first showed signs of decline in a few countries in the late 1960s, and then moved downward, at varying speeds, throughout the region from the 1970s to the 1990s. Although fertility in Southeast Asia is likely to approach the replacement level (around two births per woman) early in the twenty-first century, the force of demographic momentum and the power of compound rates of growth will continue to expand the population of Southeast Asia relative to Europe (and North America).

Southeast Asia consists of the 11 countries that lie between the Indian subcontinent and China (see Fig. 1). While there are common geographical, historical,



Figure 1
Countries of Southeast Asia

Table 1

Estimates of population, life expectancy, and fertility in Southeast Asia

Southeast Asian Countries	POP (in 000) 2000 estimate	Life Expectancy		Total Fertility Rate	
		1965–70	1995–00	1965–70	1995–00
Brunei Darussalam	328	64.9	75.5	5.9	2.8
Cambodia	11,168	45.4	53.4	6.2	4.6
East Timor	885	37.5	47.5	6.2	4.4
Indonesia	212,107	46.0	65.1	5.6	2.6
Laos	5,433	40.4	53.2	6.2	5.8
Malaysia	22,244	59.4	72.0	5.9	3.2
Myanmar	45,611	47.4	60.1	6.0	2.4
Philippines	75,967	56.2	68.3	6.0	3.6
Singapore	3,567	67.9	77.1	3.5	1.7
Thailand	61,399	56.7	68.8	6.1	1.7
Vietnam	79,832	47.9	67.4	6.0	2.6
Total	518,540	47.8	65.7	5.8	2.7

Source: United Nations. 1999. *World Population Prospects: The 1998 Revision*. Volume 1: Comprehensive Tables. United Nations, New York

and cultural features of Southeast Asia, diversity is the hallmark of the region (Reid 1988, 1993, Wolters 1999). Indonesia, one of the world's most populous countries, comprises about 40 percent of population of the region as a whole (see Table 1). The other largest countries are the Philippines and Vietnam, each with estimated populations of more than 75 million in 2000, and Thailand at approximately 60 million. At the other extreme are the microstates of Brunei and East Timor, each with less than a million people. Singapore, a city-state, has a population of only 3.5 million, and Laos is only slightly larger at 5 million (Table 1). Myanmar (Burma) at 45 million, Malaysia with 22 million, and Cambodia with a population of 11 million are the other countries in the region.

The major geographical and cultural divisions are between mainland and insular Southeast Asia and between the lowlands and uplands within each region (and country). The populations of the mainland countries generally practice Buddhism while Islam is the major religion of Indonesia, and Christianity is the predominant religion of the Philippines. While part of Malaysia (Peninsular Malaysia) is on the mainland, it is usually considered as belonging to insular Southeast Asia because the Malay population of Malaysia shares a common language, religion, and cultural heritage with much of the Indonesian population. The lowland countryside of both mainland and insular Southeast Asia consists of villages knitted together by small and medium size market towns. Wet (irrigated) rice agriculture is the predominant feature of the countryside, but rubber, oil palm, coconut, and sugar plantations (and small holdings) as well as mining remain as legacies of the colonial export economies that once dominated the region. Every country has remote highland and mountainous regions, which are often populated by cultural and religious minorities (a

minority in one country is sometimes the dominant cultural group in neighboring countries).

Located along the major sea route between the great civilizations of East Asia and South Asia, the 'plural society' has been one of the defining features of Southeast Asia, which has been relatively open and absorptive of peoples, ideas, and cultural practices from elsewhere. In the twentieth century, assimilation into Southeast Asian societies and interethnic relations became more difficult with the creation of 'national' states and the translation of cultural differences by religion, language, and region into political categories with differential status and rights. Throughout the region, the 'overseas Chinese' continue to be over-represented in commercial and professional occupations. Over the centuries, a good share of Chinese-origin peoples have intermarried and blended with local populations in many Southeast Asian countries, but there remains a segment of the Chinese population that has been residentially and socially segregated even after three or four generations of local residence (Cushman and Wang 1988, Reid 1996, Chaps. 1–3).

The new Southeast Asian economies are dominated by the major metropolitan areas, including Jakarta, Surabaya, Bangkok, Singapore, Manila, Rangoon, Kuala Lumpur, and Ho Chi Minh City. These major cities typically originated as trading and administrative areas along major rivers or coastal ports. At the present time, these cities are bustling metropolises with large-scale manufacturing, commercial, service, and governmental functions. Contemporary indicators of development in Southeast Asia, including very low levels of mortality and almost universal secondary schooling, are approaching the prevailing standards of developed countries. The economic crisis that hit the region in 1997 notwithstanding, the

prevailing trajectory of Southeast Asia appears to be heading along the lines of the modernizing East Asian countries of Japan, Korea, and Taiwan.

1. A Demographic Overview

The major demographic attributes and trends of Southeast Asia, based on the latest UN population estimates and projections, are reported in Table 1. The extraordinary pace of demographic and social change in the region is registered in the measures of life expectancy and the total fertility rate (a summary measure of annual fertility expressed in terms of the average number of births per woman) from the late 1960s to the late 1990s. Over this 30-year period, the average life expectancy in Southeast Asia increased from 48 years to 66 years and fertility declined from almost six births per woman to less than three. These rates of change are unprecedented in the historical record and have only been exceeded in the modern era by several East Asian countries. In the late 1960s, the only countries in the region with life expectancy above 60 were Singapore and Brunei, although Malaysia was not far behind. Thirty years later, every country had life expectancies above 60, except for Cambodia, East Timor, and Laos (the poorest countries in the region). The declines in fertility are no less pervasive. Even in the relatively poor countries of Myanmar (Burma) and Vietnam, fertility has dropped below three children per couple. Fertility is below the replacement level of two births per couple in Singapore and Thailand.

The broad picture of rapid fertility decline conceals significant variations from country to country. For example, there has been very little socioeconomic development and only modest demographic change in Laos and Cambodia, countries that were victims of the Indochina wars of the 1960s and 1970s. After the 'victory' of the Khmer Rouge in 1974, the mortality rate may have doubled as a result of executions and starvation during the Pol Pot regime (United Nations 1999, p. 124, Heuveline 1998). There are signs of fertility declines in the UN data for Laos, Cambodia, and East Timor, but the absolute levels of fertility remain very high in these countries. Much further along the path of fertility decline are the large countries of Vietnam and Myanmar, although economic conditions in these countries remain far below the rest of the region (Haugton 1997, Myint 1991).

Fertility dropped below the replacement level in Singapore in the late 1970s (Saw 1999). Perhaps the most celebrated case of rapid fertility decline in the region is Thailand, where fertility declined from more than six births per woman in the 1960s to less than two births per woman in less than 30 years (Knodel et al. 1996, Hirschman et al. 1994). Fertility declines have also been registered in other Southeast Asian countries with estimates of total fertility rates ranging between

three and four births in Malaysia, the Philippines, and Indonesia (Hirschman and Guest 1990a, Palmore et al. 1995). The case of falling fertility in Indonesia is particularly salient, given the size of the country and its major influence on the region as a whole.

2. Explanations of Fertility Decline in Southeast Asian Countries

The postponement of marriage has contributed modestly to the temporal decline in fertility in several Southeast Asian countries, but the major component has been declines in marital fertility (Retherford and Cho 1973, Hirschman and Guest 1990a). Declines in marital fertility are due primarily to increased contraceptive use. Indeed other factors, such as decreases in spouse mortality and reduced durations in breastfeeding, would probably have led to rising marital fertility in the absence of increased contraceptive practice. The underlying question is what has caused the increase in contraceptive use.

There are two widespread interpretations for the widespread adoption of contraception and rapid fertility declines in the region: the efficacy of family planning programs and the rapid pace of socioeconomic change in the region. These interpretations are not incompatible, and in their classic study, John Knodel and colleagues conclude that Thailand's reproductive revolution is due to the combination of four factors: rapid social and economic development, a favorable cultural setting (female autonomy and weak intergenerational influences in reproductive decisions), a latent demand for fertility control, and an effective national family planning program (Knodel et al. 1987). The challenge for the field has been to test the relative significance of these explanatory factors in different times and places. At present, the research record points to a broad mosaic of varied patterns and causes without a simple overarching interpretation.

The fundamental prerequisite for fertility transitions is a mortality transition, especially declines in infant and child mortality. In Southeast Asia, mortality declines began in the decade after World War II, as antibiotics, other aspects of modern medicine, and preventive health programs were introduced and spreading in many Southeast Asian countries. The continuing pace of mortality declines in the region is shown in Table 1. According to demographic transition theory, declines in mortality and other aspects of socioeconomic change are predicted to increase pressures on families and married couples to lower their desired number of children and to increase their 'demand' for family planning.

In almost every Southeast Asian country, there have been empirical studies that find the predicated associations between modern characteristics and lower levels of fertility. Social change, or modernization, is usually indexed by a variety of individual level characteristics,

such as female education and employment, urban residence and nonagricultural employment, exposure to the mass media, and ownership of modern goods. Such associations, however, do not explain how macrochanges (in society as a whole) lead to individual-level modernization nor whether cross-sectional associations reflect actual historical (over time) patterns of social change. Hirschman and colleagues (Hirschman and Guest 1990b, Hirschman et al. 1998) have attempted to explain fertility declines in four Southeast Asian countries (Indonesia, Philippines, Malaysia, and Thailand) from 1970 to 1990 in terms of changes in provincial level measures of social change in the status of women, the economic roles of children, and infant mortality. Although based on fertility transitions still in process, the findings are consistent with the logic of demographic transition theory, albeit with wide variations in the significance of specific independent variables over time and across countries.

Impressionistic evidence would suggest that governmentally sponsored family planning programs have been the major reason for rising contraceptive use in Thailand, Indonesia, and other Southeast Asian countries with declining fertility. Family planning programs were instituted in many Southeast Asian countries in the 1970s, just about the same time that fertility began to edge downward in the region. Several authors have attributed the decline of fertility in Indonesia to a very successful family planning program there (Freedman et al. 1981, Hull et al. 1977, Ross and Poedjastoerti 1983), and some observers attribute the slower declines in fertility in Malaysia and the Philippines to relatively less energetic family planning programs in these countries. These conjectures, however, are very difficult to evaluate empirically because the emergence and characteristics of family planning programs are correlated with other determinants of fertility decline, most notably socioeconomic development.

One analytical approach to this question is to measure the spatial pattern of fertility decline relative to the geographical location of family planning clinics and aspects of the program that may provide family planning information or contraceptives. The problem with this approach is that the location of family planning clinics may be endogenous, i.e., program resources such as clinics are directed to populations with very high fertility (just as police stations are often located in high crime areas). Moreover, contraceptives are available from nongovernmental outlets such as private clinics and drugstores. These conditions mean that empirical tests of the impact of family planning programs on fertility decline are extraordinarily difficult to estimate empirically.

Using econometric techniques to control for the endogenous role of family planning inputs, Gertler and Molyneaux (1994) find that almost all of the fertility decline in Indonesia from 1982 to 1987 was due to rising incomes and education and that very little

of the fertility decline could be attributed directly to family planning efforts. These findings do not rule out the interpretation that the Indonesian family planning programs played an important role through the provision of readily available means for fertility control (and at a modest cost) for women who were already motivated to limit their number of births, which is what Robinson and Rachapaetayakom (1993) argue was the major role of the Thai family planning program. Family planning programs may play a more influential role in the origins of a fertility transition than on its continuation.

Govindasamy and DaVanzo (1992) conclude that public policies, other than family planning programs, have been more influential in promoting more rapid fertility declines among the Chinese and Indian populations than among the Malay population of Malaysia during the 1970s and 1980s. The Malaysian government provided educational and economic subsidies (as part of an affirmative action program) for Malays that reduced the costs of children in a rapidly modernizing society. Leete (1996) adds the additional interpretation that Malay fertility has remained high because of the resurgence of Islam and pronatalist cultural values. There are, however, quite wide variations in levels and trends of Malay fertility between Singapore and Malaysia and across states in Malaysia that raise questions about a simple cultural or religious preference for high fertility (Hirschman 1986, Jones 1990).

Perhaps the single most important cultural characteristic of Southeast Asia is the relatively high status of women, especially when compared with East Asia and South Asia (Reid 1988, pp. 162–72, Van Esterik 1982). While women still face many socioeconomic obstacles in Southeast Asia, the situation is quite different from the patriarchal societies of East and South Asia. Southeast Asian kinship systems are typically bilateral with equal importance attached to the husband's and wife's families. For the most part, there is no strong sex preference for children in Southeast Asia (Wongboonsin and Ruffolo 1995) nor are there strong prescriptions on residence with the groom's family after marriage. Cultural variables are difficult to quantify, especially at the microlevel, which makes it difficult to estimate their relative role in shaping fertility levels and change, but it seems likely that greater freedom for women to pursue nonfamilial roles and to influence household decision making have been important factors in the region's rapid fertility decline.

3. The End of the Fertility Transition in Southeast Asia

If the current pace of decline continues, replacement-level fertility (two children per woman) should be reached early in the twenty-first century in most

countries and regions of Southeast Asia. Indeed, if the recent history of East Asia, and Singapore, is any guide, modern fertility levels may not stop at the replacement level, but continue to decline below the level of fertility needed to replace the parental generation. The explanations for these rapid fertility declines, although still hotly debated, are likely to be the complex interplay of rapid socioeconomic change, organized family planning programs, and spatial and social diffusion processes that vary in significance and intensity over time.

An additional factor that is likely to accelerate the decline of fertility is the postponement and perhaps decline in marriage. The most celebrated case is Singapore, where the 'the flight from marriage' and declining fertility among the highly educated population prompted the government to organize social activities to encourage unmarried persons in their twenties and thirties to marry and have babies (Leete 1994, p. 815, Yap 1995). The trend toward marital postponement throughout the region has been underway for 30 years, but the 1990 round of censuses revealed how far the trend had progressed. In Malaysia, Thailand, and the Philippines, from 13 to 15 percent of women in their early thirties remained unmarried (Jones 1995, 1997).

The figures for marital postponement are even higher in cities and among those with higher education. In several countries of Southeast Asia, the average age at marriage of women and men is older than in many Western countries. Since the historical pattern in Southeast Asia was of young and universal marriage, this is a remarkable transformation in custom. Perhaps, there may be a lag in the adjustment of social institutions to replace traditional match-making arrangements. Another hypothesis is that some highly educated men in Southeast Asia may prefer to marry women with lower educational attainment who are considered to more likely to follow traditional gender roles than highly educated women.

Although there is likely to be an adjustment in marriage patterns before high levels of celibacy are reached, the current pattern of marital postponement is certain to likely to accelerate the downward trend in fertility in most Southeast Asian countries. Indeed, the fertility transitions in Southeast Asia are among the most rapid in the developing world, and the trend may accelerate. Nonetheless, rates of rapid population growth of 2 percent per annum during much of the second half of the twentieth century and subsequent demographic momentum will lead to continued rapid growth of the population of Southeast Asia during the first half of the twenty-first century.

See also: Family Size Preferences; Family Systems and the Preferred Sex of Children; Fertility Control: Overview; Fertility Transition: Cultural Explanations; Fertility Transition: East Asia; Fertility Transition: Economic Explanations; Land Rights and Gender;

Southeast Asia: Sociocultural Aspects; Southeast Asian Studies: Gender; Southeast Asian Studies: Health

Bibliography

- Cushman J, Wang G (eds.) 1988 *Changing Identities of the Southeast Asian Chinese Since World War II*. Hong Kong University Press, Hong Kong
- Freedman R, Khoo S-E, Supraptilah B 1981 Use of modern contraceptives in Indonesia: a challenge to the conventional wisdom. *International Family Planning Perspectives* 7 (March): 3-15
- Gertler P J, Molyneux J W 1994 How economic development and family planning programs combined to reduce Indonesian fertility. *Demography* 31: 33-63
- Govindasamy P, DaVanzo J 1992 Ethnicity and fertility differentials in peninsular Malaysia: do policies matter? *Population and Development Review* 18: 243-67
- Haughton J 1997 Falling fertility in Vietnam. *Population Studies* 5: 203-11
- Heuveline P 1998 Between one and three million: toward the demographic reconstruction of a decade of Cambodian history (1970-1979). *Population Studies* 52: 49-65
- Hirschman C 1986 The recent rise in Malay Fertility: a new trend or a temporary lull in a fertility transition?. *Demography* 23 (May): 161-84
- Hirschman C, Guest P 1990a The Emerging Demographic Transitions of Southeast Asia. *Population and Development Review* 16: 121-52
- Hirschman C, Guest P 1990b Multilevel models of fertility determination in four Southeast Asian countries: 1970 and 1980. *Demography* 27: 369-96
- Hirschman C, Tan J-E, Chamratrithirong A, Guest P 1994 The path to below replacement-level fertility in Thailand. *International Family Planning Perspectives* 20: 82-7, 107
- Hirschman C, Tan J-E, Chamratrithirong A, Guest P 1998 Explaining the rapid fertility decline in Thailand. Unpublished paper, Center for Studies in Demography and Ecology, University of Washington, Seattle, WA
- Hull T H, Hull V J, Singarimbun M 1977 Indonesia's family planning story: success and challenge. *Population Bulletin* 32(6): 1-51
- Jones G W 1990 Fertility transitions among Malay populations of Southeast Asia: puzzles of interpretation. *Population and Development Review* 16: 507-37
- Jones G W 1995 Population and the family in Southeast Asia. *Journal of Southeast Asian Studies* 26: 184-95
- Jones G W 1997 The demise of universal marriage in East and South-East Asia. In: Jones G W, Douglas R M, Caldwell J C, D'Souza R M (eds.) *The Continuing Demographic Transition*. Clarendon Press, Oxford, UK pp. 51-79
- Knodel J, Chamratrithirong A, Debavalya N 1987 *Thailand's Reproductive Revolution: Rapid Fertility Decline in a Third-World Setting*. University of Wisconsin Press, Madison, WI
- Knodel J, Ruffolo V P, Ratanalangarn P, Wongboonsin K 1996 Reproductive preferences and fertility trends in post-transition Thailand. *Studies in Family Planning* 27: 307-18
- Leete R 1994 The continuing flight from marriage and parenthood among the overseas Chinese in East and Southeast Asia. *Population and Development Review* 20: 811-29
- Leete R 1996 *Malaysia's Demographic Transition: Rapid Development, Culture, and Politics*. Oxford University Press, Kuala Lumpur

- Myint N 1991 Recent levels and trends of fertility and mortality in Myanmar. *Asia-Pacific Population Journal* 6: 3–20
- Palmore J A, Made Mamas S G, Arifiyanto Y 1995 Fertility decline in Indonesia, 1971–1991. *Journal of Population (Jakarta, Indonesia)* 1: 45–66
- Reid A 1988 *Southeast Asia in the Age of Commerce 1450–1680. Vol. 1. The Lands Below the Winds*. Yale University Press, New Haven, CT
- Reid A 1993 *Southeast Asia in the Age of Commerce, 1450–1680. Vol. 2. Expansion and Crisis*. Yale University Press, New Haven, CT
- Reid A (ed.) 1996 *Sojourners and Settlers: Histories of Southeast Asia and the Chinese*. Allen and Unwin, St. Leonards, Australia
- Retherford R, Cho L-J 1973 Comparative analysis of recent fertility trends in East Asia. In: *International Population Conference*. International Union for the Scientific Study of Population, Liege, Vol. 2, pp. 163–81
- Robinson W C, Rachapaetayakom J 1993 The role of government planning in Thailand's fertility decline. In: Leete R, Alam I (eds.) *The Revolution in Asian Fertility: Dimensions, Causes, and Implications*. Clarendon Press, Oxford, UK pp. 54–66
- Ross J, Poedjastoerti S 1983 Contraceptive use and program development: new information from Indonesia. *International Family Planning Perspectives* 9: 68–77
- Saw S H 1999 *The Population of Singapore*. Singapore Institute of Southeast Asian Studies, Singapore
- United Nations 1999 *World Population Prospects: The 1998 Revision: Volume 1: Comprehensive Tables*. United Nations, New York
- Van Esterik P 1982 *Women of Southeast Asia*. Dekalb: Center for Southeast Asian Studies, Northern Illinois University, Urbana, IL
- Wolters O W 1999 *History, Culture, and Region in Southeast Asian Perspectives*, rev. edn. Southeast Asia Program, Cornell University, Ithaca, NY
- Wongboonsin K, Ruffolo V P 1995 Sex preference for children in Thailand and other Southeast Asian countries. *Asia-Pacific Population Journal* 10: 43–62
- Yap M-T 1995 Singapore's 'three or more' policy: the first five years. *Asia-Pacific Population Journal*. 10: 39–52

C. Hirschman

Fertility Trends in the Formerly Socialist Countries of Europe

The Formerly Socialist Countries of Europe (FSCE) include Albania, Bulgaria, Hungary, Poland, and Romania, as well as the newly independent states, namely the Czech Republic and the Slovak Republic from the former Czechoslovakia; Belarus, Estonia, Lithuania, Latvia, Moldova, Russia, and Ukraine from the former USSR; Croatia, Bosnia and Herzegovina, Slovenia, Macedonia, and Yugoslavia from the Socialist Federated Republic of Yugoslavia; and

finally the eastern *Länder* of Germany, which from May, 1949 to October 3, 1990 formed an independent state—the German Democratic Republic (GDR).

Relatively heterogeneous demographically, the FSCE have *grosso modo* followed a common pattern during the twentieth century, namely a late onset and high speed of the fertility transition as compared to Western and Northern Europe; a specific environment of family economics and other family policies applied under the socialist system during the four post-war decades; and a rapid transition from a centralized to a market economy since the beginning of the 1990s.

1. The Fertility Transition

The fertility transition started during the last quarter of the nineteenth century in the more industrialized and urbanized Czech Republic (which was part of the Austro-Hungarian Empire until 1918), in the former GDR (Western Prussia), and in Hungary (the Kingdom of Hungary in the Austro-Hungarian Empire). Elsewhere (Albania excluded), the beginning of a stable fertility decline may be dated to about the first two decades of the twentieth century, though exact estimation is difficult because of social troubles and wars in 1912–20 (Coale 1969, Pavlik 1991). The completed fertility (CF) of women born in 1910, a central cohort whose reproductive behavior reflected fertility dynamics during the first half of the twentieth century, was below 2.5 births per woman in the Czech Republic, the GDR, and Hungary (as it was in England and Wales), about 3.0 in Bulgaria and Slovakia (as in The Netherlands), and between 3.5 and 4.0 in Poland, Rumania, Yugoslavia, and Russia (Festy 1991). In general, the FSCE followed the West European fertility transition pattern between the two World Wars, and on the eve of World War II, Albania was the only country not to have begun its fertility transition.

2. Trends after World War II

Fertility in the FSCE in the second half of the twentieth century was marked by a notable deviation from the Western and Northern European pattern. While fertility rose throughout the West after the 1950s (see *Baby Booms and Baby Busts in the Twentieth Century*), in the FSCE it continued to decline without interruption, reaching the then European low in the early 1960s (Sardon 1998). In Hungary, the TFR fell below 2 in 1961, in Romania in 1964, in the Czech Republic in 1967, in the Ukraine in 1965, and in Russia in 1968. The TFR level necessary for generation replacement (at least 2.1) was still maintained in Poland, the GDR, Belarus, Bulgaria, Moldova, and all of Yugoslavia,

but this was mainly caused by a lowering of child-bearing age, which fell by two years between 1950 and 1965 (see *Fertility Change: Quantum and Tempo*), while cohort fertility continued to fall.

Furthermore, in Yugoslavia the relatively high national fertility level concealed strong regional and ethnic variance: while the TFR was about 3.5 in Bosnia and Herzegovina in 1965, 3.7 in Macedonia, and 3.1 in Montenegro, it was approaching replacement level in the other Yugoslav republics: 2.02 in Croatia, 2.4 in Slovenia and Serbia. Within Serbia, again an enormous difference existed between Kosovo, where the TFR exceeded six births per woman in the mid-1960s (as it did in Albania), and the other two autonomous provinces, Central Serbia and Vojvodina, where it was 2.03 and 2.13, respectively.

From the mid-1960s, when the fertility decline took new speed in Western Europe (see *Demographic Transition, Second*), the TFR stabilized for almost two decades at values of 2.0 to 2.5 in Central and Eastern Europe. Thus in the 1970s and 1980s the FSCE once again had the highest fertility levels in Europe. The stabilization of fertility in the FSCE might be explained by a particular 'socialist' type of family formation resulting from policies of social paternalism towards the family, especially toward the young family, applied throughout the FSCE since the beginning of the 1950s. Under the conditions generated by these policies, based on the principles of social egalitarianism and state guarantees of equal and universal access to education and health services, family formation (marriage and childbearing) became one of the first stages of an adult's social adaptation. It often was a condition for independent housing arrangements and it preceded a definitive choice of profession. This resulted in younger childbearing and in a strong compression of fertility around the modal age at childbearing. Furthermore, this policy stimulated the development of a two-child family system. Throughout the FSCE, the proportion of women who were childless at age 50 was very close to the natural level of sterility, i.e., about 10 percent. At the same time, families with four and more children were rare, especially in the countries that had a high level of industrialization in the pre-socialist period (the GDR, the Czech Republic, Hungary, Slovenia).

In the countries where economic modernization had not been achieved by World War II (those east of the 'Hajnal line'), the traditions of early and universal marriage as well as early childbearing persisted, because the 'gentle' socialist form of industrialization avoided the explosion of unemployment, a drop in living standards, and mass rural-urban exodus.

Another factor explaining the stabilization of (or even slight increase in) fertility was the special pronatalist policies implemented between the mid-1960s and the mid-1980s throughout the FSCE, again with the exception of Albania (Henry and McIntyre 1981). Based on an array of incentives such as family

allowances, birth payments, paid maternity and post-maternity leave, and subsidized credit, these policies aimed at stimulating second-, third- and fourth-order births (Klinger 1991). The most powerful pronatalist incentives were introduced in Hungary, Czechoslovakia, Bulgaria, and the GDR, where they substantially influenced period fertility indicators, and less strongly the fertility of female cohorts born around 1950 (± 5 years). The strongest growth in cohort fertility was observed in Hungary, where CF rose from 1.9 for women born in 1942-45 to 2.03 in the cohort of 1961. In Czechoslovakia, CF gained less, from 2.03 in the cohort of 1943 to 2.1 in the cohort of 1952-58, while in the GDR and Bulgaria, the pronatalist measures only stopped the decline and stabilized cohort fertility (at around 1.8 in the former for the cohorts born in 1952-58 and 2.05 in the latter in the cohorts born in 1945-57). Thus, all these incentives were unable to raise cohort fertility to replacement level. In 1983-86, the USSR repeated this pronatalist policy experiment, but without any restrictions on abortion, and obtained the same result in its European republics, namely a short-term increase of the period TFR to 2.2 in 1984-87, followed by a rapid decline after the fall of the Soviet Union in 1988. In Romania and Poland, the pronatalist measures were more modest. Finally, in Yugoslavia, child allowance payments were decentralized and increased for lower-income families, which suggests a social policy orientation rather than a purely pronatalist policy (Henry and McIntyre 1981).

The availability of induced abortion was another key element of the family policies of a number of FSCE countries. Following the example of a Soviet abortion law passed in November 1955, all the FSCE except the GDR and Albania liberalized induced abortion during the late 1950s. Later, the abortion law became more restrictive again in Romania (1966), Bulgaria (1968), Czechoslovakia (1973), and Hungary (1974). These restrictions had little long-term effect on fertility because they were accompanied by an extension in contraceptive use, encouraged by the governments to protect women's reproductive health (see *Fertility Control: Eastern Europe*). Only in Romania, where the level of modern contraceptive use was very low, did the strongest ban on abortion ever (implemented on 1 November, 1966 by Decree 770) result in the most spectacular rise in fertility in post-war Europe: the TFR increased from 1.92 in 1966 to 3.67 in 1967. Fertility fell quickly again, after 1969 however, and later amendments reinforcing Decree 770 (in 1974, in 1984, and in 1986) had only a very weak influence on trends in the TFR.

The main results of these fertility-stimulating and restrictive incentives were a strong shift in fertility to younger ages (the cohort MAC—mean age at childbearing—decreased to 25 ± 1 years) and a greater compression of fertility spacing toward its natural limits within the frame of the two-child system, with a

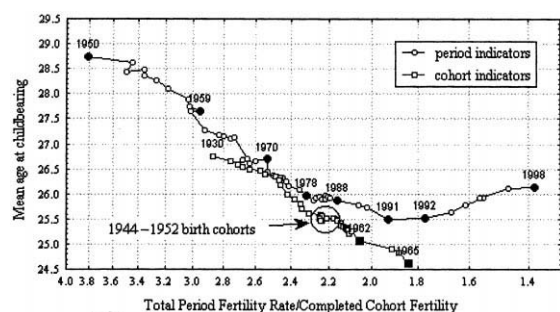


Figure 1

Evolution of period and cohort fertility indicators in the formerly socialist countries of Europe

Source: Data Base of European Demographic Observatory communicated by G Calot and J P Sardon; the estimation of the TFR in the USSR republics for the period until 1975 was performed by Alain Blum. (Personal communication)

CF in the region of 1.9 births per woman (see Fig. 1). In the GDR, where fertility compression was relatively moderate by FSCE standards, about 60 percent of fertility was concentrated in the age range of 19–25 years in the mid-1970s. By the beginning of the 1980s, this pattern of reproductive behavior had spread throughout the FSCE except for Albania, the autonomous province of Kosovo and, probably, Macedonia.

In the early 1980s, when the crisis of the socialist system emerged throughout the FSCE and was absolutely obvious in Poland and the former Yugoslavia, the first signs of a corrosion of the 'socialist' pattern of family formation also became visible. This showed up in an increase in the mean age at childbearing (Hungary, the GDR, Czech Republic, Croatia, Slovenia), followed by a decrease in TFR since the mid-1980s. In Central Serbia and the autonomous province of Vojvodina, where the economic crisis began earlier, the MAC started to rise in the early 1970s while fertility was below replacement level. Only in the European republics of the USSR, Poland, Bosnia and Herzegovina, Macedonia, and Kosovo was the fertility decrease accompanied by a reduction in the MAC until the mid-1990s.

With the collapse of the socialist system in Europe (1989–1991) and the disappearance of the social and economic support of the 'socialist' pattern of family formation, fertility plummeted in all FSCE countries. In the ex-GDR, the Czech Republic, the Slovak Republic, Romania, and the newly independent states of the former USSR, the first reaction of fertility to the social changes has been described as a paralyzing shock that led to an abrupt but brief fall in the mean age at childbearing due to a decrease in fertility rates at all ages. Subsequently MAC began to increase while

the TFR decreased. The most spectacular fertility drop occurred in the former GDR, where the TFR fell from 1.56 in 1989 to 0.77 in 1993–1994. In the rest of the FSCE, fertility followed the trends of the late 1980s, except perhaps a little faster. It seems that the economic troubles definitively brought the fertility of women born in 1962–65 down below the level that would have been reached otherwise. The fertility of younger women has been checked too, but this may perhaps be made up later (Philipov and Kohler 1999). Moreover, a rapid rise in out-of-wedlock births is observed all over the FSCE and is especially remarkable in Estonia, where the proportion of births outside marriage rose from 20.7 percent in 1985 to 51.6 percent in 1997. In the former GDR, the increase was from 33.8 percent to 44 percent, in Slovenia from 19.1 percent to 32.7 percent, in Latvia from 14.4 percent to 34.8 percent and in Bulgaria from 11.7 percent to 30.1 percent.

In summary, it seems likely that a new pattern of family formation is now emerging in the FSCE, characterized by later marriage and by late and diversified fertility similar to that observed in Western Europe. This observation is supported by the stabilization of (or even a slight increase in) period fertility observed in the former GDR, Moldova, Romania, and the Czech Republic in the second half of the 1990s, in contrast to the fertility decline still going on elsewhere in the FSCE. At the present time it is, however, impossible to draw any definitive conclusions about the near future of fertility in the FSCE. What the future will bring depends in large part on the political stability and the economic situation in these countries.

See also: Eastern European Studies: Culture; Eastern European Studies: Economics; Eastern European Studies: History; Eastern European Studies: Politics; Eastern European Studies: Society; Fertility Control: Overview; Reproductive Rights in Developing Nations

Bibliography

- Coale A 1969 The decline of fertility in Europe from the French Revolution to World War II. In: Berhman S J, Corsa L, Freedman R (eds.) *Fertility and Family Planning: A World View*. Ann Arbor: University of Michigan Press, pp. 3–34
- Festy P 1991 La fécondité en Europe de l'Est depuis 1950. *Population* 46(3): 479–510
- Henry P D, McIntyre R J 1981 *Reproductive Behavior: Central and Eastern European Experience*. Springer, New York
- Klinger A 1991 Les politiques familiales en Europe de l'Est depuis 1950. [Population policies in Eastern Europe since 1950]. *Population*. 46: 511–26
- Macura M 1997 Fertility and nuptiality changes in Central and Eastern Europe: 1982–1993. *Stanovništvo* 3–4: 11–43

- Pavlik Z 1991 Les tendances démographique longues en Europe de l'Est. [Long-term demographic trends in Eastern Europe]. *Population* 46(3): 463–78
- Philipov D, Kohler H-P 1999 *Tempo Effects in the Fertility Timing and Quantum Effects on Fertility in Bulgaria. The Czech Republic and Russia*. Max Planck Institute for Demographic Research. Working Paper 1999–008
- Sardon J-P 1998 Fécondité, bouleversements politiques et transition vers l'économie de marché en Europe de l'Est. *Population, Espace, Sociétés* 3: 339–60

A. Avdeev

Feud and Internal War: Legal Aspects

Feuding and internal war are two types of violent collective conflict. Long of interest to anthropologists, feuding and internal war may be most common in simple, kin-based societies, but, as lethal violence between modern urban gangs attests, are not confined to those settings. A central issue for scholars has been the relationship between feuding and internal war and the state's system of social control or law: are feud and internal war the antithesis of law or the functional equivalent of law? In fact, they are neither. The emergence over the past quarter-century or so of a new field of sociological inquiry known as conflict management provides a new conception of the relationship, one in which law and violence are similar, though not identical, ways of handling disputes triggered by similar, though not identical, social conditions (Black 1976, 1993).

1. The Terms 'Feud' and 'Internal War'

Scholars use two strategies to define the terms 'feud' and 'internal war,' strategies that might be labeled qualitative and quantitative, respectively. The qualitative strategy treats feud and war as distinct subtypes of violence; the quantitative strategy locates feud and war at different points on the same continuum of violence.

The work of Keith Otterbein exemplifies the qualitative strategy, the more common of the two. In an influential paper, Otterbein and Otterbein (1965) isolate three types of violence: feuding, internal and external war. Feuding the Otterbeins define as violence within a political community; internal war as violence between political communities within a society; and external war as violence between societies. Useful though these distinctions may be for some purposes, they do not appear to be necessary to answer the central theoretical question: what explains violent conflict? Otterbein and Otterbein found that the best predictor of feuding in a society is not the complexity of political organization but the presence of co-resident

related males organized for mutual self-protection or 'fraternal interest groups.' In a later paper, Otterbein (1968) reports that fraternal interest groups are also strongly associated with internal war (compare Ericksen and Horton 1992). If the same variable predicts both types of violence, much of the rationale for treating them as qualitatively distinct disappears.

The quantitative strategy is used by Donald Black. Black has developed a field of sociological inquiry known as social control or conflict management that addresses the myriad ways in which people and groups handle disputes. Diverse though conflict management behavior is, Black argues that it can be reduced to five 'elementary forms': negotiation, avoidance, toleration, self-help (or aggression), and third-party settlement (such as mediation, arbitration, or adjudication). Each elementary form can be found at different structural levels: among individuals, organizations, nations. For Black, there is therefore no sharp discontinuity between interpersonal, group, and international self-help. However, self-help does vary in the amount of violence it entails. Measures of the violence of self-help include the presence and lethality of weapons, the number of people attacked, the reciprocity of the violence, and whether the assault is accompanied by torture or mutilation. On this line of thought, feud and war (as well as other forms of self-help, such as assault, homicide, and retaliatory killing) differ not in kind but in degree: war is more violent. (Consistent with it, the remainder of this entry will refer to both feuding and internal war as feuding, except where it is important to distinguish the two.) In practice, there is often an additional contrast: the parties to war are usually large groups, such as societies (external war) or tribes (internal war) while feuds are contested by smaller groups, such as families. In the quintessential war, then, armies trade multiple rounds of lethal violence, seek to slaughter as many of the enemy as possible, and exempt nobody from attack. In the quintessential feud, antagonistic families kill at least once, abide by the principle of 'a life for a life,' and avoid targeting women, children, and non-kin.

2. Feuding and Its Alternatives

Although feuding has been practiced in many societies, there is no such thing as a feuding society. Only some conflicts give rise to feuds, even among those people well known for feuding, such as medieval Icelanders (Miller 1990), highland Albanians (Hasluck 1954), and Australian aborigines (see, e.g., Warner 1958). For instance, while homicides between families may precipitate prolonged and bitter feuds, those within families are likely to attract nothing more than some criticism and avoidance of the killer (see, e.g., Peters 1967).

Even when homicides, insults to honor, or other offenses occur within relationships that can result in a feud, a feud is rarely inevitable. In many societies, the victims's kin have the option of accepting compensation from the offender's kin in lieu of retaliation (see, e.g., Boehm 1984). Whether they agree to compensation or launch an attack (which may then trigger a counterattack) depends on a variety of factors, including their social status relative to that of the other side, and the nature and number of cross links between them (see Black 1993, Chap. 3). If they decide to seek compensation, they can arrange payment either by directly negotiating with the other side or, more frequently, by calling upon the services of a third-party skilled in the resolution of conflict, such as a mediator. Among the Nuer (of Sudan), for instance, a third-party known as the Leopard-Skin Chief traditionally intervened in disputes, such as homicides, with a view to having the parties agree to settle the case non-violently through a transfer of cattle from the killer's kin to the victim's kin (Evans-Pritchard 1940, pp. 150–76). The Leopard-Skin Chief could not force the parties to resolve the matter peaceably; nor did he referee the merits of the dispute; the only tools available to him were his powers of persuasion and his threat of supernatural sanctions if the parties did not settle. The Leopard-Skin Chief's mediation was most likely to succeed in disputes between neighbors and acquaintances: if the parties were intimates his intervention was likely to be unnecessary; if they were relative strangers his efforts were likely to be disregarded.

3. The Relationship Between Law and Feud

The noted anthropologist, Paul Bohannan (1967, p. xiii) once wrote that 'there are basically two forms of conflict resolution: administered rules and fighting. Law and war.' Later developments in the sociology of conflict management reveal that Bohannan's dichotomy is too simple. Recall Black's analysis of the five elementary forms, a scheme that goes beyond law and warfare to include negotiation, avoidance, toleration, and non-legal forms of third-party settlement. However, even though Bohannan's dichotomy neglects these additional forms, it may be correct in pitting law and war as opposites, as radically contrasting ways of handling conflict.

3.1 Law and Feud as Opposites

The idea that law and violence are fundamentally opposed is a popular one. Where law represents order and reason and civilization, feud represents disorder, passion, and tribalism. Pospisil (1968, p. 392) articulates a central tenet of this position when he states:

That law stands in opposition to feud and that the latter is the antithesis of the former ... is well documented in those societies in which feuds are stopped by legal decisions of the over-all authority who either has enough power to enforce his will or possesses the skill to persuade the quarrelling parties to accept his solutions.

In reality, however, law does not invariably strive to eliminate violence. Early states appear to tolerate feuding (see, e.g., Otterbein and Otterbein 1965). Moreover, even when law does seek to bring fighting to an end it often must do so with tactics and procedures similar to those used by the protagonists. If the parties persist in pursuing their conflict violently, the only way to stop them is to use overwhelming force. However, in employing violence to end violence, the law becomes not so much a party outside the feud as a more powerful party that wins the feud. That law's authority ultimately rests on physical force blurs the distinction between law and feud.

3.2 Law and Feud as Complementary

A second view is that law and feud are complementary institutions, at least where law is unavailable. Several eminent anthropologists, including Evans-Pritchard (1940) and Max Gluckman (1956) view feuding and the threat of feuding as a mechanism by which social order is attained and maintained in stateless societies. Evans-Pritchard, for instance, studied the Nuer who have a segmentary lineage system under which opposing kinship subunits unite to support their kinship unit in a conflict with another unit, and those opposing kinship units themselves unite when conflict occurs between larger kinship units. Feuds are least likely to occur within subunits where the parties are bound by ties of kinship and the necessity of daily cooperation; feuds are more common, and more difficult to settle, as the parties become more structurally distant and cooperate less in matters of economics or conflict. Evans-Pritchard (1940, p. 159) proposes that the 'function of the feud ... is, therefore, to maintain the structural equilibrium between opposed tribal segments which are, nevertheless, politically fused in relation to larger units.'

The idea that the feud is the functional equivalent of law in stateless societies qualifies but does not negate the idea that law and feud are opposed. Functional equivalence only applies to stateless societies but feuds and internal wars involving gangs, guerrillas, and other groups may be found in societies with well-established legal systems.

Moreover, only in some stateless societies can feud be the functional equivalent of law. There are stateless societies that have neither law nor feuding, societies in which people employ entirely peaceful methods of settling disputes (see, e.g., Howell and Willis 1989).

Finally, the idea that violence is functional sits better with feuding than with internal war: the indiscriminate nature of war can result in the political and even physical destruction of a society, rather than its continued existence.

3.3 Law and Feud as Conflict Management

What, then, is the relationship between law and feuding? The young discipline of conflict management sheds new light on this issue, revealing both the contrasts and continuities between law and violence. As modes of handling conflict, law and feuding, it is true, are opposed in certain ways. In law, the parties use words; in feuding, they use weapons. Law tends to be hostile to feuding such that in developed or strong states, feuding is met with forceful criminal punishment. Moreover, in law, it is not the adversaries who typically determine the outcome, but neutral third parties, often applying general rules. However, consider some of the ways in which law and feuding are similar:

(a) In both law and feuding the parties confront one another. They dispute through hostile encounters. In so doing, they behave quite differently from disputants who simply avoid their adversary or tolerate the grievance (see Baumgartner 1988).

(b) In both law and feuding the parties seek victory. They do not want or expect to compromise. By contrast, when people negotiate their conflicts or invoke a mediator they normally anticipate that any settlement will require them to make concessions to the other side.

(c) In both law and feuding third-party supporters typically get involved. Legal conflicts draw in lawyers and witnesses; feuds draw in warriors. In either case, supporters are often crucial to the outcome of the case (Black 1993, Chap. 7).

Black's work not only allows the close similarities and the distinct dissimilarities to be described but to be explained as well. Black argues that the explanation of why violence or law or any other type of conflict management is found lies not in the substance of the conflict but in its social structure. The social structure of the conflict consists of the social characteristics of the participants, including their wealth, social integration, reputations for deviance, and relational distance. Black (1995, p. 855, note 130) summarizes the social structure of the classic feud:

- (1) intermediate relational distance between the parties (neither strangers nor close associates), (2) relational segmentation between the parties (divided by a social chasm, without cross links), (3) functional independence of the parties (lack of division of labor), (4) solidarity of each party (intimate and homogeneous), (5) homogeneity between the parties (same ethnicity), and (6) equality between the parties (similar size and resources).

Black adds:

Narrow the distances in the model by reducing the relational separation and independence of the parties, and the reciprocity and continuity of the violence declines. Increase the distances ..., and the violence becomes more indiscriminate and warlike.

Now consider litigation. According to Black's (1976) theory of law (a theory supported by a substantial amount of evidence), the first three characteristics of the social structure of the feud above would also be conducive to extensive litigation. However, the fourth, fifth, and sixth would not: litigation is attracted more to conflicts between unequals than equals, and to people of different ethnicities than the same ethnicity. In other words, conflicts in which the parties fight and those in which they invoke law have social structures that, while not identical, are similar.

3.4 The Fate of Feud-conflicts under Law

The affinity between the social structures of law and of violence provides an answer to another puzzle: how do disputants handle the underlying conflicts when feuding becomes increasingly difficult? What happens to the feud-conflicts that legal authorities successfully prohibit being fought? Cross-cultural evidence suggests that, in the absence of violence, the contending parties generally turn to forms of conflict management similar to violence. Rarely, for example, do the parties simply tolerate the grievance. Nor do they avoid one another or even engage in extensive negotiations. Most commonly, they resort to one of two courses of action. The first is to replace fighting with litigation. Parties who once traded blows now trade legal claims in a kind of sublimated violence. For instance, before the arrival of state authority, the Mae Enga of New Guinea had a substantial amount of interclan fighting, resulting in a comparatively high rate of lethal violence. The colonial administration's efforts to end the fighting were initially unsuccessful, but eventually the Enga began to turn to litigation:

This litigation the Central Enga quickly defined as 'fighting in the courts,' and they regarded the outcomes as equivalent to military victories or losses affecting whole phratries. Thus the men of a growing clan who believed their group required more land would carefully concoct a case, often spurious, against a weaker non-fraternal neighbor. After patiently rehearsing his followers in the statements they were to make, the Big Man [i.e., Enga leader] ... took the matter to the kiap with the realistic expectation that the sheer weight of testimony would persuade the overworked officer, who had neither the time nor the facilities to sift all the evidence. The indignant defendants of course argued strenuously in rebuttal, but, given the absence of written land records and the intrinsic difficulty of determining 'historical facts' ... it is

understandable that resolute lying often carried the day. However the case was decided, the disgruntled losers would soon fabricate a counterclaim (or polish up a valid complaint) against an apparently vulnerable fraternal clan of the victors (Meggitt 1977, p. 155).

The second course of action is to make increased use of indigenous forms of third-party settlement. If, previously, feuding and mediated compensation payments were alternative ways of handling, say, homicides, mediation payments now come to be the dominant method. For instance, after coming under colonial rule, the Nuer began, with the encouragement of the authorities, to bring all homicides to the Leopard-Skin Chief to be resolved through compensation (though the killer was also prosecuted and punished separately). Thus, by settling homicide cases between lineages that once might have resulted in reprisals, the role of feud contracted while that of mediation expanded (Howell 1954, p. 61).

Whichever option—using judges or mediators—is pursued, both are possible because of the similarity in the underlying social structure of violence and litigation and, more generally, third-party settlement. The arrival of law into a setting previously stateless provides a new outlet for disputes but it does not, at least in the short term, fundamentally alter the underlying relationship between the parties. Hence, when law forbids feuding, former feudists are usually attracted more to litigation, mediation, or arbitration than to forms of conflict management bearing fewer resemblances to violence. Law and feuding, then, are neither opposites nor complements, neither strangers nor siblings. Perhaps they are best thought of as cousins.

See also: Conflict and Conflict Resolution, Social Psychology of; Domestic Violence: Sociological Perspectives; Internal Warfare: Civil War, Insurgency, and Regional Conflict; Violence, History of; Violence in Anthropology; War, Sociology of; Youth Gangs

Bibliography

- Baumgartner M P 1988 *The Moral Order of a Suburb*. Oxford University Press, New York
- Black D 1976 *The Behavior of Law*. Academic Press, New York
- Black D 1993 *The Social Structure of Right and Wrong*. Academic Press, San Diego, CA
- Black D 1995 The epistemology of pure sociology. *Law and Social Inquiry* 20: 829–70
- Boehm C 1984 *Blood Revenge: The Enactment and Management of Conflict in Montenegro and Other Tribal Societies*. University of Pennsylvania Press, Philadelphia, PA
- Bohannon P 1967 Introduction. In: Bohannon P (ed.) *Law and Warfare: Studies in the Anthropology of Conflict*, 1st edn. Natural History Press, Garden City, NY
- Ericksen K P, Horton H 1992 Blood feuds: cross-cultural variations in kin group vengeance. *Behavior Science Research* 26: 57–85

- Evans-Pritchard E E 1940 *The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People*. Oxford University Press, London
- Gluckman M 1956 *Custom and Conflict in Africa*. Barnes and Noble, New York
- Hasluck M 1954 *The Unwritten Law in Albania*. Cambridge University Press, Cambridge, UK
- Howell P P 1954 *A Manual of Nuer Law*. Oxford University Press, London
- Howell S, Willis R (eds.) 1989 *Societies at Peace: Anthropological Perspectives*. Routledge, London
- Meggitt M 1977 *Blood is their Argument: Warfare among the Mae Enga Tribesmen of the New Guinea Highlands*, 1st edn. Mayfield, Palo Alto, CA
- Miller W I 1990 *Bloodtaking and Peacemaking: Feud, Law, and Society in Saga Iceland*. University of Chicago Press, Chicago
- Otterbein K F 1968 Internal war: a cross-cultural study. *American Anthropologist* 70: 277–89
- Otterbein K E, Otterbein C S 1965 An eye for an eye, a tooth for a tooth: a cross-cultural study of feuding. *American Anthropologist* 67: 1470–82
- Peters E L 1967 Some structural aspects of the feud among the camel-herding Bedouin of Cyrenaica. *Africa* 37: 261–82
- Pospisil L 1968 Feud. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York
- Warner W L 1958 *A Black Civilization: A Social Study of An Australian Tribe*. Harper and Brothers, New York

M. Cooney

Feudalism

Definitions of feudalism ‘vary widely in their degree of comprehensiveness’ (Max Weber). Historians and social scientists of the nineteenth and twentieth centuries thus do not agree on a standard definition of the term. There is a great deal of controversy about the scope and the explanatory value of the term feudalism. Characteristic of this debate is a plurality of concepts of feudalism. The diversity of definitions and concepts is attributable to differences in both the epistemological interests and in the historical and social theories that inform the approaches of each particular scholar, school of thought, and orientation.

1. Etymology

Etymologically, the adjective ‘feudal’ and the noun ‘feudalism’ are derived from the Germanic *feod* and the Latin *feudum*, which originally meant ‘cattle, money, property.’ During the course of the eleventh century, *feudum* became a technical legal term for the fief with which the lord invests a vassal under specific conditions. Being provided with this fief obliged the vassal to serve his lord militarily. Feudal law, the sum of all recorded feudal legal norms, was known as *ius*

feudale in medieval and early modern legal terminology and as 'féodalité' beginning in the seventeenth century. During the course of the eighteenth century, structural changes of a social and political nature brought about a fundamental shift in the meaning of the term 'féodalité.' Since then, the words 'féodal' and 'féodalité' have been used to designate not only the legally vested relationship of dependence between a lord and his vassal, but also an entire societal condition: the political and social condition of the *ancien régime*. A legal term was thus transformed into a kind of political catchword in order to call into question the legitimacy of a traditional political and social order. Thus associated with socio-critical connotations, 'feudalism' entered into the languages of Europe as both a word and a concept.

2. Historical Controversies about Feudalism

In his *Esprit des Lois* (1748) Montesquieu equates the '*lois féodales*' with the breakdown of public power and the disintegration of the central authority of the French state into territorial centres of power. Voltaire (1694–1778) denounces the '*system féodale*' as a system of sovereignty in which public power was divided among 'countless little tyrants.' In *Qu'est-ce le tiers état*, a text appearing in 1789, Abbé Sieyès attacks the nobility and the church as 'agents of feudality' whose privileges were incompatible with the ideal of equality under the law. In his view, freeing France from the 'ruins of feudality' was the most important task at hand. Influenced by persistent and widespread unrest among the peasants, the national assembly decreed the 'abolition of feudalism' (*abolition de la féodalité/du régime féodal*) on August 11, 1789. In the writings of Karl Theodor Welcker (1790–1869), one of the leading thinkers of Southern German liberalism, the feudal system is a glaring counterpart to the liberal order of the early Germanic period with its public cooperations. In keeping with the polemical use of the word '*féodalité*' in revolutionary France, Welcker employs phrases like 'feudal despotism' (*tyrannie féodale*) and 'feudal anarchy' (*anarchie féodale*) in order to characterize feudalism as a form of domination which, based on the violent usurpation of power, was blatantly inconsistent with the early Germanic ideal of popular self-determination ('*Volksfreiheit*').

3. 'Feudal Rule' and 'Feudal Mode of Production' in the Work of Georg Wilhelm Friedrich Hegel and Karl Marx

In Hegel's philosophy of history, the 'feudal system' constitutes a step in the historical process which brought about the 'progress from feudal government to monarchy.' Hegel defines feudal government as a

'system of private dependence and private commitment' in which 'the general injustice' and 'the general rightlessness' assumed an institutionalized form. Under the conditions of feudal rule, the lawful power of the state was superseded by a more personalized form of power which individuals held on the basis of their 'private property.' According to Hegel, 'the wilful brutality of particularistic law' prevented the formation of an administrative and controlling authority that could act for the common good to protect the freedom and equality under the law of its civil subjects. In Hegel's view, only the legal government of a monarch can adequately guarantee freedom. 'Feudal rule is a polyarchy; it consists of nothing but masters and slaves. In a monarchy, on the other hand, there is only one master and no slaves because under a monarch, the bonds of slavery are broken and justice and law prevail. True freedom issues from a monarchy.'

In Marx's view of history, 'feudal society' with its typical 'feudal relations of property' and its 'feudal mode of production' constituted a 'progressive economic form of society that produced a new class, the bourgeoisie, which broke down the 'feudal barriers to production.' Marx does not, however, work out an explicit theory of the origins, structure, and transformation of feudalism in detail; he never systematically treats the dynamics of the feudal mode of production. For him, terms like the Middle Ages, feudalism, and feudality are interchangeable. Marx links the Middle Ages and feudalism to the idea of a premodern society that created the necessary conditions for the rise of a capitalistic society which eventually overcame and dissolved feudalism itself. In the 'Communist Manifesto,' he writes that 'the means of production and exchange on whose foundation the bourgeoisie built itself, were generated in feudal society.' 'Modern bourgeois society that has sprouted from the ruins of feudal society has not done away with class antagonisms. It has but established new classes, new conditions of oppression, new forms of struggle in place of the old ones.' Marx is interested in the Middle Ages as the epoch that preceded the 'epoch of the bourgeoisie.' He is concerned more with the political state of medieval society than with an analysis of the feudal mode of production. Like the liberal critics of prerevolutionary corporate society, Marx assumes that in the feudal society of the Middle Ages, 'the supreme command in war and justice is based on landed property.' In his opinion, past and present share common structures with regard to this connection between power and property, for in both feudal society and a bourgeois society, the form of ownership determines the form of rule. The 'supreme command in industry is based on the ownership of capital,' while rule in feudal society was based on the control of land. However, the difference between feudal and bourgeois power relations was that in the Middle Ages, control of landed property determined

both social standing and legal rights to rule; the 'constitution of private property' was also the 'political constitution.' In feudal society, 'property or membership to a privileged family or the nature of one's occupation ... were essential elements of political life in the form of landlordism, class, and corporation.' Control over land formed the basis for rule over people. Noble birth was a kind of guarantee for the privileges of a lord. Private corporations such as guilds administered justice over their members, a right which the modern constitutional state reserves exclusively for itself. Narrow 'personalized relations of dependence' characterized the 'social conditions of material production.' Bond men were personally dependent on a small class of landholding nobles; as serfs, they were obliged to pay dues to their feudal lord and provide him with certain services.

Interpreting history from a Eurocentric perspective, Marx equates classical antiquity with a slave-holding society, the Middle Ages with feudalism, and the present times with capitalism. This classification of history reflects Enlightenment philosophy's belief in progress and combines it with the view of history held by postrevolutionary historians and social scientists. In their view, '*féodalité*' is a term to describe both an epoch and a society: in the linear, evolutionary course of world history, '*féodalité*' or more exactly the '*système théologique et féodal*' was a link between the society of ancient slaveholders and the '*système industriel et scientifique*' that produced the French Revolution.

4. Feudalism as Conceived by Historians and Social Scientists: Max Weber, Otto Hintze, and Marc Bloch

Weber's concept of feudalism has nothing in common with revolutionary and liberal authors' view of feudalism as a system of anarchy, injustice, and violence. Weber integrates socio-political, legal, economic, and spiritual-religious structural elements into a single model. Weber defines feudalism as a traditional system of governance that combines economic, social, and cultural factors. The idea that a feudal society is determined exclusively by socioeconomic characteristics—its 'feudal mode of production'—is quite alien to him. In his view, feudalism is not a type of economic management, but rather a type of domination that requires an economic basis. Weber argues that as a form of organizing and exercising domination, feudalism is a political phenomenon. The structural principle of feudal power relations is the fief. Feudal relations determine the structure and form of the political order, a hierarchically graded feudal pyramid that consists of feudal lords, vassals, and their subvassals. Weber defines *enfeoffment* as 'any investment of rights, especially the right to use land or to exercise territorial domination in exchange for military or

administrative services.' The polity was not held together by—to his mind—the integrative forces of a bureaucratically controlled state, but drew on the personal relationship of loyalty between lord and vassal for stability and duration. During investiture this relationship took on the character of a legally binding contract in a symbolic act which included commendation (*homagium*) and an oath of fealty (*fides*). In the absence of a centralized bureaucracy, 'the personal feudal bond of allegiance and the feudal code of honor serve as the glue of political cohesion.' Charlemagne's reform of the army led to a feudalization of military affairs. Knights were provided with economically useful fiefs which obliged them to provide military service. A cavalry composed of heavily armed feudatories replaced an army that once consisted of freeborn peasants. The feudalization of administrative structures transformed the early and high medieval realm into a system of ruled territories (duchies, ecclesiastical and secular principalities, margravates, earldoms) whose bearers were bound to the king by feudal law as tenants in chief. The feudal system was of fundamental significance for the political-legal organization of the kingdom. As a political principle of organization that divided the power of lords, feudalism laid the foundations for a decentralization of political power. Contractually stipulated division of power caused rivalries and conflicts between the central authority of the king and local vassals. By banding together to form corporations, feudatories transformed the feudal union into a corporative state.

Weber also points out that the manor constituted the economic basis of feudalism. The manor was 'the standard object of feudal tenure.' The power of nobles and princes was founded on ownership of land, human beings, and livestock. The conditions under which feudalism emerged and its function as a system of rule lent it a social character. The social differentiation between peasants and trained warriors grew out of the emergence of an armored cavalry. Membership to the noble class of lords was necessary in order to exercise power in the feudal system. Ownership of a manor which yielded rent entitled one to a 'lordly mode of existence' and enabled one to live according to one's social standing. Honor was the most important norm influencing conduct. Work-free rental income was indispensable for a life 'from which the social forms of prince and nobleman originate.' 'Feudalism, especially in the form of free vassalage and, more importantly, in the feudal system appeals to "honour" and personal, freely rendered and sustained "allegiance" as constitutive motives of conduct.' Honor and allegiance, the embodiment of the 'feudal ethics' which are characteristic of Western feudalism, constitute the 'extreme opposite of all objective business relations.' Feudalism produced and cultivated an ethics of the nobility in the form of heroic asceticism, heroic honor, and hero-worship; these ethics contradict the objectiveness of business relations and operations. The idea of the fief

determined the various forms of 'feudal incorporation,' 'all sorts of social relations: from one's relationship to the Saviour to that to his lover.'

To continue with Weber's concept of feudalism, in the long term, 'feudal structures of power with their clearly defined rights and duties' contributed to the 'stability of the economy.' The 'constancy of the legal system' contributed to the 'emergence of a rational capitalistic system and fostered its penetration into the gaps and seams of the feudal system.' According to Weber, the feudal system also had the rationalizing effect of molding society into 'an at least relatively constitutional formation.' The feudal relationship between lord and vassal created a 'constitutional state' which did not arise on the basis of a legal system, but rather on the basis of subjective rights agreed upon by free partners. By strengthening the personal rights of the individual feudatory, a system of social relations emerged in which contract and consensus constituted the predominant forms of societal integration.

In *Character and Spread of Feudalism* (1929) Hintze follows in the steps of Weber's methodology to 'describe the ideal type that underlies the concept of feudalism.' In his effort to precisely define feudalism as a socio-historic category, Hintze follows the rules of modeling ideal types to go beyond the hermeneutical interpretation of historical texts and concepts. Hintze argues that feudalism is more than 'a self-contained and coherent system of institutions' that constitute the 'feudal state.' His concept of feudalism is thus intended to understand and include other phenomena that are not a part of the feudal system in a strict sense. He therefore tries to link political, military, and socio-economic structural elements in his definition of feudalism. He understands feudalism not only as a principle of political organization, but also as 'a principle of military, economic, and social constitution.' According to Hintze, feudal structures had an effect on the development of sovereignty and social stratification in three ways: first, in the 'selection of a highly trained and loyal class of warriors who, bound by private contract, assumed a privileged position; second, 'in the development of a manorial, rural economic form which provided for the privileged class of warriors'; and third, in the 'elevation of this military nobility to the status of local lords who exercised considerable influence or who, because of their arbitrary way of ruling, assumed a privileged position' within the loosely structured feudal state.

Hintze distinguishes between three main phases in the development of feudalism in the history of the Romance-Germanic peoples: 'first, the era of early feudalism in which the military factor predominated—it extends roughly until the end of the twelfth century; second, the era of high feudalism in which the military nobility attained the height of its political influence, either by attaining relative independence from their feudal overlord as was the case under the arbitrary rule of princes in Germany, or in the form of

a corporative federation within a more solid feudal monarchy as it occurred in England when barons joined together in parliament against the crown—it extends roughly up to the sixteenth or seventeenth century; and third, the era of late feudalism in which the main interest of the nobility was in preserving and exploiting its socio-economic position as lords of manors and estate owners—it extends up until the era of the French revolution and up until the dissolution of traditional relationships in rural areas in the nineteenth century.' Hintze sets up guidelines for applying the term feudalism to non-European states, societies, and cultures. In his view, the term feudalism is, strictly speaking, only applicable to a society in which 'military, socio-economic and political feudalism are fully developed.' Working under this premise, Hintze argues that 'feudalism in its fullest sense can only be found—albeit with considerable deviations and particularities—in Russia, the Islamic nations and Japan.' Egypt, the ancient Orient, the Roman Empire, India, and China only exhibit initial signs of a feudalism that extended equally over the political order, the manorial, agricultural economic form and a military nobility which, supported by work-free income, was bound in allegiance to the ruler. In view of the poorly developed possibilities for communication, large kingdoms required a decentralized administration and court system which were meant to function on the basis of the feudal bond of allegiance between ruler and noble vassals.

Like Otto Hintze, Marc Bloch also views feudalism as a social type. His book *La société féodale* is considered to be 'a model of analysis of social structure,' a 'paradigm of a *histoire totale*.' Bloch does not limit himself to describing an ideal type which other scholars can use as a starting point for their own research. He also analyzes and describes European social history from the mid-ninth to the mid-thirteenth centuries, never losing sight of feudal society as the main connecting thread throughout this era. He sees vassalage and the manorial system as 'leading characteristics' of feudalism as a type of society. He designates the manor as 'the fundamental unit of the feudal regime.' According to Bloch's understanding, feudalism encompasses more than just those forms of social interaction that result from vassalage and the feudal and manorial system. He sees feudalism as a society that should be reconstructed as a complete interactive system. He therefore examines the basic structural conditions—including transportation, trade, population, the system of communication, and the monetary system—that constitute the environment of society. He is concerned with describing the '*atmosphère mentale*,' the 'culture of the feudal period' that influenced the social behavior and societal life of the people by means of its 'forms of feeling and thinking' including its literature (heroic song and epic), its education and scholarship, its attitudes toward nature, time and history, and its matters of

faith and religious practices. Bloch identifies 'fear of punishment in hell' as 'one of the strongest social powers of this time.' The 'kingdom of the feudal period' had a 'holy character.'

Personal ties and relationships of dependence had a great influence on the societal interactions which Bloch designates as feudalism. 'A network of bonds of dependence that ran through all levels of human society from top to bottom gave the culture and society of European feudalism its distinctive character.' Forms of personal dependence that existed within families and clans were based on blood ties. Under vassalage and the feudal system, personal bonds were forged through freely made agreements, thus initiating a relationship in which services and duties were interchanged. Legal tradition, voluntary devotion, or involuntary coercion produced legal, economic, and social inequality on manorial estates between noble lords, serfs, and tenant farmers. Central elements of Bloch's concept of feudalism include vassalic dependence, which bound vassals to perform military services for a lord, and manorial dependence, under which villeins and serfs (*coloni, villani, mancipia, servi*) were made to pay their landlords dues and tributes and perform statute labor. Like Max Weber, Bloch sees vassalage as patterns of behavior and relations which shaped relationships within the family, between lovers, and between the believer and God.

Because the members of societies generally designated as 'feudal' were subdivided into groups which were stratified according to profession, prestige, wealth, and power, Bloch examines social divisions as well as legal and economic inequality between the nobility, burghers and peasants, high and low clerics. He views bearers of 'public power' in the 'feudal epoch' (emperors and kings, princes and prince bishops, dukes and earls) not as representatives of the so-called 'feudal state,' a system of public rule distinct from society, but rather as bearers of 'traditional powers' in terms of their social bonds, limitations, and functions. It was, after all, their duty in a legally and socially 'divided society to maintain certain principles of order and unity.'

5. History and the Concept of Feudalism: Marxist, Historical, and Social Science Perspectives

Marx and Lenin are only interested in feudalism as a societal formation that constituted a preliminary stage of capitalism. Influenced by Marx, Lenin, and Stalin, socialist historians and social scientists have developed a theory of feudal society that not only explains its origins and characteristics, but also takes the change in conditions of production as a starting point to develop a criteria for the division of history into epochs. According to the prevailing theory of Marxism, feudal conditions of productions and feudal

social conditions gradually evolved between the end of the fifth and the middle of the ninth centuries. The 'epoch of developed feudalism' lasted from the middle of the eleventh until the end of the fifteenth century. 'It was a time in which feudal conditions of production and exploitation in Western and Central Europe prevailed and influenced the economy and society, politics and class struggle, culture and ideology in spite of critical disturbances during the late medieval centuries and variations in the degree reached by capitalistic conditions of production in the most advanced European regions.' Feudalism was able to endure over a span of several epochs because the 'early bourgeois revolution' (peasants' war, Reformation) miscarried. As a 'late feudal order' characterized by 'unrestricted rule of the nobility,' 'absolute feudal monarchy,' 'fragmentation of the feudal state,' the continuation of feudal conditions of production, and the 'development of capitalistic forms of production,' feudalism continued until the end of the eighteenth century, when it was overcome and dissolved during the French Revolution by 'bourgeois emancipation.'

The question of the 'essence' and 'genesis of feudalism' grew out of an interest in the structures of the precapitalistic form of society. The following characteristics of premodern feudal society arise out of a comparison between the two systems feudalism and capitalism: low level of technique, dominance of a barter economy, and feudal and hence divided ownership of land. Strictly speaking, the owner of the land was the feudal lord. He had the absolute ownership (*dominium directum*) of the parcels of land which were lent to the peasant. The peasant was granted the right to use the land by way of a loan (*dominium utile*). The instruments of production necessary to work the land belonged to him. The villein and serf owed their lord 'surplus work' (*corvée*) and 'surplus products' (payment in kind, dues) for the use of the land. The 'extra-economic compulsion' which issued from the lord lent the exploitative feudal conditions duration and stability. Extra-economic compulsion refers to the connection between property and sovereign rights. Because the lord of the manor also had the right of jurisdiction over the peasants to whom he granted use of his land, control of jurisdiction was a constitutive element of the exploitative conditions of production.

The majority of non-Marxist historians exclude forms of manorial dependence from their concept of feudalism. They use feudalism and feudality as synonyms for the feudal system and feudal law. In their view, a definition of feudalism that includes all forms of dependence including that of the peasant on the manor is an imprecise and thus unsuitable term for ordering and explaining this historical phenomenon. They reject a definition of feudalism that is more broadly conceived to include all areas of political and social life because they say that it cannot be verified by the sources themselves. They argue that the term feudalism loses its heuristic value if the clear-cut

reference to the fief of a vassal who is obliged to perform military service is lost; in their view, extending the term further is linguistically and objectively ungrounded. Faced with the dilemma of defining feudalism either as a form of society or as the feudal system, François Louis Ganshof chooses the latter. His justification for this choice is symptomatic for all historians who, by means of an exact interpretation of the sources, claim to be able to determine 'how it actually was.'

Ganshof defines 'the feudal system as a collection of institutions,' 'that create and regulate two kinds of bonds between one free-born man, called vassal and another, called lord or seignior: the 'vassal' is obliged to obey and serve—chiefly militarily—the 'lord', and the 'lord' is obliged to provide the 'vassal' with protection and a livelihood. The lord normally fulfilled his obligation to provide the latter by granting the use of land known as the 'fief.'

It has often been pointed out that conceiving of the feudal state as a counterpart to a bureaucracy leads to a 'one-sided exaggeration of domination' and does not adequately take 'the communal, co-operative, and corporative elements in the village, the city, and in the country' into account (Otto Brunner). Research in social and constitutional history that focuses on the role of the community in towns and in the country confirms this skepticism. It is argued that there is 'no logical connection' between 'feudalism,' which over-emphasizes the rule of the nobility, and 'communalism,' which sees the community as an organ of civic and peasant self-determination (Peter Blickle). Hence by seeing the peasant as a mere object of domination and exploitation, the 'feudal' point of view ignores historical records and documents to overlook the fact that the peasants themselves were also active subjects in the long development from serfdom into freedom.

Historians who examine changes in social structures in order to explain the development of modern capitalism (Perry Anderson, Rodney Hilton, Maurice Dobb, Eric Hobsbawm, Hans Ulrich Wehler) use feudalism as a term of demarcation. Free of dogmatic fixation, the work of Ludolf Kuchenbuch, Guy Bois, and Rodney Hilton examines the sources closely and focuses directly on its subject in order to empirically substantiate and better clarify the term 'feudal mode of production.' *Feudalism and History* (1956), an anthology edited by Rushton Coulborn, attempts to turn the definition of feudalism as a 'territorial system of governance' into a category of world history that could be used for international comparisons. Drawing on this definition of feudalism, the volume comes to the conclusion that only Western Europe, Japan, and early Chou-China display evidence for the existence of a feudal social system. English and French historians make little attempt to justify their use of the terms feudalism, '*féodalité*,' and '*féodalisme*' to describe structures of medieval and early modern agrarian

societies. This usage is part of the established terminology of the field. The intense debate about feudalism in the former GDR ended after the unification of Germany. The *Jahrbuch für die Geschichte des Feudalismus* (*Annual for the History of Feudalism*) ceased publication in 1990. German historians avoid the term feudalism as an ideologically loaded construct. They consider the term manor ('*Grundherrschaft*') on the other hand, as indispensable. However, the term manor also originated as a politically tendentious term; at the end of the eighteenth and beginning of the nineteenth centuries, enlightened and liberal authors who saw the manor as an illegitimate form of aristocratic despotism used the term to stigmatize the connection between domination and ownership of property which made the peasant dependent on noble lords and removed land from circulation on the free market.

6. Feudalism as a Type of Society

Adherents of a liberal concept of the town which originated in the nineteenth and early twentieth centuries object to usage of the term 'feudalism' or 'feudality' to label medieval society as a whole because, they argue, the political order and the socioeconomic life of the town—municipalized since the twelfth century—can hardly be characterized as 'feudal.' As Otto von Gierke argues in *Rechtsgeschichte der deutschen Genossenschaft* (1868), 'free union,' a new form of incorporation that formed the basis for the medieval town, was destined to 'break the feudal system.' Max Weber recognizes the medieval town as one of the factors that 'broke the restraints of the outer and inner authorities of the feudal era.' According to Marc Bloch, social relations within the town were inconsistent with the 'feudal type.' Because of its 'liberties,' its 'economic expansion,' and 'collective independence,' medieval towns remained 'foreign bodies in feudal society.' As a community in which personally free citizens with equal rights vowed their mutual solidarity by oath, the upward bound aspiring citizens became a 'destructive element for the feudal system.' Experts in the fields of law and the social sciences have recently emphasized that assuming the form of a commune, the medieval town developed a 'principle contrary to the hierarchically structured forms of domination in feudal society' (Gerhard Dilcher). Further, the medieval town proved itself to be 'a point of crystallisation for an alternative principle of legitimacy,' for a 'legitimacy by virtue of statute, not tradition and piety,' and for a 'republican, not a feudal or estate-based model of governance' (Wolfgang Schluchter). Otto Brunner, on the other hand, contends that the medieval municipality developed into 'a special configuration in the world of European feudalism.' He argues that the bond of fealty that was characteristic for the European system of governance,

'so-called feudalism,' also determined the relationship between the town lord and the municipality. In the Marxist interpretation of history, which defines the Middle Ages as 'age of feudalism,' medieval and early modern towns are referred to as 'economic and social units of feudal society.'

In order to compare and distinguish societies with different formations and structures, it is useful to have a relatively wide range of general definitions. The term feudalism is just this kind of general definition: it is a tool of historical study in the attempt to understand and describe premodern society in its entirety. This conclusion cannot be called into question or disproved by the fact that the medieval town had its own special political, social, and economic character, and therefore contributed to the transformation of the so-called feudal society into a civil society. The towns of Old Europe were not sovereign polities; they were bound up in the constitution of kingdoms and principalities. Max Weber refers to the 'feudal town lords' to whom the town council had to pledge loyalty and obedience each time the crown changed hands or new lords came to power. Weber also points out the 'feudal character' that marked the lives of urban patricians and oligarchs and makes reference to the existence of 'feudal unions' within the towns of Northern Italy and the development of a clientele at the hands of the nobility in the towns of Southern Europe. Patrician families claimed to possess the qualities of the nobility and therefore wanted to be regarded as nobility and receive the same status. The councils of late medieval and early modern towns often saw themselves as an authority appointed by God and worthy of the loyalty and obedience of the citizens. The council saw its rule as a system of command and obedience and derived its position not from a mandate of the citizens but rather from its *summa potestas*. When towns tried to increase their power by setting up territories of their own, they acted as feudal lords themselves by exercising dominion over territory and people. The different quality of the medieval and early modern town did not preclude structural elements of a feudal provenance and character. Because empirical evidence indicates that the borders between systems are permeable, it is probably justified to include the town as a special configuration in a more open concept of feudalism.

The concept of feudalism is useful as a kind of comprehensive model which draws connections between the various constituent areas of reality in a given society (politics, economics, law, social behavior, religion, and culture) in order to show and explain how these areas function independently as well as how they interplay with each other as elements in a complete social system. If the concept of feudalism is to be used in this way, then it is necessary to expand on the meaning it has acquired during its genesis with new contents. Only by formulating the concept of feudalism anew is it possible to free it from the ideological burdens of the eighteenth and nineteenth centuries

and make it useful for our current understanding of the past.

The following features can be considered as typical structural characteristics of a feudal society: (a) Understood as political feudalism, the feudal system—in view of the limited communication possibilities—provides regions in the process of expansion which are difficult to organize politically: access to the dispensation of justice, administrative services, and military protection. Organized according to feudal law, land, administrative posts, and legal rights can be equally invested and bequeathed; the result is a system of governance determined by the interplay between manorial and cooperative powers at the level of both the kingdom and the regions. (b) Political feudalism also refers to fragmentation of executive and judicial powers. During the medieval and early modern eras, a monopoly on state power in the hands of a king was—in most cases—unknown. The monarchy and the nobility divided public power among themselves. Noble lords ruled over land and people by virtue of both their own right and royal investment. Competition for power, a consequence of divided and fragmented executive power, allowed for the development of a cooperative social structure that, in the long run, contributed to the disintegration of feudal power and social conditions. The medieval town emerged only because feudalism did not represent a fully realized structural principle with regard to politics and society. Processes of social differentiation could therefore take hold in the seams and gaps of the system, thus forming a basis for the development of the town as a new kind of settlement and social form. (c) In the chiefly agrarian structured societies of early and high medieval Central and Western Europe, the territorial sovereignty practiced by kings, noble lords, and municipal cooperatives was not one of many principles structuring society, but rather the fundamental way of ordering life and was of importance for the whole of society. (d) Decisive for the organization of social life were relations of personal dependence: the vassal was dependent on his feudal lord, the client on his patron, the serf and tenant on their ecclesiastical and secular landlord, the apprentice on this master, the craftsman, homemaker, and wage earner on his manufacturer. (e) The exercise of dominion as well as the extent and range of political participation were closely connected with one's social position. The distribution of political power and social organization were closely linked. This relationship was not only an expression of legal and social inequality, but also of the absence of the division between public and private law. The medieval *societas civilis* was constituted neither as a system of social and economic needs which was separate from the state nor as a society of citizens with equal rights. (f) Medieval and early modern societies were devoid of the inherent laws of secular orders and institutions. Christianity constituted an interpretative and formative power which encom-

passed the whole of society and pervaded and regulated all realms of life. The connective and binding power of the political order was rooted in ideas and principles of the Christian religion which attribute justice and power to God. Not only did manifestations of social inequality require legitimacy in religion, but also justice, the economy, science and art, war, and peace. The social character of religion encompassed more than just works of Christian *caritas*; it also constituted a *vinculum societatis*, a kind of bond that integrated groups and unions through common convictions concerning right and wrong.

Defining feudalism involves the difficult task of comprehending and describing both the unifying and the differentiating aspects of the premodern *societas civilis*, which became thinkable as a social unity through the medieval reception of Aristotle. However, because the urban and rural societies of the Middle Ages and the Early Modern underwent various structural changes, it is crucial from a methodological standpoint to use the term feudalism in a way that makes the difference between ideal-typical model and concrete historical reality clear.

See also: Bloch, Mark Léopold Benjamin (1886–1944); Hintze, Otto (1861–1940); Honor and Shame; Marx, Karl (1818–89); Middle Ages, The; Weber, Max (1864–1920)

Bibliography

- Anderson P 1974 *Passages from Antiquity to Feudalism*. Atlantic Highlands: Humanities Press, London
- Bader K S, Dilcher G 1999 *Deutsche Rechtsgeschichte. Land und Stadt-Bürger und Bauer im Alten Europa*. Springer, Berlin
- Bartel H (ed.) 1982 *Deutsche Geschichte, Vol. 1: Von den Anfängen bis zur Ausbildung des Feudalismus Mitte des 11. Jahrhunderts*, Köln
- Bartel H 1986 *Deutsche Geschichte, Vol. 2: Die entfaltete Feudalgesellschaft von der Mitte des 11. bis zu den siebziger Jahren des 15. Jahrhunderts*. Deutscher Verlag der Wissenschaften, Berlin
- Blickle P 2000 *Kommunalismus. Skizzen einer gesellschaftlichen Organisationsform*. R. Oldenbourg, München, Vols. 1–2
- Bloch M 1930 Artikel feudalism. In: Seligman E R A, Johnson A (eds) *Encyclopaedia of the Social Sciences*. New York (Reprint 1950), Vol. 6, pp. 203–10
- Bloch M 1939 *La société féodale. La formation de liens de dépendance*. Albin Michel, Paris
- Bloch M 1940 *La société féodale. Les classes et le gouvernement des hommes*. Albin Michel, Paris
- Bloch M 1999 *Die Feudalgesellschaft*. Klett-Gotta, Stuttgart, Germany
- Breuer St 1988 Der okzidentale Feudalismus in Max Webers Gesellschaftsgeschichte. In: Schluchter W (ed.) *Max Webers Sicht des okzidentalen Christentums. Interpretation und Kritik*. Suhrkamp, Frankfurt a. M. pp. 437–75
- Brown R A 1973 *Origins of English Feudalism*. Allen & Unwin, London
- Brown E A R 1974 The tyranny of a construct: Feudalism and

- historians of Medieval Europe. *The American Historical Review* 79: 1063–88
- Brunner O 1968 'Feudalismus'. Ein Beitrag zur Begriffsgeschichte. In: Brunner O (ed.) *Neue Wege der Verfassungs- und Sozialgeschichte*. Vandenhoeck & Ruprecht, Göttingen, pp. 128–59
- Brunner O 1975 Artikel 'Feudalismus'. In: Brunner O, Conze W, Koselleck R (eds.) *Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland*. Klett-Gotta, Stuttgart, Germany Vol. 2, pp. 337–50
- Byres T J, Mukhia H (eds.) 1985 *Feudalism and Non-European Societies*. F. Cass, London
- Demski R 1996 *Adel und Lübeck. Studien zum Verhältnis zwischen adliger und bürgerlicher Kultur im 13. und 14. Jahrhundert*. Lang Frankfurt a.M., Germany
- Dilcher G 1993 Kommune und Bürgerschaft als politische Idee der mittelalterlichen Stadt. In: Fetscher I, Münkler H (eds.) *Pipers Handbuch der politischen Ideen*. Piper, München, Germany Vol. 2
- Duby G 1978 *Les trois ordres ou l'imaginaire du féodalisme*. Gallimard, Paris
- Duus P 1969 *Feudalism in Japan*. Knopf, New York
- Fourquin G 1976 *Lordship and Feudalism in the Middle Ages*. Publisher, New York
- Heuvel G v d 1988 Artikel 'Féodalité, Féodal'. In: Reichardt R, Schmitt E (eds.) *Handbuch politisch-sozialer Grundbegriffe in Frankreich 1680–1820*. Oldenbourg, München, H.10, pp. 1–48
- Hilton R H 1985 *Class Conflict and the Crisis of Feudalism. Essays in Medieval History*. Hambledon Press, London
- Holton R H 1985 *The Transition from Feudalism to Capitalism*. St Martin's Press, New York
- Hintze O 1929 Wesen und Verbreitung des Feudalismus. In: Oestreich G (ed.) *Otto Hintze, Staat und Verfassung. Gesammelte Abhandlungen zur allgemeinen Verfassungsgeschichte*. Vandenhoeck & Ruprecht, Göttingen, Germany pp. 84–119
- Krieser H 1984 *Die Abschaffung des 'Feudalismus' in der Französischen Revolution. Revolutionärer Begriff und begriffene Realität in der Geschichtsschreibung Frankreichs (1815–1914)*. Ulstein, Frankfurt a.M., Germany
- Kuchenbuch L 1977 *Feudalismus-Materialien zur Theorie und Geschichte*. Ulstein, Frankfurt a.M., Germany
- Kuczynski J 1992 *Asche für Phönix. Aufstieg, Untergang und Wiederkehr neuer Gesellschaftsordnungen. Eine vergleichende Studie zu Feudalismus, Kapitalismus und 'Realem Sozialismus'*. Papy Rossa-Verlag, Köln, Germany
- Martin J E 1983 *Feudalism to Capitalism: Peasant and Landlord in English Agrarian Development*. MacMillan, London
- Milson St F Ch 1976 *The Legal Frame Work of English Feudalism*. S. F. G. Milson, Cambridge, UK
- Mitteis H 1958 *Lehnrecht und Staatsgewalt. Untersuchungen zur mittelalterlichen Verfassungsgeschichte*. Böhlau Nachfolger, Weimar, Germany
- Mitteis H 1944 *Der Staat des hohen Mittelalters. Grundlinien einer vergleichenden Verfassungsgeschichte des Lehnzeitalters*. Böhlau Nachfolger, Weimar, Germany
- Müller-Mertens E (ed.) 1985 *Feudalismus—Entstehung und Wesen*. Akademie-Verlag, Berlin, Germany
- Neubauer H 1968 Artikel 'Feudalismus'. In: Sowjetsystem und demokratische Gesellschaft. *Eine vergleichende Enzyklopädie*. Herder Freiburg, Germany, pp. 477–90
- Poggi G 1988 Max Webers Begriff des okzidentalen Feudalismus. In: Schluchter W (ed.) *Max Webers Sicht des okzidentalen Christentums. Interpretation und Kritik*. Suhrkamp, Frankfurt a.M., Germany pp. 476–97

- Rader T 1971 *The Economics of Feudalism*. Gordon & Breach, New York
- Roth P 1863 *Feudalität und Unterthanverband*. Böhlau, Weimar, Germany
- Sander P 1906 *Feudalstaat und Bürgerliche Verfassung. Ein Versuch über das Grundproblem der deutschen Verfassungsgeschichte*. Bath, Berlin, Germany
- Schluchter W 1979 *Die Entwicklung des okzidental Rationismus. Eine Analyse von Max Webers Gesellschaftsgeschichte*. Mohr (Paul Siebeck), Tübingen, Germany
- Schreiner K 1983 'Grundherrschaft'. Entstehung und Bedeutungswandel eines geschichtswissenschaftlichen Ordnungs- und Erklärungsbegriffs. In: Patze H (ed.) *Die Grundherrschaft im Mittelalter*. Thorbecke, Sigmaringen, Germany pp. 11–74
- Sidney P 1961 *Feudalism and Liberty: Articles and Addresses*. Johns Hopkins Press, Baltimore
- Stephenson C 1948 *Medieval Feudalism*. Cornell University Press, Ithaca, NY
- Strayer J R 1965 *Feudalism*. Van Nostrand, Princeton
- Sweezy P u. a. 1976 *The Transition from Feudalism to Capitalism*. Verso, London
- Sydney H 1921/1969 *The Fall of Feudalism in France*. Barnes & Noble, New York
- van Horst Bartel 1969 *Sachwörterbuch der Geschichte Deutschlands und der deutschen Arbeiterbewegung*. Dietz, Berlin, Germany Vol. 1, pp. 582–6
- Von Gierke O 1868/1954 *Das Deutsche Genossenschaftsrecht*. Graz Akademische Druck- und Verlags-Anstalt, Graz, Austria Vol. 1
- Ward J O 1985 *Feudalism: Comparative Studies*. Sydney Association for Social Studies in Society & Culture, Sydney
- Weber M 1921/1972 *Wirtschaft und Gesellschaft. Grundriß der verstehenden Soziologie*, 5th edn. Mohr (Paul Siebeck), Tübingen
- Wehler H-U 1987 *Deutsche Gesellschaftsgeschichte, Vol.1: Vom Feudalismus des Alten Reichs bis zur Defensiven Modernisierung der Reformära 1700–1815*. Beck, München, Germany
- Wunder H 1974 Der Feudalismus-Begriff. Überlegungen zu Möglichkeiten der historischen Begriffsbildung. In: Wunder H (ed.) *Feudalismus*. Nymphenburger Verlag, München, Germany
- Wunder H 1989 Artikel 'Feudalismus'. In: Sautier R H (ed.) *Lexikon des Mittelalters*. Artemis, München, Germany, Vol. 4, pp. 411–15

K. Schreiner

Fiducial and Structural Statistical Inference

Fiducial inference is a statistical approach to interval estimation first advocated by R. A. Fisher as an alternative to the then dominant method of inverse probability, i.e., using Bayes' Theorem. Considerable effort has gone into formalizing Fisher's notions using such concepts as statistical invariance and pivotal quantities. This entry describes elements of the fiducial approach and relates them to other currently more widely used statistical approaches to inference. Section 1 introduces some basic inferential ideas via a simple example.

1. A Simple Example

Consider a variable y that is directly available or has arisen by some preliminary reduction process and suppose that y measures θ in an unbiased manner and has error that is normal with known variance σ_0^2 . Then we can say for example that $P(y \leq \theta + 1.64\sigma_0; \theta) = 95$ percent.

Parenthetically, we note that the preliminary reduction could have occurred as part of the underlying investigation or as part of some subsequent simplification of the statistical model by one of the common statistical reduction methods, sufficiency or conditionality. If the reduction is to a sufficient statistic, then the conditional distribution describing possible antecedent data has no dependence on the parameter and the model for the sufficient statistic is used in place of the original model. If the reduction is by conditionality then there is typically an ancillary variable with a distribution free of the parameter and the given model for the possible original data is replaced by the conditional model given the ancillary (supportive) variable; see also *Statistical Sufficiency*.

With a data value y^0 the fiducial methodology would take the above probability expression and substitute the value y^0 and then treat θ as the variable for the probability statement, thus giving a 95 percent fiducial probability statement $P(\theta \geq y^0 - 1.64\sigma_0; y^0) = 95$ percent. The structural approach would consider the normal $(0; \sigma_0^2)$ distribution for the error $e = y - \theta$ and might record for example the probability $P(e \leq 1.64\sigma_0) = 95$ percent; then with data value y^0 the probability statement would be applied to the error $e = y^0 - \theta$ giving the structural probability $P(y^0 - \theta \leq 1.64\sigma_0; y^0) = 95$ percent or equivalently $P(\theta \geq y^0 - 1.64\sigma_0; y^0) = 95$ percent. In a somewhat related manner the Bayesian methodology might use a uniform prior $cd\theta$ and obtain a posterior distribution for θ that would have $P(\theta \geq y^0 - 1.64\sigma_0; y^0) = 95$ percent. For this simple example the three methods give the same result at the 95 percent level and also at other levels, thus saying essentially that with data value y^0 the probability distribution describing the unknown θ is normal $(y^0; \sigma_0^2)$.

With more complicated models the results from the three methodologies can differ and philosophical arguments concerning substance and relative merits arise. However, for one straightforward generalization the methods remain in agreement: the normal distribution can be replaced by some alternative distribution form; this is discussed in some detail in Sect. 2.

2. Fiducial Probability

Fisher (1922, 1925; see also *Fisher, Ronald A (1890–1962)*) had already introduced most of the

fundamental concepts of statistical theory, such as sufficiency, likelihood, efficiency, exhaustiveness (minimal sufficiency), when he chose (Fisher 1930) to address directly the aspiration mentioned above. He took Laplace and Gauss to task for ‘fall(ing) into error on a question of prime theoretical importance’ by adopting the Bayesian approach that ‘Bayes (had) tentatively wished to postulate in a special case’ and which was published posthumously (Bayes, *ibid*). He then proposed in a restricted context the fiducial method, as discussed.

Neyman and Pearson (1933) then gave a mathematical formulation of fiducial probability that became known as confidence intervals. Fisher (1956) however treated Neyman and Pearson’s formulation as a ‘misconception having some troubling consequences ...’; logical and philosophical arguments between the two sides were intense for many years. In particular the slight to Laplace and Gauss may well have affected the views of the more mathematical participants.

Fisher (1930) entitled his paper ‘Inverse Probability’ and examined a statistic $t(y)$ whose distribution depended on a single parameter θ . Let $P = F(t; \theta)$ be the distribution function of t , and let P itself be what we might now call a p -value for assessing θ ; of course in the usual continuous case P has the uniform distribution on $(0, 1)$. ‘If now we give P any particular value such as 0.95, we have ... the perfectly objective fact that in 5 percent of samples’ t ‘will exceed the 95 percent value corresponding to the actual value of θ ...’. Then to ‘any value of’ t ‘there will moreover be usually a particular value of θ to which it bears this relationship; we may call this the ‘fiducial 5 percent value of θ ’ corresponding to’ the given t . This led to Neyman and Pearson’s (*ibid*) confidence methodology but Fisher treated this as a misconception and he followed different directions and interpretations for the fiducial methodology; for a view on related approaches see *Estimation: Point and Interval*. For this present simple case with scalar t and scalar θ , there seems little difference between the fiducial and the confidence approaches and interpretations.

This discussion effectively ascribes a distribution to θ based on an observed t ; this is called the fiducial distribution for θ . Just as the density of t for given θ is obtained as $(\partial/\partial t)F(t; \theta)$ so also the fiducial density is obtained as $(-\partial/\partial \theta)F(t; \theta)$; the negative sign is inconsequential and is merely the result of $F(t; \theta)$ being examined typically for the case that is increasing with θ . The fiducial distribution for cases like the present can also be called the confidence distribution, which recently has a variety of uses in statistical inference. This alternative name emphasizes its more conventional role and attempts to avoid the inappropriate stigma that surrounds the fiducial concept.

For the fiducial approach in more general contexts, Fisher (1935) recommended a maximum reduction to a statistic $t(y)$ by the use of sufficiency. For example,

with a sample $y_1, \dots, y_n$ from the normal $(\mu; \sigma^2)$ distribution, the reduction would be to $t(y) = (\bar{y}, s_y^2)$. The methodology then suggests the use of a pivotal quantity $p = p(t; \theta)$ with a fixed distribution and a one-one relationship between any two of p, t, θ ; recall more generally that a pivotal quantity is a function of the variable and parameter that provides a measure of departure of variable value from parameter value, and has a fixed distribution which allows an assessment of an observed departure. Thus for the example a natural pivotal is

$$\left\{ z = \frac{\bar{y} - \mu}{\sigma/\sqrt{n}}, \quad \chi_{n-1}^2 = \frac{(n-1)s_y^2}{\sigma^2} \right\}$$

which has independent components, normal $(0, 1)$ and chi-square with $n-1$ degrees of freedom. Fisher (1956, p. 172), however, rather deviously rejected this as a legitimate part of the fiducial methodology, but he was somewhat less explicit about what would be legitimate. This pivotal reduction procedure is now however a rather familiar component of standard inference theory and in particular of confidence theory.

The final step is to invert the pivotal quantity, that is, to insert the observed values for the variables and then transform the distribution of the pivotal quantity to the parameter. For the example this gives

$$\mu = \bar{y} - \{z/(n-1)^{-1/2}\chi_{n-1}^2\}s_y/\sqrt{n}$$

$$\sigma^2 = (n-1)s_y^2/\chi_{n-1}^2$$

where the $\bar{y}, s_y^2$ have their observed values. We can then write

$$\mu = \bar{y} - ts_y/\sqrt{n}$$

where t is Student with $n-1$ degrees of freedom. This fiducial calculation closely parallels that for the confidence approach, except that the limits here are calculated from the Student distribution for μ rather than from the Student distribution of the pivotal t . In quite wide generality fiducial regions can correspond to confidence regions; it is just a matter of whether the limits are calculated before or after the data are observed, a non-issue from the Bayesian viewpoint.

There are however cases where routinely obtained confidence and fiducial regions can differ. For example, consider a sample $(y_{11}, y_{12}), \dots, (y_{n1}, y_{n2})$ from a bivariate normal distribution with means μ_1, μ_2 , variances σ_1^2, σ_2^2 , and correlation ρ , and suppose primary interest centers on the regression coefficient $\beta = \rho\sigma_2/\sigma_1$ of the second variable on the first variable. Standard regression analysis of the second variable on the first variable can produce an interval that differs

from that obtained from the joint fiducial or structural distribution for all the parameters by integrating out the unwanted parameters; see for example, Fraser (1979, pp. 189, 204, 280, and 293). Also there are other familiar cases where fiducial, structural and Bayesian methods can differ; see for example, the Behrens-Fisher and the Fieller-Creasy problems, and for some discussion see Wallace (1980).

Now suppose we generalize and examine a sample $y_1, \dots, y_n$ from some distribution with location μ and scale σ . For example the basic form $f(z)$ of the distribution might be, say, the longer-tailed logistic, or the Student with 6 degrees of freedom often cited as a realistic error pattern. The distribution for the underlying errors is then $f(z_1) \cdots f(z_n)$ and for the sample is

$$\sigma^{-n} \prod f\{\sigma^{-1}(y_i - \mu)\}$$

With a typical general error shape we do not have a simple minimal sufficient statistic as with the normal error pattern above. Fisher (1934) then recommended conditioning on an ancillary statistic and suggested the statistic

$$d_1 = \frac{y_1 - \hat{\mu}}{\hat{\sigma}}, \dots, d_n = \frac{y_n - \hat{\mu}}{\hat{\sigma}}$$

called the configuration statistic, whose coordinates are standardized residuals, standardized here with respect to maximum likelihood values. The conditional distribution of $(\hat{\mu}, \hat{\sigma})$ is fairly easily expressed in terms of likelihood. Let

$$L^0(\mu, \sigma) = c\sigma^{-n} \prod f\{\sigma^{-1}(y_i^0 - \mu)\}$$

be the observed likelihood function from data $(y_1^0, \dots, y_n^0)$, and $(\hat{\mu}^0, \hat{\sigma}^0)$ be the observed maximum likelihood values. Then the conditional distribution given the observed configuration statistic $(d_1^0, \dots, d_n^0)$ is

$$g(\hat{\mu}, \hat{\sigma} | d^0; \mu, \sigma) \frac{d\hat{\mu}d\hat{\sigma}}{\hat{\sigma}^2} = cL^0(\hat{\mu}^0 + (\mu - \hat{\mu})\hat{\sigma}^0/\hat{\sigma}, \sigma\hat{\sigma}^0/\hat{\sigma}) \frac{d\hat{\mu}d\hat{\sigma}}{\hat{\sigma}^2}$$

which can be used to give confidence intervals; also a straightforward fiducial argument following that for the normal case gives

$$L^0(\mu, \sigma) \frac{d\mu d\sigma}{\sigma}$$

for the fiducial distribution. If f is replaced by the standard normal we get the result discussed earlier and

we get it without any reference to sufficiency. This with some other results suggests (Fraser and Reid 2000) that the primary concept sufficiency can be replaced quite widely by appropriate conditioning, and with a major gain in generality of viewpoint.

Fisher (1935) did not use maximum likelihood notation for this location scale model. The notation here adapts to more general contexts (Fraser 1968, 1979) but needs to be modified if the error distribution involves a shape parameter which would then make the maximum likelihood estimates parameter-dependent.

For more general problems the fiducial can have non-uniqueness difficulties, and these tend to arise from the choice of what pivotal to use. For example with the bivariate normal discussed, different pivots can produce different fiducial distributions (Fisher 1956, p. 172, where he elusively denigrates the pivotal approach).

There are also issues connected with finding a marginal fiducial for a component parameter. For a general discussion of marginalization paradoxes connected with methodologies concerning distributions for parameters, including the Bayesian methodology, see David et al. (1973).

For a mathematical analysis that avoids some of the difficulties with the fiducial methodology, see the transformation group framework in Fraser (1961a, 1961b); this mathematical analysis led to the development of the structural approach described below. The marginalization issues still remain, however, and this is related to the common assessment technique of examining a procedure with many repetitions from the same distribution with fixed parameter value. Alternatives to this are suggested in Fraser and Reid (2000).

3. Structural Probability

In many applied problems it is possible to ascribe an underlying error distribution to the model. For example with measurements that are normal with unknown error scaling we might write

$$y_i = \mu + \sigma z_i \quad i = 1, \dots, n$$

where the $z_1, \dots, z_n$, are independent standard normal variables. If we then examine (Fraser 1966) these expressions with observed data we can note rather easily that many characteristics of the underlying realized errors can be numerically evaluated. Specifically, let

$$d_1 = \frac{y_1 - \bar{y}}{s_y}, \dots, d_n = \frac{y_n - \bar{y}}{s_y}$$

be standardized residuals; then the observed $y_1, \dots, y_n$ and the underlying realized $(z_1, \dots, z_n)$ have the same standardized residuals. In fact, $n-2$ characteristics as presented by the residuals are available concerning the

underlying errors. Most theory concerning applications of probability would then say the analysis should be conditional on these observed characteristics, that is, on what you know concerning the realized error (Fraser 1976, Chap. 4, pp. 161–2 and Chap. 11, pp. 456–66). This for quite general problems automatically gives the conditioning used by Fisher and avoids the need to invoke a conditioning principle and seek an ancillary such as the configuration statistic.

This present type of analysis carries through equally (Fraser 1966) if the normal distribution is replaced by some other error pattern. The conditional error distribution can be viewed from an invariance argument as being valid after the full data are observed. This distribution then gives the structural distribution for μ and σ . The mechanics follow closely what is presented for the fiducial case and the resulting structural distribution agrees with the fiducial distribution, $L^0(\mu, \sigma)d\mu d\sigma/\sigma$. The difference here is that a rather strict invariance structure is assumed as part of the argument. This is summarized in general notation.

The structural approach is possible in cases where the underlying error can be directly modeled and the observed response is obtained as some transformation of this. Let Z designate a general error variable with known density $f(Z)$ on say R^n . Let Y designate the response as obtained by a transformation or re-expression θ in some space Ω . Then we have

$$Y = \theta Z, \quad f(Z)dZ,$$

For the invariance we assume that a transformation θ carries R^n into R^n and that the collection Ω of such transformations is closed under composition and inversion. Also for details here we assume that the transformations are smooth and at most one transformation carries a point Z into a point Y . For some general background on transformations and invariance, see *Causal Inference and Statistical Fallacies*.

If we consider all the transformation Ω and apply them to a data point Y^0 we obtain a set ΩY^0 of images, of possible antecedent error values. If Z^0 is the unknown realization of error that produced Y^0 then we have $\Omega Z^0 = \Omega Y^0$ or equivalently we have that Z^0 lies in the set ΩY_0 . This means that a probability assessment should be made conditional on the observed set ΩY^0 . Let $D(Z)$ be a reference point in the set ΩZ and let $[Z]$ in Ω be the transformation in Ω that carries $D(Z)$ to Z . Then we have that the model $Y = \theta Z$ with data Y^0 can be rewritten as

$$[Y] = \theta[Z], \quad D(Y) = D^0,$$

giving the conditioning $D(Y) = D(Y^0) = D^0$ that appears as an observed ancillary in the fiducial framework. The conditional distribution of $[Z]$ given

$D(Z) = D^0$ then describes the reduced model $[Y] = \theta[Z]$ with data $[Y^0]$. The structural distribution of θ is then given as $\theta = [Y^0][Z]^{-1}$ using the conditional distribution $[Z]$ just described.

In most contexts the support measure dZ can be replaced by an invariant measure $dM(Z)$

$$dM(Z) = \frac{dZ}{J(Z)}$$

where $J(Z)$ is an appropriate Jacobian (Fraser 1979). Expressions are simpler if we use density $\bar{f}(Z)$ with respect to the measure $dM(Z)$: $\bar{f}(Z) = f(Z)J(Z)$.

The distributions of $[Y]$ and $[Z]$ conditional on $D(Y) = D(Z) = D^0$ are

$$c\bar{f}(\theta^{-1}[Y]D^0)d\mu[Y], \quad c\bar{f}([Z]D^0)d\mu[Z]$$

where $d\mu[Z]$ is the corresponding invariant measure on the group Ω . The inverted distribution, called the structural distribution for θ , is

$$c\bar{f}(\theta^{-1}[Y^0]D^0)dv(\theta) = cL^0(\theta)dv(\theta)$$

using the right invariant measure ν on the group Ω . In the Bayesian context the typically preferred noninformative prior is the right invariant measure $dv(\theta)$. Thus the structural distribution coincides with this preferred Bayesian posterior.

This group type model covers a wide range of regression multivariate and spherical distribution models but not in the simplified notation used here; for further details see Fraser (1968, 1979).

Recent likelihood asymptotics has been able to obtain many aspects of these results in a general asymptotic context; see Fraser et al. (1999) and Fraser et al. (1999).

4. Comment

Fiducial inference led to confidence intervals, with both having evolved from the Bayesian procedures initiated in the eighteenth century. In terms of freedom of usage, we have Bayesian inference as the more liberal. The foundational issues center on validity and interpretation of the probabilities for the parameters. Here the conditions are strictest for the structural and minimal for the Bayesian. In the nice cases they agree; and in the less standard cases they pressure the proponents to clarify their assumptions.

For some recent discussion of confidence and fiducial interconnections plus references see Barnard (1995), with complementing views in Fraser (1996). Also see Fraser and Reid (2000) for a discussion of the interplay with objective Bayesian methods.

Bibliography

- Barnard G A 1995 Pivotal models and the fiducial argument. *International Statistical Review* **63**: 309–24
- Bayes T 1763 An essay towards solving a problem in the doctrine of chance. *Philosophical Transactions of the Royal Society* **53**: 370–418 [Reprinted (1958) in *Biometrika* **45**: 293–315]
- Dawid A P, Stone M, Zidek J V 1973 *Journal of the Royal Statistical Society* **35**: 189–233
- Fisher R A 1922 On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London* **A222**: 309–68
- Fisher R A 1925 Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society* **22**: 700–25
- Fisher R A 1930 Inverse probability. *Proceedings of the Cambridge Philosophical Society* **26**: 528–35
- Fisher R A 1934 Two new properties of maximum likelihood. *Proceedings of the Royal Society of London* **A144**: 285–307
- Fisher R A 1935 The fiducial argument in statistical inference. *Annals Eugenics* **6**(4): 391–8
- Fisher R A 1956 *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh, UK
- Fraser D A S 1961a The fiducial method and invariance. *Biometrika* **48**: 261–80
- Fraser D A S 1961b On fiducial inference. *Ann. Math. Statist.* **32**: 661–76
- Fraser D A S 1966 Structural probabilities and a generalization. *Biometrika* **53**: 1–9
- Fraser D A S 1968 *The Structure of Inference*. Wiley, New York
- Fraser D A S 1976 *Probability and Statistics, Theory and Applications*. ITS, Toronto
- Fraser D A S 1979 *Inference and Linear Models*. McGraw Hill, New York
- Fraser D A S 1996 Some remarks on pivotal models and the financial argument in relation to structural models. *International Statistical Review* **64**: 23–5
- Fraser D A S, Reid N 2000 Strong matching of frequentist and Bayesian inference. *Journal of Statistical Planning and inference*, to appear
- Fraser D A S, Reid N, Wu J 1999 A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86**: 249–64
- Fraser D A S, Wong A, Wu J 1999 Regression analysis, nonlinear or nonnormal simple and accurate *p* values for likelihood analysis. *Journal of the American Statistical Association* **94**: 1286–95
- Neyman J, Pearson E S 1933 The testing of statistical hypotheses in relation to probabilities a priori. *Proceedings of the Cambridge Philosophical Society* **29**: 492–10
- Wallace D L 1980 The Behrens–Fisher and Creasy–Fieller problems. In: Fienberg S E, Henlele D V (eds.) *R. A. Fisher: An Appreciation*. Springer Verlag, New York, pp. 139–45

D. A. S. Fraser

Field Observational Research in Anthropology and Sociology

Field observational research in anthropology and sociology—or the process of doing ethnography—brings the social science investigator directly to the

scene of the human behavior she or he describes. It involves ‘being there,’ or ‘I-witnessing.’ So also, of course, do other forms of documentation: journalism, film, travel writing, explorer or missionary’s report, even background research by a novelist. What makes ethnography distinctive as a research practice is that it forms one point of a triangular process of knowledge construction which also includes comparative social theory and documentary contextualization.

The fieldnotes that result from ethnographic discovery procedures are not merely ‘written up,’ but are filtered and interpreted against theoretical propositions and the comparative record of other field observational studies; they are also grounded and enhanced by historical, ecological, demographic, economic, and other documentary sources that provide background and context. As the products of this mode of research are read—monographs and articles also referred to as ethnography—they stimulate new comparative theoretical thinking, which in turn suggests further problems and interpretations to be resolved through more field observational research. Ethnographies also regularly lead to new demands and rising standards for documentary contextualization—more history, more ecological or demographic backgrounding, more attention to state policy, economic trends or the world system. This triangle of ethnography, comparison, and contextualization is, in essence, the process by which field observational research in anthropology and sociology is utilized to explain and interpret human cultures and social life.

1. History and Scope in Anthropology and Sociology

In 1851 Morgan published his *League of the Ho-de-no-sau-nee, or Iroquois*; combining fieldwork and comparative and theoretical interests in political organization, this book was the first anthropological ethnography. It depicted the structure and operation of Iroquois society, detailing matrilineal kinship, political and ceremonial life, material culture and religion. After formation of the US government Bureau of American Ethnology in 1879, a stream of ethnographic accounts of native American societies began, including landmark field observational studies of the Zuni by Cushing, the Baffin Island Inuit by Boas, and, after 1900, ethnographies of Plains, Eastern Woodland, and California groups by Boas’ students trained at Columbia University. During the 1930s the scope broadened to include rural and urban studies in Latin America, the Caribbean, Africa, and Europe, and also in contemporary US settings.

In Great Britain’s colonies, local government official or missionary fieldworkers followed the topical guidebook *Notes and Queries on Anthropology*, drafted by Oxford University theorist Tylor and others, and first published in 1874. This armchair scholar/man-on-the-

spot division of labor ended two decades later when Oxford-trained Spencer collaborated in field observational study with local expert Gillen, producing their *Central Tribes of Native Australia* in 1899, and providing detailed coverage of cosmology, ritual, and social organization which revealed unheralded complexity, and which sparked theoretical work by Durkheim and Freud. In 1898 a team from Cambridge University conducted fieldwork in Torres Strait north of Australia, and member Rivers used genealogical data to probe community history, migration trajectories, marriage patterns, demography, and inheritance. Rivers and his students conducted further ethnographic studies in India and the Pacific, but their work was eclipsed by the more thorough Trobriand Islands field observational research of Malinowski, resulting in *Argonauts of the Western Pacific* in 1922 and its famous first chapter on method. Malinowskian ethnography dominated British social anthropology thereafter, and by the 1940s reshaped American cultural anthropology as well.

In sociology, there is less agreement about the disciplinary origin of field observational research methods. Conventionally, the monographs of the University of Chicago department during the 1920s and 1930s are cited (Madge 1962): studies of itinerant 'hobo' workers, youth gangs, dancehall employees and patrons; Italians, Poles, Jews, French Canadians, Mexicans, and Chinese; and central city 'gold coast' apartment, artist colony, and rooming house dwellers. Two decades before this research, however, Du Bois used similar methods (both he and the Chicago school were inspired by British and American urban survey and settlement house movements) to produce *The Philadelphia Negro* (1899), which depicted life in an inner city neighborhood comprised of southern migrant and northern-born African-Americans. Even earlier, Engels in 1844 studied new classes, Irish labor migration, urban spatial arrangements, symbolic attachments, and modes of observable behavior in Manchester and other rapidly industrializing British cities, although his *Condition of the Working Class in England* was not published in English until 1892.

Since the 1920s there has been much cross-fertilization between anthropological and sociological ethnography. A few of the Chicago studies were conducted by researchers identifiable as anthropologists, and the Lynds, authors of *Middletown* (1929) which credits anthropologists Rivers and Wissler for methodological inspiration, and Whyte, author of *Street Corner Society* (1955 [1943]) and influenced by Harvard anthropologists, are presented as major figures in the development of American sociology (Madge 1962). Gluckman, a Manchester social anthropologist whose fieldwork was conducted in Africa, was influential in the development of ethnographic fieldwork by British sociologists working in schools, gangs, and factories during the 1950s and later (Burgess 1984). In recent decades, both anthropol-

ogists and sociologists have undertaken field observational research in a wide range of overlapping settings, at home and abroad; so increasingly do scholars from folklore, education, environmental psychology, political science, and other disciplines. Ethnography, however, remains central to credentialing, teaching, publishing, and theory-construction in anthropology, while it occupies a more marginal position in contemporary sociology, or in other fields.

2. Selection, Access, Rapport

The selection of a particular population or site for field observational research is ordinarily related to some unanswered question or outstanding problem in social theory. Personal connections or predilections of researchers also shape this selection, but the fieldworker must justify his or her choice in terms of some significant theory to which the project is addressed. Usually this is made explicit in a written proposal for funds to underwrite the fieldwork.

While ethnography is thus lodged from conception in social theory, a comparative viewpoint also molds the fieldwork process. First, investigators acquire a cross-cultural or cross-site perspective by training and reading; when addressing any aspect of social life—marriage, leadership, ethnicity, etc.—field observational researchers recall examples of similarities and differences elsewhere. Second, the comparative perspective focuses attention on trends and transitions, not just on similarities and differences at random (which are infinite). Rather than treating each fieldwork case as unique, ethnographers place the social phenomena they study within comparative frames (foraging, horticultural, agricultural, pastoral, industrial, colonial, neocolonial economies; cooperative, competitive, individualistic societies; gender subordination, complementarity or equality; etc.) Ethnographic results in turn provoke debate, revision, and innovation in such theorizing.

While significant theories bring fieldworkers to particular locations, actors, and activities, once they arrive they begin to observe and listen. Often they must learn to listen—by learning the language, the local vocabulary, and current verbal conventions. Field observational research now turns away from theoretical discourse to the viewpoints and concepts of the people (informants, subjects, actors, consultants) themselves. Ethnographers aim to document how the people see and talk about their everyday social activities and groupings, and about the wider worlds they live in. It is their normal scenes of activity, topics of conversation, and standards of evaluation that are the objects of field observational research.

This does not begin by announcing: 'I'm your ethnographer; when can I interview you?' Investigators must be honest about their role and sponsorship (covert field observational research remains contro-

versal and rare), but their paramount aim is to listen and to move as quickly as possible into the natural settings of social life, the places people would be, doing what they would be doing, if the ethnographer was not there. This involves a continuing process of requesting and obtaining access to the many locations, activities, and groups that comprise the overall fieldwork arena. As ethnographers observe and listen in a wide-ranging manner, within parameters set by the significant theories that bring them there, they begin to understand culturally meaningful behavioral conventions, and to formulate culturally appropriate utterances. This indexes a second continuing process of building interpersonal rapport with informants, a topic that preoccupies many autobiographical accounts of field observational research—one of the most instructive being Whyte's 1955 [1943] Appendix to *Street Corner Society*.

As the ethnographic process unfolds, the fieldworker must constantly make decisions about where to be, whom to listen to, what events to follow, and what to safely ignore and leave out. These decisions are guided both by the significant theories prefiguring fieldwork, and by new theories of significance that arise in the field. These latter theories—hunches, hypotheses, ideas about connections and relationships—emerge as listening and observation proceed. They suggest which persons and activities to focus upon, what places and events to attend to, and which objects and their circulation to follow. As this occurs, the fieldwork 'funnel' narrows, to use Agar's (1980) apt metaphor. The early phase of securing access and building rapport is wide, open, and nearly all-encompassing; as theories of significance emerge, pan out, or are discarded, the funnel of informants, events, and activities narrows. Goals crystallize. Research design sharpens.

3. *Discovery Procedures: Listening*

Ethnography's greatest strength is situated listening. Interviews become useful at later stages of fieldwork, but field observational research begins by listening to what British anthropologist Richards (1939) called 'speech in action.' Here the actors control topicality—talking to each other about what they usually do—and the researcher secures access to their turf—the locations they usually occupy. Early on, as rapport begins to be established, fieldworkers deliberately place themselves in a wide sampling of such situations. Then, as the research funnel narrows and other methods are deployed, the ethnographer becomes more selective about where to listen.

The importance of situated listening was emphasized memorably by Whyte. After he asked an intrusive question during illicit gambling activities, and stopping the ongoing flow of behavior, Whyte's key

informant Doc told him, 'Go easy on that "who," "what," "why," "when," "where" stuff, Bill. You ask those questions and people will clam up on you. If people accept you, you can just hang around, and you'll learn the answers in the long run without even having to ask the questions.' 'I found that this was true. As I sat and listened, I learned the answers to question that I would not even have had the sense to ask if I had been getting my information solely on an interview basis' (1955 [1943], p. 303).

Situated listening becomes the basis for understanding both what can be talked about in later conversations and interviews, and what categories and evaluations the actors use to observe and interpret ongoing behavior.

4. *Discovery Procedures: Observing*

Many veteran field observational researchers bristle when they are told that their methods are not quantitative. Using both emic categories articulated by the actors, and etic categories formulated by the community of social science observers, ethnographers count, sample, and calculate ratios, proportions, and probabilities as part of their fieldwork activities (Johnson 1978). These countable quantities include numbers of people in events, their positions, their comings and goings; objects, inventories; and exchanges, movements, orderings, sequences, associations, assemblages, and arrangements of many sorts.

Malinowski's classic *Argonauts of the Western Pacific* (1922) incorporated his numerous observations of the interisland *kula* exchange of seashell valuables—necklaces traded for armlets—as canoes from one community arrived at another, exchanges were conducted, and then additional exchanges of items including foodstuffs, timber, greenstone, and pottery followed. Who exchanged what, the order, etiquette, volume, frequency, and other variables were all counted and analyzed. In Whyte's *Street Corner Society* (1955 [1943]) the size, composition, duration of activity, and initiation of interaction were recorded and analyzed for numerous group events; Whyte even demonstrated quantitatively a relationship between hierarchical group leadership structure and bowling scores.

Harris (1964) furnishes field observational researchers with an etic model of 'the great stream of uninterrupted behavior' that extends backward in time and unfolds continuously. He provides a technical vocabulary to operationalize recording of isolatable events and linked segments of the behavior stream—from individual body movements and their environmental effects at the most basic level, through sequences of multiactor scenes connected by actor, place, or object. Sanjek (1978) applied this approach in a multiethnic Ghanaian urban neighborhood to

collect (through interviews) sequences of multiactor scenes for 40 adults, each over four-day periods, in order to test whether their social networks were structured more by common ethnicity or by class position, and in which behavioral settings ethnicity and class were stronger or weaker.

Johnson (1975, 1978) developed an observational fieldwork approach to formulate statistically valid time allocation patterns in his research among the Machiguenga of Peru. By randomly selecting actors and times for observation at their most likely location, and recording their ongoing behavior at the time of each random visit, a powerful quantitative record of how a given population allocates its activities among production, reproduction, leisure, and other categories resulted. This and similar methods have been employed in a wide range of fieldwork settings (Gross 1984).

5. Discovery Procedures: Interviewing

Barth recounts of his highland New Guinea research, 'My field procedure was ... to rely very heavily on observation of spontaneous, unelicited word and act. I paid particular attention to this in Baktaman questions and explanations *to each other*—in men's houses, in the fellowship of hunting and working, during initiations and other rites. Only secondarily, and with much reticence, did I ask my own questions, and then only when they seemed cast in a native pattern and received easy and natural response' (1975, p. 226).

Interviewing is problematic (Briggs 1986). Human beings are apt to reinterpret or reformulate the past to make it conform with their ongoing sense of the present. Moreover, the ever present danger of an observer effect in fieldwork—that the presence of the observer affects and reshapes an actor's behavior and response—is especially difficult to control in an artificial one-to-one interview situation where normal social activity is suspended. Still, interviews are essential to field observational research. They are best left for later stages of the ethnographic encounter when, to use Agar's funnel metaphor, the speech events of fieldwork move from wide and open to narrow and more focused.

After the investigator has become at ease with the language and local codes and references, and has successfully gained access to natural conversations and amassed a stock of cultural knowledge via situated listening, deliberate intervention into speech events may begin. These steps may be graded from least to most intrusive and controlled.

(a) Still on the informants' turf, and still in the accustomed activities of daily life, the field observational researcher begins to enter natural conversations as a speaker, and begins to shift topicality to

her or his own interests. This process starts gently, by moving appropriately into rounds of chatting, gossiping, and ordinary comment. As cultural competence increases, and as theories of significance emerge, the fieldworker attempts to direct conversations by introducing questions and suggesting topics for responses from informants.

(b) After an initial period of fieldwork—a few months perhaps—arranged interviews may begin. This class of speech events is disruptive—the informant is removed from her or his turf, either to the ethnographer's household or office, or by transforming an everyday location into a scene of ethnographer–informant dialogue—an activity that would otherwise not be occurring there. Typically the earliest of these deliberate breaks in time–place behavioral flow reserve topicality for the actor. In such open-ended or discovery interviews, the informant is encouraged to move the conversation according to his or her own interests.

(c) In later and more productive interviews, the ethnographer begins to assert control. General topics are introduced, allowing informants to expand freely upon their own points of view and knowledge. In more structured ethnographic interviews, topicality is more firmly shaped and directed by the fieldworker; informant responses now move away from orations and free commentaries and to more specific responses to questions.

(d) In the most focused forms of interviewing, the ethnographer controls both turf and topicality as fully as possible. Questionnaires and interview schedules may be used, and the objective is to obtain particular types and pieces of data. These may include household interviews, psychological tests, or reports of disputes, or may encompass repeated interview sessions to secure lengthy life histories, with the fieldworker now guiding the subject according to preset standards of scope and comprehensiveness.

(e) Though not a major part of ethnographic practice, in some instances, and when still on the informants' turf, the fieldworker may ask direct and pointed questions, and attempt to secure precise pieces of data for his or her records. Interventions of this sort are especially dangerous—the inappropriateness of such seizures of topicality in everyday settings may be jarring to the actors. Typically speech events of this sort occur in the final days of fieldwork, when local rapport is at its peak, research goals are most pressing, and the fieldwork funnel approaches its vanishing point.

6. Notes and Records

The production of notes and records (Sanjek 1990) begins to move the field observational research process toward its ultimate written products. Focused in-

interview sessions with seated informants often permit direct transcription of verbal statements, but in open-ended and ethnographic interviews, brief written notes—what Ottenberg (in Sanjek 1990) terms scratch notes—are taken during the session, and these form the basis for the construction later of fuller written fieldnotes. Fieldworkers often go through this two-step process even when interviews are tape recorded, both as a backup to and index of the taped session, and because of the analytic gains many ethnographers note in transforming their scratch notes into fuller descriptive fieldnotes.

In field observational research in natural behavior stream settings, similar brief jottings may be inscribed, but major attention is directed to the event in progress. Often it is not even possible to record scratch notes, and both they and fuller fieldnote description occur later. Mead (1977) wrote about the nagging pressure to type up fieldnotes from scratch notes, and about the danger of scratch notes growing 'cold' when this was delayed, even by one day. But she also wrote of the satisfaction of being caught up with this work, and of the importance for later ethnographic writing of the insights gained in moving from scratch notes to descriptive fieldnotes.

Ottenberg sees this process as the interaction of scratch notes and headnotes—the stored memories and interpretations that arise from direct participant observation as filtered by the ethnographer's overall theoretical stance. Headnotes form an essential complement to fieldnotes (and to more formal fieldwork datasets or records). Headnotes are employed to make sense of one's fieldnotes when they are re-read later for ethnographic writing projects. The importance of headnotes is particularly evident when one ethnographer attempts to use another's fieldnotes, and quickly realizes how difficult it is to understand them without headnotes of one's own (Sanjek 1990).

Fieldnotes typically are kept in running chronological order, and may minimally be indexed separately by actor and topic. Additionally, field observational researchers often generate records, or well-defined sets of data organized by topic, person, or other category. These may include records of household composition, land holdings, ritual performances, life histories, folktales, etc. Sometimes decisions to collect sets of records are part of the initial research design, and preset forms to enter such data may be brought with the investigator to the field setting. In other cases, a decision to collect a systematic body of records develops during fieldwork, often emerging within general running fieldnotes and then being separated out in new record files.

Fieldnotes and records present ethnographers with great masses of information—hundreds, even thousands of pages—that while being accumulated may be arranged only by chronological order or by topic. Malinowski (1922) urged that ethnographers constantly re-read and begin to organize and analyze their

notes while still conducting field research, but commonly the more focused work of indexing them occurs when fieldwork is over. As ethnographers then turn to ethnographic writing, they readdress the theoretical discourse they turned away from in fieldwork. Fieldnotes and records must now be related directly to the comparative and contextual points of the knowledge-construction triangle.

7. *Canons of Validity*

Field observational research aims to maximize validity—'the degree to which scientific observations actually measure or record what they purport to measure'—rather than reliability—'the repeatability, including interpersonal replicability, of scientific observations' (Pelto and Pelto 1978, p. 33). Reliability is valued in survey, medical, and product safety research. In ethnography, restudies to assess change may be conducted, but independent investigation of the same locale or population is unlikely or impossible.

Validity depends on the intensity and depth of field observational research, but it always remains a goal rather than a property that can be independently measured (Wolcott 1995, pp. 167–70). A quantitative, and highly reliable, randomized sample study that investigates the gender, ethnicity, class, and other characteristics of the five best friends of members of a given population would be less valid than one in which cultural meanings and categories of friendship were first discovered through listening to natural conversations or eliciting such information in interviews. In the case of Indian immigrants in New York City, for example, the 'five best friends' study might miss entirely the category of *rakhi* ritual friendships between adult males and females which are affirmed in an annual festival where the man ties a flower bracelet on the woman's wrist.

If validity remains important, but elusive, how can it be assessed in field observational research results? Sanjek (1990, pp. 393–404) proposed three canons of validity by which ethnographic writing may be evaluated. The first is theoretical candor, the openness with which the ethnographer addresses the significant theories, and even more, the local theories of significance that structured the fieldwork process. A second canon calls for explicit depiction of the ethnographer's fieldwork path—the number of informants from whom information was obtained, in what ways, and their relationship both to the wider population the field observational study concerns and to each other. A third canon concerns information about the fieldnote evidence itself: not simply 'how much' and its basis in listening, observing, or interviewing, but more significantly the precise relationship of notes and records to the written ethnography. Some ethnographies utilize fieldnotes directly, even masses of them; others, for rhetorical or narrative purposes, do

not, and need not. What matters in the end is that readers of field observational research have a clear picture of what the ethnographer did and why, who they talked to and learned from, and what they brought back to document it.

Field observational research produces results that can be obtained in no other way. This ethnographic tradition entails that 'the description of people's activities, their interactions with each other, [and] their verbal behavior' be 'copious and detailed'; 'cumulative development ... consists above all in the fact that such data, provided they are adequate, never become obsolete and can be used by analysts ... in researching questions not visualized by the researchers who collected them' (Anglin 1979, p. 49).

See also: Fieldwork in Social and Cultural Anthropology; Observational Studies: Overview; Participant Observation

Bibliography

- Agar M 1980 *The Professional Stranger: An Informal Introduction to Ethnography*. Academic Press, New York
- Anglin A 1979 Analytic and folk models: The Tallensi case. In: Holy L (ed.) *Segmentary Lineage Systems Reconsidered*. Department of Social Anthropology, The Queen's University of Belfast, Ireland
- Barth F 1975 *Ritual and Knowledge Among the Baktaman of New Guinea*. Yale University Press, New Haven, CT
- Briggs C L 1986 *Learning How to Ask: A Sociolinguistic Appraisal of the Role of the Interview in Social Science Research*. Cambridge University Press, New York
- Burgess R G 1984 *In the Field: An Introduction to Field Research*. George Allen and Unwin, London
- Emerson R M, Fretz R I, Shaw L L 1995 *Writing Ethnographic Fieldnotes*. University of Chicago Press, Chicago
- Gross D R 1984 Time allocation: A tool for the study of cultural behavior. *Annual Review of Anthropology* 13: 519–58
- Harris M 1964 *The Nature of Cultural Things*. Random House, New York
- Johnson A 1975 Time allocation in a Machiguenga community. *Ethnology* 14: 301–10
- Johnson A W 1978 *Quantification in Cultural Anthropology: An Introduction to Research Design*. Stanford University Press, Stanford, CA
- Madge J 1962 *The Origins of Scientific Sociology*. Tavistock, London
- Malinowski B 1922 *Argonauts of the Western Pacific*. Routledge, London
- Mead M 1977 *Letters From the Field, 1925–1975*. Harper and Row, New York
- Pelto P J, Pelto G H 1978 *Anthropological Research: The Structure of Inquiry*. Cambridge University Press, New York
- Richards A 1939 The development of field work methods in social anthropology. In: Barlett F C et al. (eds.) *The Study of Society*. Routledge and Kegan Paul, London
- Sanjek R 1978 A network method and its uses in urban ethnography. *Human Organization* 37: 257–68
- Sanjek R (ed.) 1990 *Fieldnotes: The Makings of Anthropology*. Cornell University Press, Ithaca, NY

Whyte W F 1955 [1943] *Street Corner Society: The Social Structure of an Italian Slum*, 2nd edn. University of Chicago Press, Chicago

Wolcott H F 1995 *The Art of Fieldwork*. AltaMira, Walnut Creek, CA

R. Sanjek

Fieldwork: Ethical Aspects

Fieldwork involves a range of methodologies including ethnographic participant observation, in-depth interviews, direct observation of individual or group behavior, or a combination of both qualitative and quantitative approaches (Denzin and Lincoln 1998). Social scientists conducting fieldwork confront unique ethical dilemmas in the course of implementing their investigations, particularly when researchers are actively engaged with participants in their natural environments. Ethical dilemmas occur when there are conflicts about what is considered to be an appropriate response to a situation involving the investigators, research participants, or the production and use of research materials. A full treatment of these issues is beyond the scope of this paper. This article briefly explores ethical issues associated with informed consent, the protection of confidentiality, the development of personal relationships with research participants, involvement in illegal activities, and intervention in on-going social activities. Finally, a framework for resolving ethical dilemmas in fieldwork is outlined.

1. Informed Consent

National and international guidelines for ethical conduct in research with human subjects identify specific requirements for informed consent: the provision of information, comprehension of information, and voluntary participation (Faden and Beauchamp 1986, Weissstub 1998). Voluntary participation in research depends upon the individual's ability to understand not just the meaning of the research, but the impact it may have on his or her life. Excessive financial compensation or unrealistic promises may constitute coercion, especially if someone is vulnerable because of social factors such as poverty or ethnic background. Voluntary participation may be inhibited when there are social status and social class differences between fieldworkers and research participants—differences based on gender, age, education, income level, religious affiliation, and ethnic identity. The institutional affiliation of the field researcher may also pose a threat to participants. Additionally, the fluid and unbounded nature of many social science field studies complicates

the straightforward approach to informed consent associated with other forms of social science methodologies.

A number of factors influence informed consent in fieldwork—the specific goals of the research, the cultural context of the research project, and communication issues that influence comprehension of information. In international settings, the challenges associated with implementing informed consent in fieldwork are heightened because of language barriers that may inhibit effective communication, particularly regarding the translation of scientific concepts. Misunderstandings and miscommunication about social science research are more likely to occur when fieldworkers and participants speak different languages. Language barriers may be reduced through the use of an interpreter. However, translators must negotiate not only language, but also cultural and contextual factors. Some of the problems associated with interpretation in social science fieldwork include: the inability to translate equivalent expressions easily across languages; paraphrasing that results in omissions or erroneous substitutions of terms; varying levels of comprehension; and the influence of conflicting cultural beliefs and values concerning the meaning and purpose of research.

Differences in beliefs about who has the authority to give permission to participate in research also influence the consent process. Beliefs about personhood, individual autonomy, and decisional capacity are embedded within the social and cultural patterns of family ties and community obligations. In Western industrialized nations, where personal autonomy is emphasized, research participants are expected to make decisions for themselves or through designated surrogates. However, in many non-Western settings, religious or tribal leaders or a person's extended family may play a significant role in decisions concerning research participation. Cultural differences in beliefs about the nature of personhood and the location of decisional authority for consent have been problematic for social science investigators conducting fieldwork in international settings (Levine 1991). Negotiating informed consent with designated authorities in non-Western populations requires fieldworkers to move beyond narrow definitions of personhood, autonomy and self-determination. While respect for persons demands cultural sensitivity to local customs, social scientists conducting fieldwork are never authorized to conduct research without regard to potential harm and without attempts to be informative throughout the duration of the study (Levine 1996). Moreover, although a tribal chief, village elder, or community leader may express approval of a research agenda, sensitivity to cultural customs is secondary to honoring individual choice regarding participation (World Health Organization and the Council for International Organizations of Medical Sciences 1993).

1.1 Informed Consent in Observations of Group Interactions

In fieldwork involving direct observation of group activities, arrangements usually are made prior to initiating the study to inform group members that the fieldworker will be present in the course of routine activities. In closed systems such as a hospital unit, a community center, or other institutional setting, informed consent should be obtained from all those who are at the facility on a regular basis. Individuals who are not present on a regular basis but whose behavior may be observed in public group activities should be alerted to the presence of the fieldworker if it is feasible to do so. In some group observations conducted in the course of fieldwork, it may not be possible or necessary to obtain informed consent from every person present. For example, at informal gatherings such as family or community events, or in public places where individuals interact informally, it would be intrusive for the fieldworker to explain the study to everyone present.

2. Protection of Confidentiality

Protection of confidentiality and privacy for research participants is an important aspect of fieldwork (Cassell and Jacobs 1987). Fieldwork in social science investigations often involves the collection of sensitive information concerning an individual's personal life, including information on physical conditions, mental illness, or behavior regarding illegal activities. Individuals or communities could be harmed if information was inadvertently or purposely disclosed. In some cases, information could be used to ostracize or marginalize research participants, particularly if research participants suffer from a stigmatizing illness such as HIV/AIDS, or if they are involved in activities that may be viewed as socially deviant.

Data collection strategies associated with fieldwork require careful attention to the representation of individuals and communities in written transcripts derived from interviews and observations (Fluehr-Lobban 1991). Descriptive data recorded in field notes, daily logs, or diaries should be entered using a coding system that protects the confidentiality of the research participants. Additionally, field data should also be secured outside of the research setting in order to safeguard the identity of participants. Pseudonyms, rather than the names of participants, must be used in published manuscripts that include case narratives. This gesture, however, will not necessarily protect confidentiality because it may be possible to identify the individuals involved based on the details of the case and the research setting. Nevertheless, at each stage of the research—data collection, data analysis, and data reporting—every effort should be made to protect the privacy of the study participants. In

obtaining informed consent, investigators are obliged to advise participants in fieldwork about the methods designed to protect their privacy.

3. Relationships with Research Participants

Social scientists involved in fieldwork have the opportunity to form close personal relationships with research participants because of the collaborative nature of ethnographic approaches such as participant observation. The development of trust between fieldworkers and the individuals involved in the research enhances the potential for collecting important information relevant to the study. However, in some circumstances, investigators may find themselves in compromising situations because of their personal relationships in the field.

3.1 Intimate Relationships

Although it is rarely addressed in the literature (for an exception, see Kulick and Willson 1995), fieldworkers may become intimately involved with research participants. Sexual relationships with field participants are more likely to occur in the context of extended fieldwork. Traditional anthropological fieldwork—where the investigator may spend months or years in an environment far removed from familiar social networks (often in a different country)—represents the type of setting that contributes to the development of intimate personal relationships with research participants. Despite what some might view as the natural inclination to form a sexual attachment with a close companion, the prevailing opinion among social and behavioral scientists is that sex with participants in the field should be avoided. Sexual relationships between researchers and study participants may diminish the relative objectivity of the investigator.

More problematic is the potential for coercion of socially vulnerable individuals. Moreover, attitudes concerning sexual behavior are socially constructed and the meanings attached to sexual intimacy may vary significantly between the fieldworker's home community and the research setting (Singer et al. 1999). The social complications that can result from a sexual relationship (e.g., jealousy, accusations of infidelity, gossip) jeopardize the integrity of the research and place the investigator at risk for problematic interactions with the broader community.

3.2 Illegal Activities

Investigators conducting fieldwork with populations such as illegal immigrants, or with individuals involved

in activities such as drug use or the sex trade, may be asked to participate in behavior that is legally questionable. For example, fieldworkers investigating the use of illegal drugs may be asked to share drugs with study participants or to assist in procuring or holding drugs or drug paraphernalia. While some researchers may believe that participation in these types of activities strengthens the bonds of trust between a fieldworker and research participants and thus increases the likelihood of securing information that is meaningful and necessary to achieve the study goals, there are strong reasons for avoiding participation in illicit or illegal behavior. As with sexual involvement with study participants, engaging in illegal activities seriously threatens the integrity of study findings and places the researcher at risk socially, professionally, and in some cases, physically. Fieldworkers may find themselves targeted for blackmail or other forms of coercion. Additionally, fieldworkers who participate in illegal behavior diminish the potential to secure funding for future research. A full consideration of the ethical dimensions of involvement in illegal activities cannot be explored in this essay.

4. Fieldworker Intervention in Ongoing Social Activities

Questions concerning the intervention of a fieldworker in problematic situations that occur in the course of conducting research are challenging. The fieldworker's decision should be a judgment based on the purpose of the study, the context, and the specific event that occurs. When investigators continually intervene in field situations that are directly related to the issue being investigated, the possibility that their research will have an impact is jeopardized since they would be unable to document the events being studied. For example, fieldwork with injection drug users to investigate HIV risk behavior would be inconclusive if researchers attempted to stop practices known to facilitate the transmission of the virus. The fundamental problem is one of advocacy—when to be an advocate for research participants and when to maintain the more circumscribed role of the investigator (Bosk 1985). In general, non-intervention is the accepted mode for fieldwork.

5. Recommendations for Resolving Ethical Dilemmas

Investigators who confront moral dilemmas in the course of conducting fieldwork must consider carefully the full range of issues involved in the problem. A systematic approach to ethical decision-making begins with a robust description of the research dilemma.

This would include details about the purpose of the study, the research design, the sponsors of the study, and the individual participants and community involved in the research. The researcher then must consider the cultural and social values represented in the ethical problem. The relevant values of the study participants, the study community, the professional community, and the study sponsors should be outlined. The researcher must determine the principal value conflict. At this point in the process of decision-making, the fieldworker should consider whose values are threatened and who are the most vulnerable to potential harms. This process of reflection facilitates the identification of the key issues involved in the ethical dilemma. When the primary ethical issues are determined, the investigator should outline a full range of strategies and consider the potential risks and benefits associated with each solution. The decision regarding a course of action should maximize respect for the individual and group values identified. The vulnerability of research participants and the communities within which they live should be of paramount importance in resolving the ethical dilemma encountered.

See also: Experimenter and Subject Artifacts: Methodology; Fieldwork in Social and Cultural Anthropology; Linguistic Fieldwork; Research Conduct: Ethical Codes; Research Ethics, Cross-cultural Dimensions of; Research Ethics: Research; Research Subjects, Informed and Implied Consent of

Bibliography

- Bosk C 1985 The fieldworker as watcher and witness. *Hastings Center Report* 15: 10–14
- Cassell J, Jacobs S (eds.) 1987 *Handbook on Ethical Issues in Anthropology*. Special Publication of the American Anthropological Association, Number 23, Washington, DC
- Denzin N K, Lincoln Y S (eds.) 1998 *Collecting and Interpreting Qualitative Materials*. Sage Publications, Thousand Oaks, CA
- Faden R R, Beauchamp T L 1986 *A History and Theory of Informed Consent*. Oxford University Press, New York
- Fluehr-Lobban C (ed.) 1991 *Ethics and The Profession of Anthropology*. University of Pennsylvania Press, Philadelphia PA
- Kulick D, Willson M 1995 *Taboo: Sex, Identity and Erotic Subjectivity in Anthropological Fieldwork*. Routledge, New York
- Levine R J 1991 Informed consent: Some challenges to the universal validity of the Western model. *Law, Medicine, and Health Care* 19: 3–4
- Levine R J 1996 International codes and guidelines for research ethics: A critical appraisal. In: Vanderpool H (ed.) *The Ethics of Research Involving Human Subjects: Facing the 21st Century*. University Publishing Group, Frederick, MD
- Singer M, Marshall P, Trotter R T, Schensul J J, Weeks M R, Simmons J E, Radda K E 1999 Ethics, ethnography, drug use, and AIDS: Dilemmas in federally funded research. In: Singer M, Marshall P, Clatts M (eds.) *Integrating Cultural, Observa-*

tional, and Epidemiological Approaches in The Prevention of Drug Abuse and HIV/AIDS. US Department of Health and Human Services, National Institutes of Health Publication 99–4565, Washington, DC

Weisstub D N (ed.) 1998 *Research on Human Subjects: Ethics, Law and Social Policy*. Pergamon, Oxford, UK

World Health Organization and Council for International Organizations of Medical Sciences (WHO-CIOMS) 1993 *International Ethical Guidelines for Biomedical Research Involving Human Subjects*.

P. Marshall

Fieldwork in Social and Cultural Anthropology

1. Origins of a Method

Fieldwork is a term which has been employed for nearly a century by social/cultural anthropologists, for their major methodological tool and a profound professional experience that leaves its mark on their lives throughout their careers. It designates a complete concentration on observations in their chosen site of research, for a considerable period of time, usually of no less than six months and up to a few years. Originally, that engagement involved a total separation from the researcher's ordinary life, and his/her immersion into a remote and alien social environment. Anthropological fieldwork is known also as the 'ethnographic method.'

The idea of fieldwork owes its origin to the interests of scientists at the turn of the nineteenth century concerning the evolution of the human species, as much as to the expansion of European colonialism that brought under its rule vast territories inhabited by unknown 'exotic' peoples. But no less, that movement coincided with a growing awareness in the United States about the vanishing cultures of native Americans. The new science of anthropology that intended to satisfy these interests was at its first stages mostly speculative, based on secondary sources of data, collected by missionaries, traders, colonial administrators, travelers, adventurers, etc. The practice of pre-fieldwork anthropology was nicknamed 'arm-chair anthropology.' Its most celebrated, as well as discredited representative, remains James Frazer whose monumental work was all based on library material.

The invention of fieldwork as the characteristic method of anthropology, as well as of its corollary aspect of 'participant observations,' has often been attributed to Bronislaw Malinowski, the British–Polish scholar, who is also considered among the founders of the school of functionalism in anthropology. His claim to fame as fieldworker is based on his long stay, of more than two years, during World

War I, among the Trobriand Islanders. Although he was not the first to employ the method engaging participant observation, nevertheless, his linguistic skills, the meticulous ethnographies he produced, and the introductory chapter on his method of inquiry that opened his most famous work (*Argonauts of the Western Pacific* 1922) have become the hallmark and quintessential model of fieldwork for almost half a century.

The fieldwork projects carried out by Malinowski, his colleagues, and the later generations of younger anthropologists entailed observations of the minutiae of everyday life in the various domains of social behavior and cultural traditions. In sum, it was supposed to record the totality of the social experience in the studied society. The essence of fieldwork was encapsulated, for example, in the following statements by representatives of three generations in British anthropology:

This goal is, briefly, to grasp the native's point of view, his relation to life, to realize *his* vision of *his* world (Malinowski 1922, p. 25).

First, in my estimation, is the point that fieldwork in the empirical mode remains the *sine qua non* both for the testing of theory and, what is more important, for making new discoveries (Fortes 1978, p. 24).

'Good' fieldworkers are those who are prepared to have good 'trips,' that is, suspend as far as possible their own social conditioning in order to have sensory and mental knowledge of what is really happening around and to them (Turner 1985, pp. 205–6).

In the United States, the method of ethnographic fieldwork was initiated by Franz Boas and his students most of whom had until World War II concentrated their work in the not too far away reservations of Indian tribes. Their trips to the fields of research had rarely involved a long departure from their ordinary life and work. They were also far more dependent on key informants who provided them with data on specific subjects. Consequently, they were not as deeply engaged in the routine of everyday life of the people they studied. The ethnographies by Ruth Benedict are good examples of the achievements and deficiencies of the tradition of short fieldwork trips and reliance on informants. Margaret Mead, probably the most celebrated among Boas' students, who went to Samoa in a style of research closer to the British tradition (1928), was later seriously criticized for her poor work as fieldworker. It was mostly since the mid-1940s that American anthropologists have been equally engaged in intensive fieldwork projects away from home. Compared with the British tradition, American anthropologists, since the early stages, have developed a distinctive orientation in their fieldwork projects, more broadly directed toward inquiries of culture and personality (represented in later years, for example, in the works by Clifford Geertz and his colleagues).

The method of fieldwork was adopted in most Western European countries, though in France the interests of leading anthropologists (Lévi-Strauss in particular) were far more concentrated in oral and textual traditions such as tribal myths of creation. They attributed little importance to the ethnographer's experiences *in vivo*. The ethnographic tradition was also adopted outside Europe and the United States.

2. The Practice of Fieldwork

As a first step in their project, fieldworkers in the Malinowskian tradition have always been acutely aware they must develop *rapport* and friendly relationships with the people in their chosen field. They are expected to listen attentively, and witness whatever happens around them in the field, without disturbing the natural flow of events. As they consolidate their position with locals, they usually add information through more active modes of communication (such as initiating conversations, interviews, collecting statistical data, taking photographs, etc.), but they are careful not to instigate a flow of information into preconceived patterns of their own making. They invest much of their time recording this material in their notebooks or computers. The bulk of that store accumulates into their most treasured fieldwork product—'fieldnotes' (to be later transformed into an ethnography). The hallmark of anthropology, fieldwork had also come to differentiate anthropology from sociology. That major difference became more prominent a few decades later when anthropologists were no longer confined to the study of remote Third World societies.

The practice of fieldwork has raised many, as yet not fully resolved, problems of methodology and ethics. Considered a 'method,' it is assumed to maintain a standardized guide for professional behavior, as well as technical rules concerning the collection and recording of data. But throughout the history of anthropology, the craft of fieldwork was practiced, mostly by individuals, who went on their own to explore new research sites where they confronted conditions and constraints not envisioned by their teachers and colleagues. For many years, the teaching of anthropology in many universities did not include specific courses in methodology. The apprenticeship in anthropology was gained through more informal methods such as a close familiarity with the relevant corpus of ethnographies (and particularly those associated with the candidates' departments, for example, 'The Manchester School'), attending presentations and seminars by colleagues who returned from the field, communication with those who had already gone through the experience, and finally, through the good or bad fortunes of the novice on his arrival in the field and his success in establishing rapport with his native hosts. No surprise, fieldwork had been often described

as the major *rite de passage* for students of anthropology. Many anthropologists have been engaged in one major fieldwork project only. But it seems that more ethnographers have gradually become involved in a few fieldwork ventures throughout their career.

Ethnographic work in the various sites of fieldwork, in the early days and today, has been based on an ethos of trust. Anthropologists, mostly engaged in a lone journey, have never been able to implement a procedure of testing the accuracy and truthfulness of fieldnotes reported by colleagues and students. Rarely did fieldworkers choose to return to a field studied already by another ethnographer. The few cases of contested reports have gained much attention, in particular, the controversy between Oscar Lewis and Robert Redfield who studied the same community of Tepotzlan, and Derek Freeman's refutation of Margaret Mead's observations in Samoa. But long-term research, carried out by anthropologists whose continuing interest and commitment to a particular field has been supported by sufficient funds or easy access, also offered a strategy of verification of observations made at earlier trips.

Although anthropologists have been expected to master the language of the society they studied, that prerequisite was rarely fully achieved. Constraints of time and linguistic skills often impelled them to rely on 'informants' as interpreters until they gained a better command of communication in the native languages. The length of stay in the field was usually determined by budgetary as much as by personal circumstances. But an informal norm of one year seems to have dominated fieldwork projects.

How fieldworkers present their role, and to what extent their subjects comprehend the purpose and consequences of their work, remains an unsolved ethical problem. However, some norms that pertain to a universal code of behavior have been institutionalized at an early stage, such as the disguise of the community studied. More rigorously maintained is the tradition of disguising the identity of individual participants. But no regulations were designed in regard to the rights of privacy in recording and disclosing of intimate information collected during fieldwork. Anthropologists were left to make their own decisions concerning ethical issues under the assumption of a deeply felt responsibility toward 'his/her people.' This dilemma was addressed, for example, by Geertz (1968, p. 157): 'A professional commitment to view human affairs analytically is not in opposition to a personal commitment to view them in terms of a particular moral perspective.' In any case, fieldworkers were expected to record in their fieldnotes, without censoring, as much as they could remember from their daily observations and conversations.

During many years of practice, however, anthropologists informed their colleagues, also in a more public fashion, about the circumstances they en-

countered during fieldwork. The early publications had mostly revealed the personal tribulations of individuals in their respective fields (e.g., Powdermaker 1966, Freilich 1970). But certain specific subjects of methodology had engaged a more rigorous professional treatment. For example, 'the extended case-method' (Gluckman 1958), though mostly a method of ethnographic presentation, suggested a follow-up strategy that stimulated the fieldworker to concentrate his/her observations, thereby revealing the hidden connection between disjointed events. Another theme that gained much attention concerned the various types of informants one might meet at entrance to a new field, and the strategies and precautions one might consider working with them (e.g., Casagrande 1960, Shokeid 1988). Detailed manuals and reference materials for the various aspects of fieldwork have gradually developed into books of methodology so long absent (e.g., Epstein 1967, Wax 1971, Agar 1996, Ellen 1984, Shaffir and Stebbins 1991). A more meticulous approach came to the fore examining the methods of recording and safekeeping of fieldnotes, the most valuable product of fieldwork (Sanjek 1990). Also, conferences dedicated to specific methodological issues have become frequent events (such as 'fieldwork at home').

3. *Transformations in Fieldwork Research*

The idea of fieldwork had gone through immense changes since the days of Malinowski and the generations that established the ethnographic tradition of participant observations in remote, apparently homogeneous societies, clearly identified within territorial and cultural boundaries. The consequences of World War II, the decline of colonialism on the one hand, but the expansion of anthropology in academia on the other, made both the entry and the budgeting of long trips to remote continents an impractical goal for the growing number of students. Instead, many candidates turned their interest to various attractive fields within the borders of their own societies. No doubt, there had been previous successful projects in the United States and Europe of ethnographic studies in fields 'at home' (for example the most celebrated work in an Italian lower class Boston neighborhood by Whyte 1970). But these few research endeavors were unusual cases, in a tradition that was directed to confront 'other cultures' away from home.

The new search for fields close to home seemed at first to involve only a transfer of sites, such as the study of the apparently clearly defined urban ethnic neighborhoods in the United States (e.g., Lewis 1966, Hannerz 1969). However, that novelty demanded a change in the style and intensity of participant observation (Plotnicov 1973). But of more profound consequence, the next change of spatial sites of fieldwork came to also include institutions that offered

specific services to permanent or transient populations (for example, Myerhoff's (1978) work in a community center for aged people). That transformation opened the door to a choice of fields that, hitherto, have not been considered suitable for observation (for example, fieldwork in the scene of gay life, such as by Newton 1972, Read 1980, Shokeid 1995). In these fields, the fieldworker is often unable to observe the participants throughout a wide spectrum of daily activities. Moreover, in the fields of anonymous participants, the observer's ability to record behavior and communicate with his subjects of research had been very limited in scope. These new fields were instrumental in the development of the method and theory of *symbolic interaction*.

The transformation of methods could also be identified in the choice of problem-oriented ethnographic projects. Even when the relevant site was actually a bounded territorial unit, it engaged a modified strategy of fieldwork (e.g., Ginsburg 1989). But another major change of sites entailed a complete eradication of the bounded spatial characteristics of fieldwork. These unbounded sites included observations of participants in particular occupations (e.g., Agar 1986, Hannerz 1998), observing cultural performances (e.g., Heilman 1983), as well as observing the members of dispersed social categories (e.g., Ortner 1997). These fields, no longer designed to record the totality of personal and communal life experiences, have been variously defined: *delocalized*, *multisited*, *postcommunity*, *focus groups*, and *transnational spaces*.

The ongoing change of the spatial research entities have not necessarily transformed some of the major characteristics of fieldwork. Anthropologists in the new fields continue to initiate their observations mostly as individuals, men or women, engaged in their lone projects. It seems, however, there are less couples involved in these projects, compared with the tendency of married anthropologists to take along their families to the more remote field sites.

In recent years, however, some issues not considered before, concerning the fieldworker's feelings and demeanor, have been openly raised in conferences and edited volumes. Anthropologists preparing for a fieldwork project have never been requested to consult about their personality traits and social capabilities to accommodate what might become stressful field circumstances. But since Malinowski's widow published his personal diaries (1967), revealing the prejudices and the emotional stress experienced in the field by the mythical founder of fieldwork, a more relaxed attitude was gaining place in the professional discourse about fieldworkers' personal problems. This attitude has been closely related to the genre of reflexivity that since the late 1970s gained popularity in ethnographic text (e.g., Rabinow 1977). The lone fieldworker was no longer perceived through the image of the infallible man or woman, expected to withstand successfully all obstacles in a valiant test of personality that proves

he/she are capable of penetrating the social codes of alien cultures in far away territories.

The new fieldwork sites in Western urban societies contributed to that growing awareness and doubts about some long established taboos that separated the fieldworker from the subjects of his observations (e.g., Kulick and Willson 1995, Lewin and Leap 1996). The border between the observer and the observed that was naturally upheld in the old fields, where the alien researcher had been conspicuously distinguished (by race and other visible marks of personal characteristics), are no longer easily maintained in many of the new sites of fieldwork. The photographs displayed in many ethnographies of the founding generation, showing them in a landscape of exotic trees and simple huts, posing in the company of minimally clothed natives, have become a symbolic emblem for the changing circumstances of fieldwork.

It is probably an irony that by the end of the twentieth century, a century that gave rise to modern anthropology, fieldwork, its most inventive method and major signifier of professional identity, seems to have lost its incontestable paramount position. Critical views of the validity of positivist methodology, as well as the growing influence of postmodernist currents in cultural studies, are challenging the authority of ethnographic texts based on participant observations (e.g., Marcus and Fisher 1986). These intellectual movements, popular in recent years, inspire and legitimize anthropological projects that are based on other types of data (such as textual and mass media creations) which can be collected without the need for fieldwork. It is too early to predict whether the new genre of arm-chair anthropology will have a more lasting impact on the role of fieldwork and its practitioners.

See also: Anthropology; Anthropology and History; Anthropology, History of; Cultural Critique: Anthropological; Cultural Relativism, Anthropology of; Ethnography; Ethnology; Field Observational Research in Anthropology and Sociology; Psychological Anthropology; Qualitative Methods, History of; Symbolic Interaction: Methodology; Thick Description: Methodology

Bibliography

- Agar M H 1996[1980] *The Professional Stranger*, 2nd edn. Academic Press, San Diego, CA
- Agar M H 1986 *Independents Declared: The Dilemmas of Independent Trucking*. Smithsonian Institution Press, Washington, DC
- Casagrande J B (ed.) 1960 *In the Company of Man: Twenty Portraits by Anthropologists*. Harper, New York
- Ellen R F 1984 *Ethnographic Research: A Guide to General Conduct*. Academic Press, London
- Epstein A L (ed.) 1967 *The Craft of Social Anthropology*. Tavistock, London

- Fortes M 1978 An anthropologist's apprenticeship. *Annual Review of Anthropology* 7: 1–30
- Freilich M (ed.) 1970 *Marginal Natives: Anthropologists at Work*. Harper & Row, New York
- Geertz C 1968 Thinking as a moral act: ethical dimensions of anthropological fieldwork in the new states. *The Antioch Review* 28: 139–58
- Ginsburg F D 1989 *Contested Lives: The Abortion Debate in an American Community*. University of California Press, Berkeley, CA
- Gluckman M 1958[1940] *Analysis of a Social Situation in Modern Zululand*, Rhodes–Livingstone Paper No. 28. Manchester University Press, Manchester, UK
- Hannerz U 1969 *Soulside: Inquiries into Ghetto Culture and Community*. Columbia University Press, New York
- Hannerz U 1998 Reporting from Jerusalem. *Cultural Anthropology* 13: 548–74
- Heilman S C 1983 *The People of the Book: Drama, Fellowship and Religion*. University of Chicago Press, Chicago
- Kulick D, Willson M (eds.) 1995 *Taboo: Sex, Identity and Erotic Subjectivity in Anthropological Fieldwork*. Routledge, London
- Lewin E, Leap W L (eds.) 1996 *Out in the Field: Reflections of Lesbian and Gay Anthropologists*. University of Illinois Press, Urbana, IL and Chicago
- Lewis O 1966 *La Vida: A Puerto Rican Family in the Culture of Poverty—San Juan and New York*. Martin Secker and Warburg, London
- Malinowski B 1922 *Argonauts of the Western Pacific*. Routledge & Sons, London
- Malinowski B 1967 *A Diary in the Strict Sense of the Term*. Routledge & Kegan Paul, London
- Marcus G E, Fischer M M J (eds.) 1986 *Anthropology as Cultural Critique*. University of Chicago Press, Chicago
- Mead M 1928 *Coming of Age in Samoa*. Morrow, New York
- Myerhoff B 1978 *Number Our Days*, 1st edn. Dutton, New York
- Newton E 1972 *Mother Camp: Female Impersonators in America*. Prentice-Hall, Englewood Cliffs, NJ
- Ortner S 1997 Fieldwork in the postcommunity. *Anthropology and Humanism Quarterly* 22: 61–80
- Plotnicov L 1973 Anthropological fieldwork in modern and local urban contexts. *Urban Anthropology* 2: 248–64
- Powdermaker H 1966 *Stranger and Friend: The Way of an Anthropologist*, 1st edn. W.W. Norton, New York
- Rabinow P 1977 *Reflections on Fieldwork in Morocco*. University of California Press, Berkeley, CA
- Read K E 1980 *Other Voices: The Style of a Male Homosexual Tavern*. Chandler & Sharp, Novato, CA
- Sanjek R (ed.) 1990 *Fieldnotes: The Makings of Anthropology*. Cornell University Press, Ithaca, NY
- Shaffir W B, Stebbins R A (eds.) 1991 *Experiencing Fieldwork: An Inside View of Qualitative Research*. Sage, London
- Shokeid M 1988 Anthropologists and their informants: marginality reconsidered. *Archives Européennes De Sociologie* 29: 31–50
- Shokeid M 1995 *A Gay Synagogue in New York*. Columbia University Press, New York
- Turner V W 1985 *On the Edge of the Bush: Anthropology as Experience*. The University of Arizona Press, Tucson, AZ
- Wax R H 1971 *Doing Fieldwork: Warnings and Advice*. University of Chicago Press, Chicago
- Whyte W F 1970[1943] *Street Corner Society*. University of Chicago Press, Chicago

M. Shokeid

Figurative Thought and Figurative Language, Cognitive Psychology of

1. A Definition

Different types of linguistic expressions (metaphors, idioms, proverbs, etc.) form figurative language. Since the seminal work of Lakoff and Johnson (1980 but see Black 1979), the term 'figurative' is also used to denote conceptual entities that contribute to shaping our understanding of the world. Figurative language in general, and metaphor in particular, plays a central role not only in everyday discourse and in linguistic change but also in reflecting and shaping how people think in a broad range of domains (Cacciari 1998).

1.1 Literal vs. Figurative Language

When a speaker/reader encounters a sentence, she/he is generally able to judge whether it is intended literally or not. The ease with which people deal with this distinction in everyday discourse belies the difficulty and complexity of the distinction between literal and figurative language in the linguistic and psycholinguistic literature. No universally accepted criteria for discriminating between literal and nonliteral expressions have yet been devised. Many researchers expressed serious doubt as to whether a sharp distinction can (or should) be drawn between these two varieties of language use and proposed to substitute it with the notion of levels of conventionalization of meanings and sentence uses.

Despite the pervasiveness of figurative language, many language researchers still assume that it represents a peripheral issue that language comprehension and production models can easily ignore. For instance, current psycholinguistic models of production are almost silent as to why a speaker would select a figurative expression instead of a 'corresponding' literal expression (if any). Likewise, standard semantic theories still assign a rather peripheral and theoretically uninteresting role to figurative language (with the important exception of cognitive linguistics). Two long-standing assumptions contribute to this error.

(a) Figurative expressions can always be paraphrased literally since they only represent a 'stroke of style' reflecting, as Black (1979, p. 27) ironically stated, the 'incidental pleasure of stating figuratively what might just as well have been said literally.'

(b) The opposition between literal language as the realm of clear shared meanings, and figurative language as the realm of idiosyncratic, obscure meanings.

The cognitive psychology of figurative language does not consider figurative language as an ornament added to everyday straightforward literal language but it views it as a powerful communicative and

conceptual tool. This shift in the view of figurative language, and particularly of metaphor, from a strictly linguistic entity to a conceptual structure has opened a new wave of studies on the mental representation and the comprehension of figurative language. The main stance of this new viewpoint consists of treating metaphor as a bridge that forces the human mind to construct higher-order linkages between the entities referred to in metaphorical assertions. This bridging function is very widespread since metaphors can be used to connect a variety of different domains: abstract entities; abstract concepts to sensory-perceptual experiences and to concrete objects; perceptual experiences belonging to different sensory modalities (as in synaesthetic metaphors).

2. *The Comprehension of Figurative Language*

Contemporary cognitive research on figurative language comprehension began in the 1970s with a predominant focus on metaphor only recently extended to other linguistic expressions, particularly to idioms. Despite the fact that a vast family of expressions form figurative language, not all tropes have been investigated with a theoretical and experimental effort similar to that employed for metaphors and idioms. For instance, proverb comprehension has been scarcely investigated (see Honeck 1997); analogies have almost always been treated as a separate problem generally studied independently from metaphor, even though the two tropes are often confused and sometimes actually difficult to differentiate (see Gentner and Jerzierski 1993). The privilege accorded to metaphor is not surprising if one considers that, since Aristotle, it has been considered the trope *par excellence* often even serving as a collective label for any and all rhetorical figures.

3. *Metaphors as Comparisons*

Metaphor, according to the still most influential definition that comes from Aristotle, is said to be a form of transfer of a name from one object to another. The 'ground' of a metaphor (i.e., its content) can be the set of properties that two concepts share (e.g., in 'Dew is a veil' both are transparent, covering, shimmering). In some cases, it can be a superordinate category that subsumes both concepts (e.g., as in 'the tooth of the mountain' where the mountain peak and the tooth partake of the category 'sharpened forms'). In other cases, it may be necessary to reconstruct the underlying analogy involving both concepts (e.g., as in Virginia Woolf's 'She allowed life to waste as a tap left running').

The comparison view of metaphor has generated a class of comprehension models that can be characterized as feature- or attribute-matching models (for a review, see Cacciari and Glucksberg 1994). Matching

models adopt the assumption that in attributive metaphor, of the form 'An A is a B,' the A and B concepts are represented either as sets of features or as positions in an n -dimensional semantic space. Regardless of the specific representational assumptions, matching models assume that metaphors are first transformed into comparison statements, and then the features of the vehicle are compared to, or mapped onto, those of the topic.

The two basic assumptions – that metaphors are implicit comparison and that the comparison is derived from a sharing of features – are each problematic, for different reasons, and have been extensively criticized insofar as they do not contribute to definitively clarify what metaphors are and how they are understood. These assumptions are in fact unable to account for several factors: first, that metaphorical comparisons, unlike literal ones, are nonreversible or asymmetrical, that is, they change their meaning if the order of the topic and the vehicle is modified. Second, they do not provide any explicit mechanism for distinguishing between relevant and irrelevant features or properties nor for introducing new features into the mental representation of metaphor's topic. In processing terms, then, they posit that in order to understand a metaphorical assertion, one should first transform it into an explicit comparison. Hence this process should take longer than with a comparable literal expression. On the contrary, a consistent body of evidence suggests that attributive metaphors are understood as quickly as literal sentences when embedded in appropriate contexts, a result which confirms that identical comprehension processes are involved irrespective of language use. More importantly, the feature-matching process may be too limited to capture the full range and impact of metaphors.

Metaphors may have more to do, as Levinson (1983, p. 150) noted, with 'the contingent, factual (real-world) attributes of the referents of the metaphorical focus than with the semantic features that can be claimed to express its meaning.' Metaphor does something more than simply calling our attention to pre-existing similarities between the semantic features of the topic and the vehicle. Whether similarity is pre-existing or is created by the metaphor depends of course on the nature and quality of the metaphorical assertion. In many cases, good metaphors certainly made us perceive resemblances previously unperceived.

4. *Metaphors as Categorizations*

A metaphor can imply a comparison, but it is not in itself merely a comparison. Provided that it is not an implicit comparison, how can we characterize a metaphorical statement? Two theoretical answers have contributed to radically change the cognitive studies on metaphor representation and comprehension: the

Conceptual model (Lakoff and Johnson 1980, Lakoff 1993, Gibbs 1994) and the Property Attribution model (Glucksberg and Keysar 1990, 1993).

These models share the idea that metaphor's use and understanding involves not only linguistic or discourse processes but also reasoning and conceptual processes. Metaphor is based at the same time on the use of acquired categories and conceptual schema and on the creation of new categories and conceptual schemas. These models differ in important respects: first of which the locus of metaphoricity. According to the Property Attribution model proposed by Glucksberg and Keysar, metaphors belong to language use since they are linguistic entities expressing assertions of categorization similar to those expressed with literal language. They are used not simply to call the reader's attention to superficial similarity between concepts but to label categories that have no conventional names of their own. Metaphors create new categories employing the predicate to name the class of things that it exemplifies. When one says, 'My job is a jail,' she/he wants to suggest that his/her job is part of the metaphorical category of situations that are involuntary, unpleasant, and so forth, and to attribute specific properties to the metaphor topic 'job.'

In nominal metaphors, the metaphor topic and the vehicle play distinctive roles in that the topic provides constraints on what is likely to be attributed, while the vehicle provides properties to be attributed to the topic. Metaphorical vehicles are located at two different levels of category abstraction: the literal concrete level (the subordinate level), where they are members of a class of entities, and the more general, superordinate, attributive category level. A metaphor vehicle such as 'jail' can refer simultaneously to a concrete physical structure, and to the metaphorical category of situations that are involuntary, unpleasant, confining, and so forth. The literal 'jail' contains all the properties of the superordinate category 'jail' plus the properties that are specific to a jail *per se* (e.g., having painted walls, being made of concrete, etc.).

According to the Conceptual model proposed by Lakoff and colleagues, the locus of metaphor is not in language, but in the way in which people conceptualize one mental domain in terms of another. 'The general theory of metaphor is given by characterizing such cross-domain mappings' (Lakoff 1993, p. 203). The linguistic expression of metaphor is a simple epiphenomenon of a more general feature, i.e., that we think in metaphorical ways. Lakoff contends that everyday metaphor is characterized by a huge system of thousands of cross-domain mappings that motivates the idea that metaphor is primarily conceptual. It is not just 'a matter of language, but of thought and reason. The language is secondary. The mapping is primary, in that it sanctions the use of source domain language and inference pattern for target domain concepts. The mapping is conventional, that is, it is a fixed part of our conceptual system' (Lakoff 1993, p.

208). People use figures of speech in everyday conversation because they conceptualize much of their experience through the figurative schemes of metaphor, idioms, metonymy, and so forth.

These two models also differ in their theoretical scope: whereas Glucksberg and Keysar want to provide a psycholinguistic model of how people use and comprehend metaphors in discourse, Lakoff's aim is to single out a general account of human thought independently from any processing validity (but see Gibbs 1994).

5. Idioms: What They are and How They are Comprehended

The task of defining what an idiom is, how it is acquired and understood is still rather controversial although enough evidence has been cumulated to formulate satisfying models of the comprehension processes (for a review see Cacciari and Tabossi 1993, Gibbs 1994). One of the problems in defining idioms is that idiomatization is a process: a given structure is not idiomatic once and for all, but gradually acquires its idiomaticity. A very general definition is that idioms are composed of freely occurring units whose global meaning cannot be reduced simply to the meanings of these units.

Idioms behave at the same time as words and as larger-than-words chunks. From a processing viewpoint, this has been taken to imply that they were multiword lexical units accessed as such from the mental lexicon. The 'Lexical look up' models of idiom comprehension share the idea that idiom meanings are apprehended by direct memory retrieval, not by linguistic processing. They differ in terms of representational assumptions: according to Bobrow and Bell (1973), idioms are stored in a separate mental idiom list that is accessed and scanned when the literal analysis of a given string fails. This assumption has been rejected on the basis of the robust finding that idioms are understood as quickly as (if not more quickly than) comparable literal expressions. According to Swinney and Cutler (1979), idioms are represented as long words together with all ordinary words of the mental lexicon.

The received view is that idioms are noncompositional strings of words. Despite this, it has been noted that even in the most frozen of idioms (e.g., 'kick the bucket'), the semantic of the single constituent words is still available even though these idioms behave much as words, with their meaning acquired and stored in memory, as for other memorized strings of words (lines of poetry, snatches of songs, etc.) one cannot 'slowly kick the bucket' due to the semantics of 'to kick' that implies an abrupt action (Nunberg 1978, Glucksberg 1993). On more general grounds, this reflects the fact that language comprehension is nonoptional: in spoken language, the

meanings of the words are activated and discarded all the time since understanding occurs automatically without conscious control by the subject.

The recent psycholinguistic literature has shown that during the comprehension of idioms there is activation of the literal meaning of the words forming them, despite what is predicted by the direct look up models. According to the Configuration model (Cacciari and Tabossi 1988, Tabossi and Zardon 1993), idioms are no different from any other type of familiar, memorized string of words. When during a conversation a listener encounters one of such strings (e.g. 'being in seventh heaven'), the word meanings are activated up to the point (and even sometime later) in which the string itself is recognized as an idiomatic configuration. A configuration is made up of the same lexical units that need to be activated during the comprehension of literal discourse. This implies that there is only one processing of an idiomatic string up to the point in which, after the 'keyword' of the idiom is activated, the configuration can emerge. In some idioms, the keyword can appear very early, for instance, immediately after the main verb (e.g., as in 'spilling the beans' or 'breaking the ice'); in other cases it can be located at the end of the idiomatic configuration (as in 'being in seventh heaven' or 'looking for a needle in a haystack'), changing the time course of the comprehension process.

6. The Role of Semantic Transparency

The role of syntactic and semantic factors in idioms' interpretation has emerged in the last few decades corresponding to a new wave of studies that has concerned two main areas. The first one concerns the interplay between lexical and syntactic information, that is, the syntactic flexibility of idioms. A consistent body of evidence suggested that the number of syntactic variations an idiom can undergo still retaining its idiomatic meaning may be predicted on semantic grounds: the more the idiomatic meaning can be analyzed, the more flexible the idiom is. The second area has to do with the interplay between lexical and semantic/conceptual systems, that is, with the problem of semantic transparency of idiom meanings. The semantic transparency of an idiom's meaning has been defined as the extent to which the meanings of the words forming an idiom contribute to its overall figurative interpretation. Enough evidence already exists to reject the received view that idioms are semantically empty strings, suggesting that semantically transparent idioms (e.g., 'spill the beans', 'skate on thin ice') are easier to comprehend and to acquire for both children and adults.

These studies argued that adults and children could reliably differentiate among types of idioms (for a review, see Cacciari and Tabossi 1993), for instance among:

(a) Analyzable-opaque: the relation between an idiom's elements and the idiom meaning may be opaque, but the meaning of the individual words can nevertheless constrain both interpretation and use (e.g., 'kick the bucket,' 'by and large,' 'speak one's mind').

(b) Analyzable-transparent: there is a clear semantic relation between the elements of the idiom and the components of the idiom's meaning usually because of metaphorical correspondence between the idiom's elements and the components of the idiom's meaning. The elements of an idiom map onto the components of the idiom's meaning (e.g., in 'spill the beans,' 'spill' corresponds to the act of revealing or letting out, and 'beans' to the materials that had been heretofore concealed or otherwise unknown).

(c) Quasi-metaphorical idioms: here the literal referent of the idiom is itself an instance of the idiomatic meaning. A particularly relevant case or instance is used to name all possible other situations that one might want to describe using the semantic characteristics epitomized by the idiom referent. For instance, 'carrying coals to Newcastle' is simultaneously an ideal or prototypical instance of a useless transfer of something and also a phrase that can refer to any instance of bringing something to a place that has a surfeit of that something (as in 'Selling rice to Japan is like carrying coals to Newcastle'). The literal referents of this type of idiom are relevant and appropriate to their idiomatic meanings.

Quasi-metaphorical idioms may serve the same communicative function of metaphor vehicles in expressions such as 'My lawyer is a snake' or 'My job is a jail.' In these metaphors, vehicles such as 'snake' or 'jail' serve as exemplars of their metaphoric categories—vicious people and confining, unpleasant situations, respectively—and simultaneously as names for those categories (Glucksberg and Keysar 1990). Consider a metaphorically intended term such as 'snake.' It provides a set of candidate properties (many and different) that can be plausibly attributed to a person. Usually, mutual knowledge among the speakers and/or the discourse context provide the relevant knowledge about the person to which such a vehicle ('snake') refers so that only a subset of properties will be attributed to him/her. Metaphorically speaking, a person can be a 'snake' in several respects: she/he can be vicious, dangerous, hit you unexpectedly, and so on. Topics such a lawyer or a supposedly good friend would require, for instance, the instantiation of a different set of attributive properties (or values) insofar as each of these two topics put constraints on what is more likely to be attributed to them.

Linguistic processing and idiom look-up can occur in parallel. Ordinarily, people understand idioms following the basic principles of language comprehension to court. Language is full of noncompositional units that are not idiomatic at all. Consider, for

instance, nominal compounds such as 'carpet sweeper' or 'vacuum cleaner': a carpet sweeper sweeps carpets but a vacuum cleaner does not clean vacuums. As a processing default, people retrieve an idiom as a configuration from the lexicon. Still, the meanings of the words forming it and its underlying structure (be it metaphorical, analogical, metonymical, etc.) are available and might play a role if necessary for discourse purposes. In these cases, people can use their intuitions on the structure of an idiom as well as on the conceptual templates underlying it to elaborate on an idiom's meaning.

7. Sense Creation and Extension

The notion that semantic properties of the constituent words of an idiom play an important role in the comprehension and production of idioms has been extended to also account for the problem of meaning extension, i.e., polysemy (Glucksberg 1993). The words forming familiar idioms might become polysemous in that they have at least two meanings: their default context-free literal meanings, and the meanings that are induced by the idiom context. With repeated use, familiar idioms might become able to induce polysemy adding the idiomatic meaning to each word's set of possible meanings. The verb 'spilling,' for instance, might acquire as a secondary meaning that of 'revealing.'

Figurative language is arguably the most powerful source for meaning creation and sense extension. Words evolve additional meanings through it. When such an evolution is consistent and stable enough within a speech community, then a new lexical entry can be added to the list of possible meanings of a word. At that point, lexicographers will officially confirm a new sense by adding it to the dictionary.

Words as well as idioms do not acquire new senses randomly. Most often idioms derive their meaning from metaphors that for some reason emerge in language. According to many authors, this process is not at all arbitrary (for a review, see Cacciari 1998). As Quinn (1991, p. 79) noted, 'the invention of new metaphorical expressions is constrained...by the necessity of finding metaphors that make sense in terms of the cultural models and by preference for metaphors that do a particular good job of this.' Hence, metaphors that become conventionalized (the first step toward idiomatization) presumably represent 'satisfying instantiations' of culturally shared models.

Language comprehension may always involve a mixture of sense selection and sense creation. As Clark and Gerrig (1983) noted, in general speakers consider the information they share with the addressee, evaluate the salience and coherence of potential interpretation against this common ground, and select or create specific interpretations, 'all with deceptive ease' (1983, p. 607). This 'deceptive ease' reflects human's ability to simultaneously process information coming from

different domains: lexical, syntactic, semantic, pragmatic, and so forth. Figurative language, as an old tree, extends its roots across these domains, taking its nourishment where it springs. That's why it is such a crucial hinge among different language systems.

See also: Creativity and Cognition; Cognitive Development in Childhood and Adolescence; Comprehension, Cognitive Psychology of; Language Acquisition; Language and Thought: The Modern Whorfian Hypothesis; Scientific Reasoning and Discovery, Cognitive Psychology of; Semantic Similarity, Cognitive Psychology of; Sentence Comprehension, Psychology of; Word Meaning: Psychological Aspects

Bibliography

- Black M 1979 More about metaphor. In: Ortony A (ed.) *Metaphor and Thought*. Cambridge University Press, Cambridge, UK, pp. 19–43
- Bobrow S, Bell B 1973 On catching on to idiomatic expressions. *Memory and Cognition* 1: 343–6
- Cacciari C 1998 Why do we speak metaphorically? Reflections on the functions of metaphor in discourse and reasoning. In: Katz A, Cacciari C, Gibbs Jr R W, Turner M (eds.) *Figurative Language and Thought*. Oxford University Press, Oxford, UK, pp. 119–57
- Cacciari C, Glucksberg S 1994 Understanding figurative language. In: Gernsbacher M A (ed.) *Handbook of Psycholinguistics*. Academic Press, New York, pp. 447–77
- Cacciari C, Tabossi P 1988 The comprehension of idioms. *Journal of Memory and Language* 27(6): 668–84
- Cacciari C, Tabossi P (eds.) 1993 *Idioms. Processing, Structure and Interpretation*. Laurence Erlbaum, Hillsdale, NJ
- Clark H H, Gerrig R J 1983 Understanding old words with new meanings. *Journal of Verbal Learning and Verbal Behavior* 22: 591–608
- Gentner D, Jezierski D 1993 The shift from metaphor to analogy in Western science. In: Ortony A (ed.) *Metaphor and Thought*, 2nd edn. Cambridge University Press, Cambridge, UK, pp. 447–80
- Gibbs W R 1994 *The Poetics of Mind. Figurative Thought, Language and Understanding*. Cambridge University Press, Cambridge, UK
- Glucksberg S 1993 Idiom meaning and allusional content. In: Cacciari C, Tabossi P (eds.) *Idioms. Processing, Structure and Interpretation*. Erlbaum, Hillsdale, NJ, pp. 3–26
- Glucksberg S, Keysar B 1990 Understanding metaphorical comparisons: beyond similarity. *Psychological Review* 97(1): 3–18
- Glucksberg S, Keysar B 1993 How metaphor works. In: Ortony A (ed.) *Metaphor and Thought*, 2nd edn. Cambridge University Press, Cambridge, UK, pp. 401–24
- Honeck R P 1997 *A Proverb in Mind: The Cognitive Science of Proverbial Wit and Wisdom*. Erlbaum, Mahwah, NJ
- Lakoff G 1993 The contemporary theory of metaphor. In: Ortony A (ed.) *Metaphor and Thought*, 2nd edn. Cambridge University Press, Cambridge, UK, pp. 202–51
- Lakoff G, Johnson M 1980 *Metaphors We Live By*. The University of Chicago Press, Chicago
- Levinson S C 1983 *Pragmatics*. Cambridge University Press, Cambridge, UK

- Nunberg G 1978 *The Pragmatics of Reference*. Indiana Linguistic Club, Bloomington, IN
- Quinn N 1991 The cultural basis of metaphor. In: Fernandez J W (ed.) *Beyond Metaphor. The Theory of Tropes in Anthropology*. Stanford University Press, Stanford, CA, pp. 56–93
- Swinney D A, Cutler A 1979 The access and processing of idioms. *Journal of Verbal Learning and Verbal Behavior* 18: 523–34
- Tabossi P, Zardon F 1993 The activation of idiomatic meaning in spoken language. In: Cacciari C, Tabossi P (eds.) *Idioms. Processing, Structure and Interpretation*. Laurence Erlbaum, Hillsdale, NJ, pp. 145–62
- Wasow T, Sag I, Nunberg G 1983 Idioms: An interim Report. In: Hattori S, Inoue K (eds.) *Proceedings of the XIIIth International Congress of Linguistics*. CIPL, Tokyo, pp. 102–5

C. Cacciari

Film and Video Industry

The production, and distribution of cinema occurs throughout the world but since the 1920s has been dominated by eight or less giant Hollywood studios. These vast multinational enterprises create films that are presented in virtually every country in the world, frequently capturing a major share of the business despite competition from locally made and distributed motion pictures. Almost continuous technological change—led by the Hollywood oligopolists—did not lessen Hollywood's industrial power, and so all other film industries sought to ape Hollywood's continuous success. The coming of sound and color, and the innovations of pay-TV and home video (played on VCRs) simply extended Hollywood's influence, and increased its profits. Analysis of the Hollywood studio system has led to praise or damnation of this omnipresent, inescapable, and highly influential industry.

1. The Film and Video Industry

A handful of companies have defined the Hollywood film industry, dominating the US and world markets. They have weathered a world war, and a Great Depression, innovated wide screen and color technologies, made peace with television, learned to exploit home video, and were more powerful at the end of the twentieth century than ever before.

The coming of sound solidified control by five studios over the world cinema market: Paramount Pictures, MGM, Twentieth Century Fox, Warner Bros., and RKO. During the 1930s and 1940s by controlling the key cinemas in the downtowns of nearly all major US cities, the 'Big Five' took in nearly all box-office take. Only after they granted their own theaters first-runs, did they then permit smaller,

independently owned theaters to scramble for the remaining bookings months, or even years, after a film's premiere.

Paramount represented the most profitable, powerful, and traditional business-like Hollywood company. More than any other member of the Big Five, Paramount relied on its chain of more than 1,000 theaters to maintain its corporate might. Not surprisingly, a former theater operator, Barney Balaban, stood at the top of this corporate colossus, hiring more lawyers and MBAs than movie stars.

Metro-Goldwyn-Mayer (MGM) technically functioned as a successful unit of Loew's, a theater chain concentrated in the Eastern USA. MGM, more than any other studio, relied on the star system, producing films from its massive Culver City, California-based 'factory' of 27 sound stages (for interior shots), and a 168-acre back lot (for outdoor scenes).

Twentieth Century Fox ranked behind Paramount and MGM with neither as large a theater chain or studio facility. Studio boss Darryl F. Zanuck used what he had to authoritatively produce films for world distribution. His decisions were final, frequently belittling to all subordinates: 'Don't say yes until I've finished talking!'

Warner Bros. represented the only family-run operation among the major movie studios, with eldest brother Harry as President, middle brother Abe supervising worldwide distribution, and baby brother Jack heading studio production. The innovation of sound pushed Warners into the Big Five, yet financially the company struggled through the 1930s and 1940s.

Radio-Keith-Orpheum (RKO) was formed so that radio's RCA could sell its sound equipment to Keith-Orpheum theaters. While the studio created many memorable films—such as *Citizen Kane* and the Astaire-Rogers musicals—RKO's only regular profitability came from the distribution of Disney animation. TV killed RKO.

The Big Five tolerated some competition—from Universal Pictures, Columbia Pictures, and United Artists—as the 'Little Three' depended on the Big Five for access to the top theaters in the USA. The suburbanization of the US, and the coming of television allowed Universal, Columbia, and United Artists to prosper because the advantage of owning theaters disappeared because of antitrust action by the US government. The former Big Five, save RKO, tried to re-invent themselves.

Warner Bros. began its transition into what became Time Warner when in 1956 the founding brothers sold out, and a series of new owners embraced television production. In 1969, Kinney National Services, a New York conglomerate, purchased Warners, and Steven Ross, son-in-law of Kinney's founder, began to fashion the modern Hollywood-based prototypical multinational media conglomerate. In 1989 Ross merged Warner with Time, and the largest media

company in the world was created. A decade later AOL took control.

Paramount stubbornly hung on until 1966 when Charles Bluhdorn's Gulf + Western Industries conglomerate, after a 2-year battle, took over. Through the 1970s and 1980s the 'new' Paramount thrived as part of a classic conglomerate, but in another 2-year struggle, was taken over in 1993 by another media conglomerate, Sumner Redstone's Viacom, Inc.

Twentieth Century Fox began its reinvention in 1956 when longtime studio boss Darryl F. Zanuck resigned. His replacements struggled, but did lead the way to licensing feature films to major television networks. NBC's 'Saturday Night at the Movies' began TV's showcasing of feature films. The surprise hit *Star Wars* (1977) moved Fox to the top of the Hollywood studio hierarchy, and within a decade led to the studio's purchase by Australian media mogul Rupert Murdoch. From the Fox studio base, Murdoch created the Fox TV network.

MGM never was able to reinvent itself, even with its merger with United Artists. The two other members of the Little Three did far better. Columbia went through several ownership changes (including during the 1980s when Coca-Cola owned it), but it was not until 1989 that stability came when Japanese electronics giant Sony purchased Columbia. During the 1990s Sony never was able to exploit its Hollywood studio to sell more VCRs, but did learn to create vast profits.

Universal also struggled until 1962 when the MCA talent agency acquired it. Thereafter, under Lew Wasserman, Universal ascended to the top of Hollywood movie and television studio hierarchy as Wasserman innovated the made-for-TV movie, and the modern film blockbuster—*Jaws* (1975). Wasserman retired in the early 1990s, and in 1995 Canadian liquor giant Seagram took control.

Disney joined Warners, Paramount, Fox, Sony, and Universal to form the 'Big Six' of the 1990s. Beginning in 1984, Michael Eisner leveraged Disney's weak position into major studio status, a significant change in an industry with so much long-term stability. Indeed, how had this small number of companies survived and thrived for so long—with no real newcomers joining their ranks?

First, even with all the social technological and industrial changes, the major companies alone continued to be able to distribute films around the world. At considerable expense, all maintained offices not only in the USA, but in every region of the world. In most countries around the world, owners of theater chains knew the Hollywood majors alone could continuously deliver popular films, and so even theater owners outside the USA preferred Hollywood films to native productions. The Hollywood studios' record at the box office satisfied theater owners—a conservative lot with most of their assets invested in real estate. Economies of scale prevented rivals from building and maintaining equal world distribution networks.

Second, modern Hollywood fashioned the blockbuster strategy to turn television into a friend, not a foe. Through the final quarter of the twentieth century, the Hollywood oligopoly learned to generate profits from creations that premiered with blitz TV advertising, which in turn was married to a multirevenue generating machine to sell musical sound tracks, novelizations of the stories, 'action figures' from movie characters, and endless sales on video. Each movie aspired theatrically to become a smash hit because if it succeeded then the feature turned into a 'product line,' designed to fill all entertainment needs and desires from toys to theme park rides to campaigns to sell more McDonald hamburgers.

The blockbuster strategy in turn led to ever escalating costs of production so it became more difficult for rivals to join the exclusive Hollywood club. By the close of the twentieth century the average Hollywood feature cost in excess of \$50 million to produce, and half again that amount to publicize. *Titanic* doubled those average cost figures, but then earned in excess of \$2 billion in revenues. Potential filmmakers knew that if they ever aspired to have a world audience see their creation and to attain undreamed of wealth, they needed to be 'green lighted' by one of the Hollywood major studios.

Alone, Hollywood could keep the revenues flowing year after year—particularly from home video. While theaters still premiered new films, it was 'downstream' on video where the bulk of the monies were generated. The theater functioned as the 'voting booth' where hits were made, and necessary publicity generated. But once the film had fully milked the box-office it went into release on pay-TV and home video that created more than two-thirds—in the long run—of the average film's revenues. Hollywood executives carefully planned—in classic price discrimination fashion—to drain all possible revenues from each of these release venues, only releasing it to home video when they reasoned all the revenues from theatrical and then pay-TV had been procured. Each window in this sequence was an exclusive; a new window opened only when all value of the previous window had been captured. This exploitation of video represented Hollywood's newest advantage in the film industry.

2. *The Rest of the World*

While the Hollywood majors maintained their economic power and influence, filmmakers outside Hollywood long struggled to find their place in their national cinemas, often requiring the assistance of their governments. Only sporadically have large nation-states—France, Italy, Japan, and India in particular—been able to mount a studio system as an alternative and serious rival to Hollywood in their own countries, let alone elsewhere.

Globally, only in the rare nation does Hollywood not capture more than half the business. The Hollywood majors maintain offices around the globe, where their representatives maintain constant contact with the heads of the dominant theater chains, pay-TV outlets, and home video retailers. Hollywood's regular production of hit films provides a strong incentive for foreign theater, pay TV, and home video companies to consistently deal with them rather than take a flyer with some new company—even a native company.

But other nations did try. During the 1920s European nation-states—led by the United Kingdom, France, and Germany—sought to build their own studio systems, and were successful for short periods of time, until their leading filmmakers left for Hollywood. During the 1920s, the new Soviet Union tried to build an alternative national cinema industry under Communism. The 1930s ended these experiments. At the close of World War II, the focus shifted to state sponsored television, but still in Japan and India studio systems did succeed, even while conceding Hollywood a vast share of the monies paid for cinema attendance.

After the advent of the blockbuster, it became harder for foreign governments to subsidize a native cinema industry. Most chose to focus on television instead. But Germany and Australia, for example, did for a time in the 1980s create enough different types of popular productions long enough to be discovered as the latest 'new wave.' Japan had long had a studio system, but the coming of television and the takeover of Hollywood studios by Sony, and temporarily, Matsushita, ended the importance of that native industry. But India, for example, was still able to maintain native popular cinema, complete with a star system and defined genres like Hollywood. But these films were intended for domestic consumption, with minimal export possibilities. Community viewing in cities and poor areas alike still flourished.

Hollywood has long maintained an influential trade association—the Motion Picture Association of America noted for its ratings of films, but far more influential around the world for lobbying to smooth the way for international distribution. The MPPA lobbied the US State Department to keep cinema trade 'free' to permit the Hollywood oligopoly to continue pulling billions in revenue from the foreign [read: non-US] market. Foreign governments, on the other hand, resented US pressure, and formally and informally sought to protect the national film industries, and their cultural integrity and identity.

3. Industrial Analysis

The common thread of Hollywood power and influence, and its growing presence around the world, have led to much analysis of the Hollywood film industry. Three different interpretations have dominated.

3.1 Marxist Industry Analysis

Marxist critics have long seen this influential cultural industry as a classic example of monopoly capitalism, focusing on how these long lived corporations colluded to devise ways to maintain their power and cultural imperialism. While Wall Street celebrated Twentieth Century Fox and Paramount's co-financing of *Titanic*, Marxist critics would see this as yet another example of studios working together—not competing—to maximize profit. Since there were at most eight players, they did not have to worry about going out of business, but simply seeking how much their exploit and keep in profits.

The long-term propensity toward concentration of ownership, endemic to monopoly capitalism, was simply exemplified by the Hollywood motion picture industry. That is what history should teach us, and that is how Hollywood has behaved even in the face of new technologies. Marxists focus on the expansion of giant corporations and the spread of their influence to television, music, toys, theme parks, and other entertainments. Market concentration and anticompetitive behavior describe the operation and ownership of the Hollywood movie industry.

Marxist industry analysis usually focuses on international distribution, seeing the trend toward globalization as evidence of cultural and economic imperialism. While Hollywood publicity focuses on the production of films—its stars, stories, and special effects—international distribution has always been a key to corporate longevity. No other national film industry has been ever been so far-reaching, endangering indigenous, nonprofit seeking culture. They correctly emphasize that Hollywood taught the mass culture businesses to focus on and exploit the advantages of globalizing as seen in the work of Wasko (1994) and Guback (1969).

3.2 Neoclassical Industry Analysis

Neoclassical economists emphasize that movies face competition from all forms of mass culture. While there are and have been a few very powerful studios, consumer sovereignty means that no Hollywood company can ever force a customer to pay to see a film—even *Titanic*. These analysts focus on the forces and interplay of supply and demand, seeing the movie theater as a classic voting booth, where top grossers win and unpopular films are pulled almost immediately. They stress that there is a great deal of product diversity that permeates the entire landscape of the mass media. This is why genres—from science fiction to westerns, from dramas to comedies—ebb and flow in relative popularity. With the rise of the VCR, for example, the choice of titles to rent or buy needs to be measured in the thousands. Like the Marxists who come to their analysis with a predetermined

anticapitalist assumption, these neoclassical analysts enter the debate with a procapitalism assumption as seen in the work of Litman (1998).

3.3 Industrial Organization Analysis

Industrial organization economists start with neither pro- or anticapitalistic assumptions, and instead seek first to define 'who owns the media,' then analyze what economic conduct emerges from that industrial structure, then detail how the industry performs given specified criteria, and finally recommend possible public policy corrections to make the industry do 'better.' Industrial organization economists have long recognized the host of advantages that enabled them to maintain their considerable economic power and keep out the competition.

For example, cross-subsidization enabled a Hollywood corporation—with interests in a number of media markets—to extract profits from a thriving area to prop up another less financially successful area. Single-line corporations do not have this luxury and so aspiring single-line Hollywood operations invariably fail as Gomery (1986) discusses.

Or, to take another example, reciprocity enables a diversified member of the Hollywood major studios to choose to whom they will sell and then only deal with those companies that cooperate with other units of the media conglomerate. For example, Universal Pictures might not sell movies to Time Warner's HBO unless Time Warner's cable franchises book Universal's (partially owned) USA television network.

Horizontally, year after year, the Hollywood oligopoly has generated considerable profits from a wide spectrum of mass-media enterprises, including theme parks, music, and television production. There is no such thing as a single-line major movie company. The major Hollywood companies also control the process of production and distribution—as they long have—but also more and more of the increasing avenues of presentation, from pay-per-view channels to home video to digital home viewings.

Bibliography

- Alexander A, Owers J, Carveth R 1968 *Media Economics*. Erlbaum, Mahwah, NJ
- Balio T 1985 *The American Film Industry*. University of Wisconsin Press, Madison, WI
- Bruck C 1992 *Master of the Game: Steve Ross and the Creation of Time Warner*. Simon & Schuster, New York
- Conant M 1960 *Antitrust and the Motion Picture Industry*. University of California Press, Berkeley, CA
- Gomery D 1986 *The Hollywood Studio System*. St. Martin's, New York
- Gomery D 1992 *Shared Pleasures: A History of Movie Presentation in the United States*. University of Wisconsin Press, Madison, WI

- Guback T 1969 *The International Film Industry*. Indiana University Press, Bloomington, IN
- Kindem G 1982 *The American Movie Industry*. Southern Illinois University Press, Carbondale, IL
- Lardner J 1987 *Fast Forward: Hollywood, the Japanese, and the Onslaught of the VCR*. Norton, New York
- Litman B 1998 *The Motion Picture Mega-industry*. Allyn and Bacon, New York
- Morin A 1996 *Film Policy: International, National, and Regional Perspectives*. Routledge, London
- Neale S, Smith M 1998 *Contemporary Hollywood Cinema*. Routledge, London
- Wasko J 1994 *Hollywood in the Information Age*. University of Texas Press, Austin, TX
- Wyatt J 1994 *High Concept: Movies and Marketing in Hollywood*. University of Texas Press, Austin, TX

D. Gomery

Film: Genres and Genre Theory

Genre, a term notoriously difficult to pin down, refers to a type of film or other art object where instances of the type share similarities in form and style, theme and content, as well as communicative function. A film genre is thus based on a set of conventions that influence both the production of individual works within that genre and audience expectations and experiences.

1. The Study of Film Genres

The concept of genre has been important since the beginning of the history of film, as a category for the industry and thus for the production, marketing, and distribution of films; for film critics and the academic, esthetic, historical, and theoretical analysis of film; and for audiences as a framework for the selection and experience of films. The study of film genres is therefore placed in a triangular structure relating to the institutional, social, and cultural context of films; to the esthetic, formal, thematic, and stylistic dimensions; and to audiences and reception understood both in an empirical social and psychological sense.

The development of studies of film genre from classical to modern film theory have shifted between focus on different aspects of this triangular structure. Studies of particular film genres, e.g., the Western or the family melodrama, may stress the social and cultural aspects of genre, as ideological or ritual expressions of American history, values, and mythology, or as part of the Hollywood studio system and the institutional development of production norms. Or studies may stress the formal, narrative, visual, iconographic, and thematic structure of particular groups of films in order to find similarities and variations and historical changes and developments.

But studies of genre have also focused on psychological aspects of genre such as gender and the imaginary, using Freudian, Lacanian, and feminist terminology or, as it has been the case in recent thinking about genre, from a cognitive and emotional theory of basic prototypes of film genres. But many theories of genre do not stress just one part of the triangle but clearly indicate that genres are unstable and fuzzy categories. Rather than defining film genres as sharp, distinctive, and essential categories, modern genre theory tends to define genres as prototypes, and to see the establishing of genres as processes related to institution, text, and audience, and historical changes.

1.1 Auteur and Genre

Although film genres are found as categories from the beginning of film history, used by both production companies and film critics, the systematic theoretical and historical reflections on film genres were developed in the 1950s as European auteur theory met US genre films. The early theories of film were often very strongly tied to a more individualized and esthetic understanding of film as art, and therefore often viewed mainstream cinema and genres as commercially dominated formulas, that did not allow artistic freedom and innovation for the auteur. But although this discussion is often repeated even today, film genre theory in general has explicitly questioned this normative division between an industrialized product and mainstream cinema on the one hand and film art and the auteur concept on the other.

In the 1950s French critics centered around *Cahiers du Cinéma* and the New Wave stressed the variations and auteurism of classical genre films in their writings on classical Hollywood cinema. Bazin, despite his central role in the formulating of a realist film theory and an auteur theory, was also engaged in describing the special qualities of classical US genre-films, such as the Western. In his two seminal articles (Bazin 1971a, 1971b) published in French in 1953 and 1954, he discussed the Western, and especially John Ford's Westerns as US film art *par excellence*. In these articles, Bazin saw the Western as a visual and narrative embodiment of American history and myths, and in his definition of what constitutes a genre, he focused on both thematic elements and elements of cinematic technique and style, narrative structure, and *mise-en-scène*. He also initiated later discussions on the historical transformation of the Western, defining both its pure classical forms and newer forms, which he called 'superwesterns' and 'novelistic Westerns.'

Bazin's articles mark a transition from a normative esthetic notion of genre films as just degraded, commercialized, and industrialized art. He combined auteur perspectives and genre perspectives into a more general esthetic and cultural perspective on all forms of film. This was also the intention behind a number of

later US contributions to the general understanding and re-evaluation of US film genres. Braudy (1976) not only explicitly criticized the division between film as art and film genres, but also claimed that mainstream genres can express broader cultural, social, and esthetic values than classical works of art can. Buscombe ([1970] 1995) takes a broad discussion of literary genre theory as his starting point, and specifically uses Wellek and Warren's distinction in *Theory of Literature* ([1949] 1956) between 'outer form' and 'inner form' and their cinematic equivalents, thus opening a more structural and esthetic specification of genre definition.

1.2 Semiotics and the Ritual and Ideological Turn

The move from auteur to genre from the late 1960s was sustained and further developed through the influence of semiotics, structural anthropology, and literary genre theory to a broader ritual and ideological approach to the study of film and film genres. The ritual approach involved both a more systematic analysis of cinematic texts and a form of institutionalized practice expressing and negotiating fundamental cultural values. The ideological perspective involved a critical study of cinema and cinematic genres and the reproduction of dominant forms of ideology from sociological, Marxist, Freudian, and feminist perspectives. The two approaches do not necessarily exclude each other but in the ritual approach there is a tendency to see the relation between genre and audience as one of negotiation and exchange, whereas the ideological perspective more often stresses the principle of dominance.

In Cawelti's broad ritual study of popular literature and film (Cawelti 1976) and in Wright's classical ritual and structural analysis of the Western (Wright 1975), genres are seen as fundamental story types that represent and embody cultural images and myths with both universal and historical elements, and which are recognized by the audience. Genres and the act of genre reading balance between mimesis and escapism. Cawelti's idea is that genre formulas work because they are able to express and negotiate historical conflicts and mental and cultural oppositions and contradictions in a way that is accepted, understood, and appreciated by a large audience, not just dominated by ideology. He rejects deterministic theories in favor of symbolic or reflective theories, where assimilation, negotiation, and change take place but without denying that formula stories reflect and affirm dominant forms of thought and attitudes.

Cawelti defines five universal story types that can be used and combined in different ways: adventure, romance, mystery, melodrama, and stories of alien beings and states. In Schatz's (1981) study of classical Hollywood genres he talks about genres in a fundamental semiotic way in which he sees 'film genre as

a specific grammar or system of rules of expression and construction and the individual genre film as a manifestation of these rules' (Shatz 1981, p.19). Basic film genres form the *langue* and individual genre films form the *parole* of genre filmmaking and film language, and he defines the grammar in terms of a genre's iconography, characters, settings, and plot structure. Shatz's analysis is a study of some of the most stable film genres in the classical Hollywood system, the individual films, and the directors, but he also points to two very basic types of supergenres: genres of order (Western, gangster, detective) and genres of integration (musical, screwball comedy, family melodrama) with two very different narrative strategies, settings, types of hero, and solution models of conflicts that either express violent elimination and male dominance or emotional, familial integration and couple or female orientation.

Shatz's approach points to problematics also addressed in the more ideological and feminist genre theories developed at the same time, but with a background in French film semiotics or the influential UK journal *Screen*. One of the seminal articles articulating the ideological approach is Comolli and Narboni's ([1969] 1992) article, printed originally in *Cahiers du Cinema* and in English in *Screen* (1972). Every film is political and ideological, they state, in so far as no film or filmmaker can escape the commodity form and the capitalist economy. According to Comolli and Narboni, filmmaking at large is nothing but an expression of the prevailing ideology, which is working behind the back of both the filmmaker and the audience. But the article also defines a kind of ideological genre taxonomy, in which films are defined in relation to whether and how they try to oppose the dominant ideology. Among the seven genres defined are: the majority of films, which are just unconscious instruments of ideology; films that have an explicit critical content and or/form; films that have an ambiguous secret message and form; and films like documentary films, that are seen as only illusionary breaks with the dominant ideology.

Ideological genre analysis found dominance of ideology disguised behind classical realism or inscribed in the apparatus of cinema, and therefore viewed the audience as rather passive. However, ideological analysis did look for both a countercinema and counterstrategies to the dominance of ideology in genre film and mainstream cinema. But whereas the ritual approach saw qualities in the mainstream film and its relation to the audience on a more direct level, the ideological approach often investigated ambiguities and contradictions in genre films dominated by ideology. Writers belonging to the ideological and semiotic project (e.g., Wood [1977] 1995 and Klinger [1984] 1995) therefore also found elements of artistic personality still visible in genre films and often described subversion of the dominant tendencies in mainstream film.

1.3 Gender and Genre

The story of genre theory and gender starts with the feminist project based on the ideological and Freudian-Lacanian-inspired genre theory that sought to find the dominant. Mulvey's ([1975] 1992) much quoted essay combined elements from all aspects of the ideological genre theory, but with the aim of demonstrating the male dominance of the feminine in mainstream Hollywood. She pointed to the male gaze and the feminine characters as the looked-upon object, and she also described the asymmetric relation of active/passive roles. But Mulvey's article did not just open the feminist critique of male dominance, but also the demand for another esthetic, a feminist countercinema. Eventually it also led to the investigation of subversive female readings of classical genre films, most notably 'women's films,' e.g., melodrama or film noir as it is very broadly demonstrated in Gledhill's (1987) anthology on melodrama and women's films and Kaplan's ([1978] 1998) on women and film noir, in which prominent feminist film genre critics, e.g., Mary Ann Doane, Tania Modleski, and Annette Kuhn, figure. Later feminist writers have directly entered more action-oriented and excessive 'body genres,' such as Williams studying pornography, horror, and melodrama (Williams 1989) or Clover's (1992) thorough study of the horror genre. And just as feminist criticism has opened towards seemingly more 'masculine genres', gender studies of film genres have also included studies of gay and lesbian films or gay and lesbian aspects of mainstream genre films.

1.4 Genre, Reception, and History

In literary genre theory (Todorov [1970] 1975) the distinction between theoretical genres is defined in more general terms by criticism and theory, and historical genres are defined as genres actively used and recognized by critics and readers. This distinction in a way is a starting point for Neale's (1980, [1990] 1995) and Altman's ([1984] 1995, 1999) contributions to a genre theory. Neale (1980) represents screen theory as combining ideological and semiotic-esthetic aspects. But he expands the concept of genre to a very multifaceted process. Genres are both social and textual institutions and, according to Neale, they work to create a coherence and narrative closure, which positions the viewer safely inside a well-known structure: 'a genre is, simultaneously, a coherent and systematic body of film texts, and a coherent and systematic set of expectations' (Neale 1980, p.37). Altman ([1984] 1995) also clearly continues the semiotic project of genre definition, but at the same time seeks a merging of the ritual and ideological approach and the more general description of genres as a grammar and language. His suggestion in the article is

thus to combine the semantic approach to genres, defining its 'building blocks', and a syntactic approach defining the actual structures in which the semantic elements are placed. But at the same time he argues for a stronger historical complexity in the study of genres, and for a stronger emphasis on the study of genres in historical contexts and to actual interpretative communities.

This line of argument focusing on reception and the historical process of 'genrification,' and genre as a kind of 'game' between critics, producers, and audiences in a concrete historical process, is strongly underlined in Neale ([1990] 1995) and Altman (1999). Taking some historical examples of how different early film genres were defined in industrial and critical terms and how long it took for well-established genres to stabilize as both institutional and audience categories, Neale argues for genre studies with more real historical integrity, actually studying the discourses on genre used in a historical context, rather than only constituting general, theoretical genre concepts. Along the same lines, Altman focuses very much on the unstable and fuzzy nature of genres and the differences in audiences' readings of individual genres and individual films. He adds a pragmatic dimension to his earlier semantic-syntactic dimensions to include the user-contexts aspect. So key terms in this reception-oriented historicizing of film genre theory are different readings, multiple users, and conflicting understandings of genres, instead of genres understood as relatively stable categories based on both general textual similarities and a social and cultural contract or common understanding.

1.5 Cognition, Emotion, and Genre

A move in the opposite direction back to a more general definition of genres can be found in the strong new trend toward cognitive and pragmatic studies of genres in the 1990s. The earlier semiotic and ritual sociocultural approach to genres defined dominant structures based on more general patterns of mental, generic procedures, that partly explain why mainstream genres are so strong and dominant. This point of view was based on structural arguments that were not very firmly anchored in a psychological dimension. Cognitive-pragmatic genre studies also define very basic patterns and schemata in our active construction of narrative forms and basic generic formulas. But the cognitive-emotional theories focus on more internal psychological dimensions, but do not exclude a pragmatic-historical dimension, since all such mental aspects of communication are situated and embodied in a specific historical and communicative situation. The cognitive-emotional genre theory also opposes the dominant Freudian, Lacanian, or Jungian-inspired psychology in the screen tradition of genre studies. Important contributions to this new cognitive-emo-

tional genre theory is Bordwell's (1985) description of different narrative prototypes and the way they cue the spectator and Carroll's studies of horror (1990) and his later article on emotion and genre (Carroll 1999), in which melodrama, horror, and suspense are described from a cognitive-emotional perspective.

But the most ambitious attempt to formulate a cognitive and emotional theory of genre is that of Grodal (1997). Grodal defines a new typology of very basic genres based on the viewer's active-passive relation with the film, the kind of distanciation or identification involved, and the typical moods and emotional responses activated during viewing: associative lyricism, canonical narratives of action, obsessional paratelic fictions, melodramas, horror, schizoid fictions, comedy, and metafiction. Behind these basic forms we then find a number of more concrete, historical subgenres. The basic assumption of the book is that genres of visual fiction are mental structures that to a large degree are modeled on the basis of a simulated reality of actions and emotions to which we react with much the same experiences and capacities as in real life experiences. In his book Grodal describes both the typical embodied flow, during the experience of visual fiction with reference to cognitive-emotional psychology, and the reason for the strength and popularity of the prototypical genres of visual fiction and their relation to different mental and emotional activities and experiences. Grodal's book is, of course, not strongly oriented toward the cultural and historical aspects of moving images, but there is no essential opposition between a cultural, historical, and stylistic approach to visual fictions and Grodal's position, as long as one realizes that certain aspects of the experience and processing of moving images cannot be described as culturally constructed in any short-term sense.

2. Film and Genres

How many film genres are there, and how do we define them and distinguish one genre from another? This is the basic question raised by film genre theory, and this is also the question raised in more concrete, historical analysis of particular film genres and their development. Theoretically there are at one end theories that try to define genres as relatively stable and accepted narrative, stylistic, and thematic structures that transcend individual films and clearly influence and determine both production and audience response; at the other there are theories that stress the unstable and processual nature of genres and the fuzziness of genre. Some theories argue for few and very basic generic prototypes with a variety of possible subgenres, that may vary and change historically, whereas practical criticism and the history of film genre categories often show that you can construct genres on the basis of many criteria, for instance on content and dominant

type of narrative action (war films, porno films), on specific stylistic features (film noir, musicals, animated films), on the basis of the expected audience (women's films, children's films), on the basis of emotional response (comedy, melodrama), etc.

Although genres are often related very clearly to the dominance of US cinema worldwide, national variations of genres are also very important and often have more universal influence, e.g., the Asian martial arts films or the British heritage film or costume drama. But although historical, national, and other variations are very common and the number of actual genres defined in some critical context is probably very large, then some basic narrative modes and generic formulas are stronger and more permanent and universal than others.

2.1 Basic Film Forms and Universal Formulas

Kracauer (1960) defined two basic types of film: narrative films (story films), and nonnarrative films (nonstory films) which he divided into 'films of fact' and experimental/avant garde films. He considered narrative film to be the most important and basic form. In a similar way Williams (1984) has suggested that before we talk about specific subgenres, we should perhaps talk about only three, very basic or principal film forms: narrative film, avantgarde/experimental film, and documentary film. This division is also supported by Bordwell's (1985) division of films belonging to the more canonical, classical narration which dominates both Hollywood mainstream formats and national European mainstream films; art cinema narration with its more fragmented and less linear models of narration found in, for instance, European neorealism, New Wave cinema, and modernist and postmodern films; and finally nonnarrative forms such as documentary. All three forms have very distinct institutional backgrounds and are normally very easily recognized by the general audience, because distribution and indexing of art cinema, for example, is linked to special cinemas, festivals, and a special critical discourse, and documentary films have a totally separate distribution and are indexed much more for their special reference to reality.

2.2 Classical Hollywood and Beyond

Very often the classical Hollywood genres form the basis of what is generally understood as the fundamental genres, and these would most typically include melodrama, comedy, action-adventure, science fiction, horror, Western, film noir, crime genres, and musical. But each of these categories has subcategories, such as the crime genre, which is often further divided into gangster films, thrillers, hardboiled detective films, spy films, police films, etc. A popular

Internet site like the Internet movie database lists 11 general genre categories: action and adventure, art house and international, classics, comedy, drama, horror and suspense, kids and family, musicals, science fiction and fantasy, special interests and studio specials, and documentary. But each of these categories again has a large number of subcategories. Action and adventure, for instance, includes war films, Westerns, and disaster films. Neither the historical, critical, nor the popular taxonomy of genres are defined in very rigid and stable forms, although some basic forms are clearly present. As indicated in the ritual approach, the cognitive approach, and in most studies of the classical Hollywood system, it is however possible to define a number of more general generic forms and prototypes, which then often have different subforms and furthermore can change and vary historically.

Several models of typical historical developments of genres have been suggested in film genre theory. Neale ([1990] 1995), for instance, discusses the evolutionary model, used by Schatz (1981). In this model genres move from an experimental stage to a classical stage and to stages of more opaque, baroque, ironic, or even metatextual modes in which the basic formula elements are so familiar that directors and audiences have a more playful and reflexive relation with them. Neale points to another model in which genres occupy shifting positions as dominant, and run through phases of canonization, automation, and reshuffling. In recent theories on genre mixing and intertextuality it is furthermore often indicated that it is an increasingly common strategy by both the producer and the director to try to mix genres, to have maximum audience effect, and that furthermore the general audience today is much more used to genre mixing and intertextual and metatextual elements even in mainstream films.

2.3 Nonfiction Films and Genre

Many theories and discussions of film genres relate mainly to fictional film; however, both nonfiction film genres and the distinction between fiction and nonfiction are very important for the understanding of the fundamental film genres. Historically the term documentary was defined by John Grierson in 1926 as 'creative treatment of actuality.' This definition points to the fact that documentary films, like all nonfiction genres, have a special and more direct reference to reality than even the most realistic fictional film, but that on the other hand all documentary films are also a product of rhetorical and esthetic choices. Plantinga (1997) gives a pragmatic definition of the distinction between nonfiction and fiction: Nonfiction is related to those discourses that make explicit claims about reality and fiction to discourses that do not, a distinction between an assertive and a fictional mode, in which the

spectator treats the reference from text to reality in two fundamentally different ways.

Historically, esthetically, and rhetorically the documentary film has developed many different ways of representing and talking about reality. Barnouw (1993) uses different communicative roles to define different rhetorical purposes in documentary: explorer, reporter, advocate, poet, promoter, and observer. Renov (1993) in his theorizing about documentary forms makes a distinction between four fundamental functions of documentary films: to record, reveal, and preserve; to persuade or promote; to analyze and interrogate; and to express. Combinations of each of these communicative roles or purposes of course result in different subgenres of documentary.

Documentary genre theory offers several ways of defining the basic genres, but one of the most widely cited is Nichols's (1991) definition of four fundamental modes of representation: *the expository documentary* (with direct address to the viewer, often through authoritative voice-over and based on explicit arguments); *the observational documentary* (often based on simply edited visual observation and people expressing views in their observed reality, but without explicit commentary); *the interactive documentary* (based on a more active intervention and interviewing into reality, but still without an authoritative voice and comment); and *the reflexive documentary* (with a dominance of more self-reflexive and poetic strategies foregrounding the film's own strategy and relation to reality). If these modes can be taken as basic prototypes, then of course different subgenres can mix these modes and rhetorical strategies and it should also be added that mixing of fictional forms and documentary forms also has a long history. Docudrama and drama documentary or even more more complicated fact-fiction formats have developed strongly in recent film and TV culture adding to the esthetic and rhetorical variations of documentary forms or 'faked' documentary forms (Rosenthal 1999).

See also: Art and Culture, Economics of; Art, Sociology of; Culture, Sociology of; Film and Video Industry; Television: General; Television: Genres; Television: History

Bibliography

- Altman R [1984] 1995 A semantic/syntactic approach to film genre. In: Grant B K (ed.) *Film Genre Reader*. University of Texas Press, Austin, TX, pp. 26–41
- Altman R 1999 *Film/Genre*. British Film Institute, London
- Barnouw E 1993 *Documentary. A History of the Non-fiction Film*. Oxford University Press, New York
- Bazin A 1971a The evolution of the Western. In: Bazin A (ed.) *What is Cinema? II*. University of California Press, Berkeley, CA, pp. 149–58
- Bazin A 1971b The Western: Or the American film par excellence. In: Bazin A (ed.) *What is Cinema? II*. University of California Press, Berkeley, CA, pp. 140–49
- Bordwell D 1985 *Narration in the Fiction Film*. University of Wisconsin Press, Madison, WI
- Braudy L 1976 *The World in a Frame: What We See in Films*. Anchor, New York
- Buscombe E [1970] 1995 The idea of genre in the American cinema. In: Grant B K (ed.) *Film Genre Reader*. University of Texas Press, Austin, TX, pp. 11–26
- Carroll N 1990 *The Philosophy of Horror, or Paradoxes of the Heart*. Routledge, New York
- Carroll N 1999 Film, emotion and genre. In: Plantinga C, Smith G M (eds.) *Passionate Views. Film, Cognition, and Emotion*. Johns Hopkins University Press, Baltimore, pp. 21–48
- Cawelti J 1976 *Adventure, Mystery and Romance*. University of Chicago Press, Chicago
- Clover C 1992 *Men, Women and Chain Saws: Gender in the Modern Horror Film*. Princeton University Press, Princeton, NJ
- Comolli J-L, Narboni J [1969] 1992 Cinema/ideology/criticism. In: Mast A, Cohen M, Braudy L (eds.) *Film Theory and Criticism*. Oxford University Press, New York, pp. 682–9
- Cook P, Bernink M (eds.) 1999 *The Cinema Book*, 2nd edn. British Film Institute, London
- Gledhill C 1987 *Home is Where the Heart Is. Studies in Melodrama and The Woman's Film*. British Film Institute, London
- Grodal T 1997 *Moving Pictures. A New Theory of Film Genres, Feelings and Cognition*. Oxford University Press, Oxford, UK
- Kaplan A [1978] 1998 *Women in Film Noir*, rev. edn. British Film Institute, London
- Klinger B [1984] 1995 Cinema/ideology/criticism revisited. In: Grant B K (ed.) *Film Genre Reader*. University of Texas Press, Austin, TX, pp. 74–91
- Kracauer S 1960 *Theory of Film*. Oxford University Press, London
- Mulvey L [1975] 1992 Visual pleasure and narrative cinema. In: Mast A, Cohen M, Braudy L (eds.) *Film Theory and Criticism*. Oxford University Press, New York, pp. 746–58
- Nichols B 1991 *Representing Reality*. Indiana University Press, Bloomington, IN
- Neale S 1980 *Genre*. British Film Institute, London
- Neale S [1990] 1995 Questions of genre. In: Grant B K (ed.) *Film Genre Reader*. University of Texas Press, Austin, TX, pp. 159–87
- Plantinga C 1997 *Rhetoric and Representation in Nonfiction Film*. Cambridge University Press, Cambridge, UK
- Renov M (ed.) 1993 *Theorizing Documentary*. Routledge, New York
- Rosenthal A (ed.) 1999 *Why Docudrama? Fact-Fiction on Film and TV*. Southern Illinois University Press, Carbondale, IL
- Schatz T 1981 *Hollywood Genres: Formulas, Filmmaking and the Studio System*. Random House, New York
- Todorov T [1970] 1975 *The Fantastic*. Cornell University Press, Ithaca, NY
- Wellek R, Warren R [1949] 1956 *Theory of Literature*. Harcourt, Brace & World, New York
- Williams A 1984 Is a radical genre criticism possible? *Quarterly Review of Film Studies* 9(2): 121–5
- Williams L 1989 *Hard Core: Power, Pleasure, and the 'Frenzy or the Visible.'* University of Berkeley Press, Berkeley, CA
- Wright W 1975 *Sixguns and Society: A Structural Study of the Western*. University of California Press, Berkeley, CA

Wood R [1977] 1995 Ideology, genre, auteur. In: Grant B K (ed.) *Film Genre Reader*. University of Texas Press, Austin, TX, pp. 59–74

I. Bondebjerg

Film: History

As the art of film enters its second century, film historians not only encounter a longer and more heterogeneous film-historical past, but do so from within contexts unimaginable during the 1930s, when an important wave of book-length film histories appeared. Most notably, film history is now a field of study within higher education, with specialized journals, professional organizations, and conferences. An indication of film history's evolved condition is the extent to which film-historical research today proceeds through explicit revision of the findings and assumptions of established film histories.

Contemporary film historiography's self-critical orientation, together with the discovery and restoration of lost films, has enabled a fundamental rethinking of key periods of film history. Examples include recent studies of silent-era cinemas that had preceded or outsurvived the formation of Hollywood's so-called classical cinema. In attending to silent cinema's differences from later Hollywood practice, the new historiography has complicated established interpretations of film history as a progressive movement toward worldwide adoption of classicism. New interpretations of other periods in film history are certain to occur as films continue to be found, identified, reconstructed, and made available to scholars, especially films produced outside of film history's familiar research domains of Western Europe and the USA. Indeed, as new case studies accumulate, it is questionable whether world cinema can be said to have a history, if the term history is understood to imply a single narrative that might encompass the heterogeneity of the world's films and film practices.

1. *Film as Art and Industry*

Critics have often distinguished film from traditional arts, such as painting, music, and literature, by insisting that film is simultaneously an art and an industry. This familiar claim points to differences between the industrialized processes of filmmaking and the craft-based traditional arts that have differentiated the historiography of film from that of the other arts. Here three basic points can be made. First, as products of an industrial system, films lend themselves to a variety of forms of analysis, ranging from traditional art-historical approaches, centered on issues of esthetics and filmmakers' biographies, to

diverse economic, technological, sociopolitical, and quasi-anthropological investigations (see Allen and Gomery 1985, Lagny 1992). Indeed, film history's object of study is rarely defined exclusively in esthetic terms, as is evident in the use of nonesthetic events, such as the end of World War II, to demarcate periods in film history. Even film histories concerned mainly with esthetic questions typically also situate films in economic, technological, and social contexts. In fact, certain film histories conceptualize film primarily as a technological, economic, and/or social phenomenon, to the point of excluding esthetic considerations altogether. Such has been the case, for instance, in studies of the economics of national film industries.

Second, the evidently industrial nature of film production has encouraged modernist conceptions of historical change, according to which film history moves forward, in a single direction, with new films necessarily counting as a technical advance over old films. The histories of the established fine arts tradition have been interpreted in terms of cyclical patterns of change, according to which artistic movements rise, fall, and emerge again, in the manner of the life cycle of a natural organism. In contrast, filmmaking, so evidently a function of technological, industrial, and commercial conditions and forces, also invites conceptualization as an ongoing, forward-moving process, forever open to additional improvement. This conception of historical change as a modernizing process underlies both the film industry's celebration of the latest advances in film technique and critics' nostalgia for the lost artistry of the filmmaking of earlier periods.

Third, film history's progressivist outlook has been supported by the relatively short duration of the cinema's history. The histories of painting, poetry, and music have typically been studied irrespective of present-day artistic developments. Film history, however, has been concerned with an art whose past coincides with a recent and relatively short span of historical time. Thus, film historians have often examined the cinema's past in light of its latest trends. Indeed, on occasion, current developments in film practice have enabled new directions in film-historical inquiry. For instance, during the 1930s, the conversion to sound cinema had made it seem as if 'silent' films belonged to the historical past, thereby providing justification for the creation of film archives in Germany, the UK, the USA, France, and elsewhere.

2. *Beyond Classical Cinema*

Film history often displays a retrospective orientation in which the cinema's past comes up as a prelude to the cinema of the historian's present. Examples can be found in film histories that understand world cinema as having evolved toward the adoption of the so-called classical film style associated with studio-era Holly-

wood. Such adoption is generally taken to have occurred by the late 1930s, when filmmakers, in film-producing countries throughout the world, began making sound movies in imitation of Hollywood's films and filmmaking methods. Popular throughout the world, Hollywood films appear to have met transcultural expectations concerning story structure; moreover, it has been argued that the style of Hollywood films conforms to universal perceptual dispositions (see Bordwell et al. 1985). For instance, in a classical film, the duration of a shot is never too short or too long, but instead just long enough for the viewers, regardless of their cultural experience, to register the shot's significance to the unfolding of the film's narration.

Although Hollywood's studio system had collapsed by 1960, classicism continues to provide the context for inquiries into recent narrative cinema, as if classicism's apparent conformity to naturalized perceptual and cognitive habits has made it difficult to see how mainstream film practice might evolve beyond it. Yet since the 1970s, such evolution appears to have been happening, as the contemporary commercial cinema, with its emphasis on spectacular effects and tendency toward episodic narration, increasingly exhibits nonclassical attributes. Thus, following the pattern whereby today's film trends prompt reconsideration of the cinema's past, the nonclassical salience of current mainstream cinema has served to cast new light on established areas of film-historical inquiry. Such areas have begun to include studio-era Hollywood itself, as is evident in recent critical studies of the nonclassical aspects of Hollywood genres such as the musical, film noir, melodrama, and comedy (see Altman 1999).

3. A Revised Chronology

Since the 1970s, film historians, working in various research fields, have attempted to understand past films and film movements as if classicism's status as the pre-given *telos* or goal for world cinema can no longer be taken for granted. An indication of this development can be seen in a basic shift in historical method. Film historiography has traditionally displayed what might be called a 'vertical' interest in the position of a film, or a film-historical period, within a developmental sequence. In contrast, the new historiography can be said to be defined by 'horizontal' approaches, in which films are analyzed as embedded within contemporaneous contexts of production, exhibition, and reception.

3.1 Early Cinema, 1895–1906

Recent research on the early cinema of the period from 1895 to about 1906 provides a prototypical example of

contemporary film historiography's horizontal orientation. The period of early cinema begins in the mid-1890s with the first screenings of films by inventors of motion picture cameras and projectors. Canonical examples include screenings by the Lumière brothers in France and the Skladonowsky brothers in Germany in late 1895, and, in early 1896, by Birt Acres in London and Thomas Edison in the USA. The period continues until roughly 1906, when the sites of film exhibition were no longer mainly fairgrounds, vaudeville houses, and rented theatres, but came also to include nickelodeons and other permanent venues intended specifically for film screenings.

Accounts of early cinema typically center on the development of techniques of fiction filmmaking, especially devices of film editing. The films of inventors such as Edison and the Lumières were made up of only a single shot, usually lasting the length of a standard reel of film. But by 1899, in response to a greatly increased demand for films, film companies in France, the UK, and the USA began producing relatively lengthy narrative films made up of multiple shots. Historians have analyzed these films as attempts to solve the problem of how separate shots might be staged and connected together so as to create the impression of a continuous story. In established accounts, early multiple-shot films such as *The Great Train Robbery* (made by Edwin S. Porter in 1903) mark important moments in the development of the system of stylistic conventions associated with studio-era Hollywood (see Jacobs 1939).

Since the late 1970s, however, a very different history of the cinema of 1895–1906 has been written (see Elsaesser and Barker 1990). Instead of examining early cinema as a forerunner of the commercial narrative cinema of the late 1910s, the new historiography has analyzed early films in the context of their function within vaudeville routines, illustrated song slides, public lectures, and other forms of popular spectacle and fairground-like attraction. Examining how live music, spoken narration, and other exhibition practices affected the viewing experience, recent studies of early cinema bring forth dimensions of early films occluded within traditional developmental accounts. An example is the theatrical or presentational aspect of early cinema, which became a focus of historical study in the 1980s. Investigating early cinema's practices of direct address, adapted from vaudeville, the fairground, and other forms of popular theater, historians have come to understand early films less as primitive attempts at classical storytelling and more as performances within the context of live, spectacle entertainment at the turn of the century.

3.2 The Transition to Classicism

The cinema of the period 1906–1917 has traditionally been conceptualized as a transition from early cinema

to classicism. Key industrial developments include: a shift in editorial control away from exhibitors and toward producers; the creation of exchanges which rented film prints to exhibitors rather than sold them outright; the organization of film production according to a hierarchical division of labor that gave producers authority over directors and camera operators; the establishment of Hollywood as a production center; and the cinema's further expansion worldwide, marked, for instance, by the emergence of Denmark and Italy as major film-producing countries. Basic stylistic changes include: the dominance of the fiction film over the travelogs and documentary 'actualities' that had been an important part of early cinema; a new emphasis on celebrity performers or 'stars'; and a continuing increase in film length, leading, by 1914, to the establishment of the feature-length film as the norm for commercial movie-making.

As with other silent-era periods, established histories of the 1910s have emphasized the development of devices of film editing. Since the 1910s, critics noted that editing enabled forms of acting unique to cinema, in which actors' performances were constructed through a planned assembly of numerous shots rather than captured in a lengthy, uninterrupted run of the camera. During the 1920s, film theorists in France and the USSR argued that editing, in differentiating film from theater, had served to establish film as a singular art.

Accounts of the cinema of the 1910s have long centered on the work of the director D. W. Griffith. Dubbed the father of narrative cinema, Griffith has been credited with having conceived of the scene as a composite of shots, staged, filmed, and assembled so as to construct a unified dramatic world. In some 500 short one- and two-reel films, directed by Griffith between 1908 and 1912, devices such as the closeup, cross-cutting, and the eyeline match are employed as elements within what film theorists have described as a storytelling 'language' unique to cinema. Films such as *The Lonely Villa*, made by Griffith in 1909, feature elaborate last-minute rescues, made up of dozens of shots. Such sequences include events occurring in as many as three separate locales, cut together in rhythms that accelerate or slow down in relation to the intensity of the dramatic action. Innovations associated with Griffith became essential to US film practice, thus providing the basis for the classical style that would take form in Hollywood circa 1917.

During the 1990s, some historians have challenged the traditional emphasis on editing practices, arguing that proper appreciation of the cinema of the 1910s requires attention to techniques of *mise-en-scène*, i.e., the staging of dramatic action, through arrangement of actors, settings, and props within the camera's visual field (see Brewster and Jacobs 1997). Such attention seems especially warranted in the case of films made outside the USA, where innovations in film technique often were prominent in areas other than

editing. For instance, recent archival discoveries have shown that the film industries of Italy, Russia, France, Denmark, Sweden, and Germany produced hundreds of narrative films whose stories, in crucial respects, were told through techniques of staging. With scenes that play out in extended long takes, European films of the 1910s commonly feature far fewer shots than comparable US films, instead presenting story events through careful choreography of multiple actors' movements relative to their positions within the film's frame. The ensemble performances, in which several dramatic actions may occur simultaneously within a single framing, serve to situate character interactions within a shared environment. Thus, the narration of European films of the 1910s might be said to exhibit a psychological or social dimension unavailable to the contemporaneous US cinema's stress on linear, physical action.

3.3 The Decade of National Cinemas

By 1918, with the operation of Europe's film industries having been interrupted by World War I, Hollywood had established itself as the world's foremost film exporter, with distribution branches in South America, Australia, the Far East, and Europe. In response to Hollywood competition, Europe's film industries imposed quotas and other trade restrictions on Hollywood films, and drew inspiration from modernist trends in the arts to invent distinctive national film styles. The major art cinemas of the 1920s have long been a focus of film historiography. Such cinemas include expressionism and the *Neue Sachlichkeit* ('new objectivity') in Germany, impressionism and surrealism in France, and constructivism in the USSR. In France, where there was no clear industrial division between commercial cinema and art cinema, avant-garde techniques were integral to many mainstream films.

The experimental film movements evolved within a new international film culture, organized around specialized journals, film clubs, and art-cinema theaters. An unprecedented intellectual interest in film became manifest in a proliferation of theoretical writings and manifestos. Theorists stressed the need to overcome the apparent theatricality of established film styles, while numerous filmmakers, in Europe and elsewhere, undertook radical explorations of film's artistic potential. Thus, the films of the major European film movements of the 1920s frequently exhibit far more editing technique than had earlier European films. The new art films also often featured experimental cinematographic devices, from filtered lenses to superimposed images, double-exposures, and split-screen compositions.

Certain European film movements explored film's potential for radical alternatives to conventional scenography. Prominent examples include: the mon-

tage films of the USSR, such as Sergei Eisenstein's *The Battleship Potemkin* (1925) and Dziga Vertov's *Man with a Movie Camera* (1929), famous for their radical experiments in film editing; and French Impressionist films, from Louis Delluc's *Fièvre* (1921) to diverse films by directors such as Germaine Dulac, Jean Epstein, and Abel Gance, who employed multiple superimpositions and other optical effects to simulate diverse moods and affective states. Also important were German films such as Robert Wiene's *The Cabinet of Dr. Caligari* (1920) and Fritz Lang's *Metropolis* (1927), acclaimed internationally for their spectacular, studio-created settings. In addition, German films such as F. W. Murnau's *The Last Laugh* (1924) were known for the use of optical effects and elaborate moving-camera shots to suggest a character's embodied or psychological 'point of view.' By the mid-1920s, with increasing numbers of films and filmmakers crossing national borders, an international avant garde had taken form, encompassing filmmakers in Europe, North America, and Japan. 'International' films, such as Teinosuke Kinugasa's *Page of Madness* (1926) and Carl Dreyer's *La passion de Jeanne d'Arc* (1928), confounded national labels with their unusual mix of styles.

Questions remain concerning how new research into national cinemas besides those of France, Germany, the USSR, and the USA might alter the historiography of world cinema during the 1920s. For instance, during this period, Japan and India were among the world's largest film-producing countries. Yet, as with many other national cinemas during the years between the world wars, the films were rarely distributed beyond the national and regional film markets, and few of them have been preserved by distributors and archives. Thus, historical assessments of world cinema during the interwar years remain incomplete and provisional.

3.4 The Conversion from Silent to Sound Cinema

The introduction of recorded sound to cinema has attracted much historical analysis and commentary. Beginning in 1927, the unexpected success of Alan Crosland's *The Jazz Singer* (1927) provoked debate worldwide concerning the future of cinema. By the summer of 1928, as the film industry in the USA began its conversion from silent to sound cinema, observers described it as nothing less than an artistic, technological, and economic revolution. Historical accounts of the conversion period have traditionally emphasized various traumas, such as: the sudden collapse of established actors' careers, the flood of new stage and technical personnel into the film industry, the technical difficulties of sound filmmaking, the esthetic limitations of early sound films, the refinancing of the film industry by large banks, and the curtailment of avant-garde experimentation brought about by the high costs of sound-film production. The

impression of crisis encouraged partisan assessments, whether in the form of Hollywood's celebration of sound cinema as a great technological advance or critics' nostalgia for the vanished art of silent film. In either case, film history had seemed to start over again with the introduction of recorded sound.

Concerning film esthetics, sound conversion is said to have produced two main consequences. First, early sound films appeared to signal an artistic regression from the sophisticated art films of the 1920s. The available technologies had restricted the style of many early sound films to relatively long takes, stationary framings, and immobile actors who recited dialog in styles familiar to the popular stage. For many modernist critics, sound cinema amounted to a return to the allegedly theatrical cinema of the years prior to World War I. Established histories of the early sound years have focused on exceptional films such as Rouben Mamoulian's *Applause* (1929), René Clair's *Sous les toits de Paris* (1930), and Fritz Lang's *M* (1931), whose innovations in multitrack recording, postsynchronization, playback, and other techniques are credited with having restored movement to cinema.

Second, the technical demands of sound filmmaking are said to have pushed filmmakers everywhere toward adoption of film classicism, thereby producing a worldwide homogenization of film style. For critics familiar with the distinctive national film styles and transnational film movements of the 1920s, sound films often seemed to belong to a common family, regardless of where they were made. Even the USSR, famous during the 1920s for its counter-Hollywood film culture, was, by 1934, producing films that looked and sounded much like commercial narrative films made elsewhere.

In assessing sound's impact on film style, it is important to distinguish between filmmaking in Hollywood, which had undergone full industrialization prior to sound conversion, and filmmaking in other countries. One issue concerns the industrial conditions for film style. Whatever the personal, technical, and administrative upheaval, sound conversion did not fundamentally change Hollywood's processes of serial manufacture. Thus, sound conversion occurred in Hollywood under circumstances that enabled a continuation of silent-era practice. Most notably, Hollywood's sound-era editing style remained essentially consistent with established methods, involving many of the same techniques for constructing a unified, linear narrative. After a brief period of experimentation, even the new sound techniques were brought into line with classical precedent (see Crafton 1997). For instance, in 'backstage musicals' such as Harry Beaumont's *The Broadway Melody* (1929), song-and-dance sequences no longer came forward strictly as recordings of self-contained stage performances but instead functioned within narratives concerning the offstage lives of show-business personalities. By the summer of 1931, with

the industry-wide adoption of multitrack recording and mixing, the foregrounding of sound technique as an artistic novelty had been definitively abandoned in favor of the stylistic balance characteristic of Hollywood films of the late silent period.

In many other countries, however, sound conversion had entailed fundamental industrial change, and hence radical consequences for film esthetics. Small countries such as Canada were unable to sustain the industrial production of feature films, and instead came to produce mainly short films, such as documentaries and animated films. In countries such as India and Mexico, where filmmaking had been minimally industrialized, complete restructuring of the industry occurred, and new, sound-era genres, such as musicals that drew on indigenous entertainment traditions, came to define the national film output.

Research into the worldwide conversion to sound remains in its early stages. Relatively little is known about the output of many important film industries of the period. Recent research on Japanese cinema provides a case in point. Although Japan's large film industry dates from the 1920s, research on the history of Japanese cinema began to flourish only in the 1970s, with the rediscovery of the prewar films of Yasujiro Ozu, Kenji Mizoguchi, and other prominent art-cinema directors. Assessments of world cinema during the 1930s are certain to undergo revision pending further research into other important film industries of the period. In this event, the familiar notion that world cinema became homogenized during the 1930s may prove to have been a function of the restricted corpus of films then available to historians.

3.5 Postwar Internationalism

The decade after World War II is known for the emergence of a revived international art cinema. Prominent institutional developments include: the creation of international film festivals in cities such as Cannes, Venice, Berlin, Edinburgh, and Tokyo; increased government funding of film archives, thereby enabling expansion of the International Federation of Film Archives; new film journals devoted to international coverage; and the opening of special art-house movie theaters in cities throughout the world. These developments had prewar antecedents, and observers noted parallels between the new art cinema and the international avant-garde of the 1920s. Nonetheless, the new film culture differed in the extent to which its global network functioned independently from nation-state institutions.

An example can be found in the worldwide reception of Italian neorealism, whose status as the pre-eminent postwar film movement depended crucially on critical acclaim from outside Italy. Films such as Roberto Rossellini's *Rome, Open City* (1945) and Vittorio De

Sica's *Bicycle Thieves* (1948) drew admiration from filmmakers from every continent. Most notably, in evident contrast to the commercial cinema's linear and formulaic plots, the neorealist films emphasized the acts and events of ordinary life, including chance occurrences, routine, nondramatic actions, and inconclusive chains of events. Neorealist films were also acclaimed for the documentary-like 'realism' manifest in the use of nonprofessional actors, improvised performances, actual locations, and natural lighting.

Neorealism has commanded attention from historians as a beginning for prominent subsequent movements and tendencies in world cinema. Indeed, neorealism is said to have affected filmmaking in Asia, Africa, and Latin America more so than any other film movement of the sound era. For many filmmakers, neorealism offered a coherent, low-budget alternative to the style of commercial cinema, with great potential both for naturalistic social representation and for sociopolitical allegory. Filmmakers who have cited Italian neorealism as a personal inspiration include: Fernando Birra and Tomás Gutiérrez Alea, leading directors during the 1960s in the film industries of Argentina and Cuba, respectively; Satyajit Ray, the famous Bengali director from the Indian subcontinent; and Nelson Pereira dos Santos, founder of the *Cinema Novo* movement in Brazil.

The emergence of innovative film movements in Latin America, Asia, and, by the late 1960s, in Africa, with the Senegalese director Ousmane Sembene, had been conditioned by two basic developments. First, in 1947, movie attendance in North America had begun to fall, and would continue to do so through much of the 1960s. The decline became apparent just as Hollywood faced an important 1948 decision by the US Supreme Court, which required that the five major studios sell off their theater chains, and that both major and minor studios cease a variety of other monopolistic practices. In response to the economic and political adversity, the major Hollywood studios produced fewer films throughout the 1950s, targeted now toward specific markets. Second, during this same period, many countries experienced booms in film attendance, especially in the so-called developing world, where television had not yet emerged as an entertainment alternative. Responding to the demand, major film industries emerged in Egypt and Hong Kong, while the new nations of Burma, Pakistan, South Korea, and the Philippines also began producing films on an industrial scale. Also during the 1950s, India became one of the world's largest film producers, making hundreds of films exclusively for the South Asian film market. By the mid-1960s, some 60 percent of all of the world's feature films were made in Asia, Latin America, and Africa.

Coinciding with the worldwide boom in film production, 'new waves' and 'young cinemas' took form in diverse countries, beginning in the late 1950s and continuing through the 1960s. Young filmmakers

worked to distinguish themselves from the established filmmakers of their native countries by emulating new currents in world cinema. Most prominent was the French New Wave, known for films such as François Truffaut's *Les quatre cent coups* (1959) and Jean-Luc Godard's *Breathless* (1960), which featured inventive adaptations of conventions drawn both from Hollywood and European genre films and also from the international art cinema. Important new waves also emerged in Japan, Czechoslovakia, and, indeed, in virtually all countries, in one form or another, during the 1960s. A famous example is the Young German Cinema, whose self-conscious opposition to Germany's established film culture opened the way for the international prominence during the 1970s of directors such as R. W. Fassbinder, Werner Herzog, and Wim Wenders.

3.6 Film Historiography Since the 1970s

The new film culture of the 1960s would have a profound impact on the writing of film history. An example can be found in the notion of the film director as artist or *auteur*. This notion had been formulated in Paris in the 1950s by a group of young film critics, a remarkable number of whom would go on to become key directors of the French New Wave. These critics argued that film directors should be seen as performing the kind of creative role traditionally accorded to novelists, playwrights, painters, and composers. A precedent could be found in certain currents in film critical writing during the 1920s, which had recognized filmmakers such as Charlie Chaplin, F. W. Murnau, and Erich von Stroheim as artists. But the auteur critics extended the sphere of art to encompass a remarkable diversity of films, including, in some cases, even low-budget sound-era genre films. Provocatively comparing movie directors to Mozart and Shakespeare, the auteur critics expanded the canon of great film artists to include current commercial directors such as John Ford, Orson Welles, Alfred Hitchcock, and Howard Hawks.

The 1970s might be said to have opened a new era in film historiography. During the 1970s, the integration of film study into university curricula in Western countries opened up new contexts for film-historical research. For instance, the notion of the director as artist came to inform the academic study of film, leading to the publication of numerous monographs on a wide variety of filmmakers. Moreover, professional approaches to film study, involving the critical re-examination of established film history, also took form. A case in point is the new historiography of silent cinema. Whereas previous film historians had often relied on other historians' accounts, together with personal memories of films, the new generation of archivists and professional scholars studied large collections of film prints, comprised of recognized classics as well as many lesser known or forgotten

works. Additional historical sources included a wide variety of print documents, including film reviews, trade publications, film companies' catalogs, and paper prints of films from the US Library of Congress. The result of this research was a thorough reconceptualization of the early silent period. This work, in turn, is likely to serve as an exemplar for historians working on later periods, such as the period of sound conversion, which, like early cinema, demands both empirical research and historiographical reflection.

As the cinema now enters its second century, new technological changes can be expected to enable new ways of seeing the cinema's past. For instance, as computer-generated digital images come to replace the cinema's photographic image, cartoons and other nonphotographic animated films may come to appear central to film history. In any case, as old films continue to be restored and made available to scholars, significant changes in film historiography continue to occur. At annual silent-film festivals, such as those held in Bologna and Pordenone, large bodies of films, representing the output of diverse directors, actors, studios, film movements, and national cinemas, are made accessible to critics and historians. Each year, new discoveries stimulate revision of established scholarship, while at the same time underscoring the incompleteness of the historical record, and hence the certainty that film history will remain a work in progress.

See also: Art and Culture, Economics of; Art, Sociology of; Culture, Sociology of; Film: Genres and Genre Theory; Television: History; Television: Industry

Bibliography

- Allen R, Gomery D 1985 *Film History: Theory and Practice*. Knopf, New York
- Altman R 1999 *Film/Genre*. British Film Institute, London
- Bardèche M, Brasillach R 1935 *Histoire du cinéma*. Denoël et Steele, Paris
- Bazin A 1958 *Qu'est-ce que le cinéma?* Editions du Cerf, Paris
- Bordwell D, Staiger J, Thompson K 1985 *The Classical Hollywood Cinema: Film Style and Mode of Production to 1960*. Columbia University Press, New York
- Brewster B, Jacobs L 1997 *Theatre to Cinema: Stage Pictorialism and the Early Feature Film*. Oxford University Press, Oxford, UK
- Cherchai-Usai P 1995 *Burning Passions: An Introduction to the Study of Silent Cinema*. British Film Institute, London
- Crafton D 1997 *The Talkies: America's Transition to Sound, 1926-1931*. Scribner, New York
- Elsaesser T, Barker A (eds.) 1990 *Early Cinema: Space, Frame, Narrative*. British Film Institute, London
- Gunning T 1991 *D. W. Griffith and the Origins of American Narrative Film*. University of Illinois Press, Urbana, IL
- Hill J, Church Gibson P 2000 *World Cinema: Critical Approaches*. Oxford University Press, Oxford, UK
- Houston P 1994 *Keepers of the Frame: The Film Archives*. British Film Institute, London

- Jacobs L 1939 *The Rise of the American Film*. Harcourt Brace, New York
- Lagny M 1992 *De l'histoire du cinéma: Méthode historique et histoire du cinéma*. A. Colin, Paris
- Salt B 1992 *Film Style and Technology: History and Analysis*, 2nd expanded edn. Starword, London
- Sklar R 1993 *Film: An International History of the Medium*. Prentice Hall and Harry N. Abrams, New York
- Thompson K, Bordwell D 1994 *Film History: An Introduction*. McGraw Hill, New York
- Vasey R 1997 *The World According to Hollywood, 1918–1939*. University of Wisconsin Press, Madison, WI

C. O'Brien

Finalization in Science

The finalization model addresses the question of how internal (cognitive) and external (social) regulations interact in the dynamics of modern science. It assumes that paradigm building in scientific disciplines will eventually lead to stable theories that solve the fundamental problems in the field and are unlikely to be repealed by a Kuhnian scientific revolution (Kuhn 1962). With the advent of paradigmatic theory, research programs in disciplines become responsive to social and political goals on the level of theory formation. Therefore, cognitive autonomy in the sense of a compelling internal logic of the growth of knowledge is a transient feature and hence more characteristic of an immature science than a mature one. (The finalization model is explicated in Böhme et al. 1983; it has been reassessed in contributions to a symposium for *Social Science Information* Vol. 36, see Leeming 1997.)

1. The Internalism–Externalism Debate

The internalism–externalism debate—a debate that has intrigued historians and sociologists of science since the 1960s—fell into disrepute in the 1980s when mainstream studies of science became preoccupied with the analysis of the social construction and cultural embeddedness of knowledge (see *Laboratory Studies: Historical Perspectives*). These studies have tended to discard any distinction between the cognitive and the social aspects of science (Shapin 1992). However, a (relative) distinction must be upheld if one is to explain not only how science can be shaped by social context, but also how society can be transformed through science (see *Knowledge Societies*).

Modern science emerged from the struggles over cognitive order that waged throughout the seventeenth century, eventually establishing the notion of, and a niche for, the pursuit of natural knowledge which sets empirical facts categorically apart from religious belief, social choice, or political preference (see

Webster 1976). The ‘edge of objectivity’ (Gillespie 1960) is culturally biased: modern science applies an operational (experimental) conception of truth that reflects a social interest in prediction and control. Despite that bias, however, it yields universally valid knowledge. Accordingly, when observations in the laboratory fail to identify any interesting epistemological difference between the pursuit of knowledge and the pursuit of power (Knorr Cetina 1995, p. 151), this tells more about sociology than about science. The finalization model acknowledges the premise of social construction, but still emphasizes the institutional and cognitive differentiation and specificity of modern science as a system of knowledge production. It considers the internal logic of the system—albeit as an emergent property and evolving process.

2. The Rise of Scientific Internalism and its Transition

2.1 Paradigmatic Science

Historically, scientific fields develop through phases in which the mechanisms of internal problem generation prevail. The *locus classicus* of such growth, both in cognitive and institutional terms, are academic disciplines or specialties (e.g., mechanics, electrodynamics, physical chemistry, genetics, biochemistry) that pursue research programs in which the fundamental laws governing the phenomena in their respective object areas are elaborated and confirmed. In this phase of theory formation or paradigm building (Kuhn 1962), frontiers of research are largely determined through explanatory ideals and strategies within science. The flow of problems in understanding the nature and mechanisms of light, electricity, matter, heredity, or cell differentiation regulate the dynamics of the disciplines. The full logic of this internalism is not well understood and remains a matter of ongoing debate among philosophers and sociologists of science (see Hacking 1983, Pickering 1995). The search for more fundamental and more comprehensive theories to explain natural phenomena of increasing complexity seems to be an element of that logic.

2.2 Postparadigmatic Science

The finalization theory adopts Thomas Kuhn’s phase model of disciplinary development (preparadigmatic, paradigmatic, normal science) with one major modification: it assumes that the dynamics of paradigmatic science can lead to mature theories which provide a fundamental and, in a certain sense, conclusive understanding of the phenomena in their domain. To assume that all theories will eventually be falsified or superseded through scientific revolutions, and completely replaced by new paradigms, fails to account for the extraordinary stability of modern theories. For at least

100 years, consolidated theories of motion, heat, light, electricity, matter, or heredity have not been rendered obsolete, despite the Kuhnian revolutions that established new paradigms. Werner Heisenberg (1971) proposed the concept of 'closed theory' to capture the stability of classical mechanics and electrodynamics in the transition to relativity and quantum mechanics (see also Böhme 1980).

To the extent that disciplines achieve theoretical maturity, the core of their dynamics will shift from research that strives for the discovery of new fundamental laws to research that applies these laws to the wealth of phenomena in their domain. Science is a truly 'endless frontier' (Bush 1945) not so much because fundamental problems are multiplied beyond limits and paradigm change is always imminent, but rather because existing paradigms can be extended to an indefinite number of applications. Such applications are termed 'postparadigmatic' science in the finalization model.

2.3 *The Social Direction of Theory Development*

Applying paradigmatic theories is not a simple exercise in deduction. The general principles of hydrodynamics or quantum mechanics or organic chemistry are valid for fluids or solid matter or the processes of plant nutrition, respectively, but they cannot be applied to them without a series of steps of theory specialization through which the principles are extended and adapted to the complexity of real objects. These steps involve such fundamental achievements as the boundary layer concepts in hydrodynamics, or the band concept in solid-state physics or the metabolic cycles in biochemistry. Postparadigmatic science (Kuhnian 'normal science') is not puzzle solving in a trivial sense, it is theory development at the frontiers of basic research. The point the finalization model raises is that internal regulation of science becomes weak at these frontiers. While problem generation may flow from the natural complexity of real phenomena in the domain of general theories, it can also flow from the task of coping with the boundary conditions defined by technical and social tasks. For example, fluid mechanics had been developed into a theory of flow around airplane wings by Prandtl early in the twentieth century; more recent efforts in fusion research extended the theories of magneto-hydrodynamics to the special conditions required for controlled fusion in a reactor: stability and containment of a plasma in toroidal geometry.

3. *Limitations of the Model*

The finalization model is biased on physics as the exemplary scientific discipline and on theory building as the prevalent pattern of scientific internalism. While

the model may cope with the fact that there are other internal frontiers in physics, such as the exploitation of instruments for discovery (Hanson 1965), a critical point is that the meaning of theory may vary among disciplines. Such variation will be reflected in what sociologists describe as the 'epistemic culture' of a field (see Knorr Cetina 1999). Further analysis is needed to determine whether models of scientific growth derived from the case of physics are transferable to genetics or physiology or ecology. The case of agricultural chemistry suggests that theoretical maturity may not be a necessary condition for goal oriented theory development (Krohn and Schäfer 1983).

Finalization is not the only option for the development of disciplines that have reached the paradigmatic stage. 'Normal science' can also pursue the extension of available theories to objects of purely academic interest. In addition, there is also the possibility that the paradigmatic dynamics in the discipline is renewed; non equilibrium theories that go beyond classical mechanics are one example. Such renewal does not challenge the assumptions of finalization as long as new theories do not supersede established paradigms, but indeed provide a more precise understanding of their domain of validity.

Applied science must not wait for finalization to become feasible. To a great extent the mobilization of science for external objectives, today as in the past, does not occur on the level of targeted basic research and theory development; rather, less ambitious strategies are pursued that aim at the quick exploitation of findings and methodologies available in science. Application strategies are pervasive in science. They even intrude disciplines in the paradigmatic phase. However, it seems that the priority and the logic of paradigm building remain unaffected in this case. A study of the relation of early molecular biology to medicine indicates that external orientation can operate as long as it converges with the internal frontiers of the discipline. In addition it is possible to transfer the findings and techniques of basic research to contexts of application (Hohlfeld 1983).

Finalization is a pattern of goal orientation in science that becomes feasible with the growth of knowledge in scientific disciplines. However, critical social and political demands for scientific knowledge often transcend the boundaries and resources of existing disciplines. Ecological prognosis of climate change resulting from the burning of fossil fuels is a case in point. While such demands may be rejected as 'transscientific' from the perspective of traditional disciplines (Weinberg 1972), they do find resonance in what has been called 'post-normal science' (Funtovicz and Ravetz 1993). This is basically a recursive effort to combine scientific analysis with political judgement of how one should responsibly proceed in view of the fact that hypotheses about imminent danger cannot at present be tested by scientific means. See also van den Daele and Krohn (1998) as to how the methodology of

research is translated into a policy design for dealing with the inherent risks of the implementation of new technology.

4. The Science for Society Perspective

Ever since Bacon and Descartes, modern science has been hailed as a ready tool for social and political goals. Historically, however, it took time until the sweeping promises began to materialize. Now, at the beginning of the twenty-first century, the applied mode of science seems to dominate (Gibbons et al. 1994). Pressure to cooperate with various consumers of science, mainly industry, is extended even to academic research. Such pressure does not just reflect a change of political climate or the emergence of a new narrative of the science–society relationship (Elzinga 1997): science itself has become more susceptible to external direction. The finalization model provides an understanding of this susceptibility by analyzing the supply-side of science, so to speak. With theoretical maturity, social and political objectives can be pursued through targeted research at the core of scientific disciplines. Postparadigmatic sciences flourish in many fields of physics (National Academy of Sciences 1972, p. 404); the same would hold for chemistry and considerable parts of molecular genetics. The advancement of science is bound to bring more disciplines to a postparadigmatic stage.

Finalization affects the social relations of scientific disciplines. New issues of responsiveness and accountability arise. Social goals for basic research may be defined by scientists sensitive to social issues, but must ultimately be justified in political, not scientific, terms. The generation of problems in postparadigmatic research can be easily converted to a political process that entangles scientists in negotiations with stakeholders in society over the future development of their fields. Finalization, however, implies a commitment to theory formation. While bound to social objectives, it thus operates at some distance to problem definitions derived from the immediate instrumental needs of companies, governments, or regulatory agencies. Such distance may be needed if the true potential of science for society is to unfold. Theory building provides broader horizons of both explanation and technology, in light of which present definitions of problems and goals may have to be reconsidered.

The finalization model suggests that the distinction between social context and cognitive contents is indispensable for the analysis of the social relations of science. Within a conceptual framework that reduces knowledge to social construction and discards any reference to scientific internalism, one cannot account for the changing patterns of scientific growth and their impact on society. These points deserve more consideration in contemporary sociology of science. To

that end, it is not enough to provide a new language that can be applied equally to natural events, experimental data, and theoretical arguments on the one hand, and to social actors and processes on the other (see Latour 1993). The notion that knowledge is categorically separate from power, interest, and social relations must be rehabilitated. This will not jeopardize the social studies of science: it will only eliminate some of the misleading epistemological claims from sociologists who observe science, and perhaps spare us more ‘science wars’ (Gross and Levitt 1997) from researchers who do science.

See also: Innovation, Theory of; Kuhn, Thomas S (1922–96); Science and Social Movements; Scientific Culture; Truth and Credibility: Science and the Social Study of Science; Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects

Bibliography

- Böhme G 1980 On the possibility of closed theories. *Studies in the History and Philosophy of Science* **11**: 163–72
- Böhme G, van den Daele W, Hohlfeld R, Krohn W, Schäfer W 1983 Finalization in science. The social orientation of scientific progress. In: Schäfer W (ed.) *Boston Studies in the Philosophy of Science*, Vol. 77. Reidel, Dordrecht, The Netherlands
- Bush V 1945 *Science, the Endless Frontier*. US Government Printing Office, Washington, DC
- van den Daele W, Krohn W 1998 Experimental implementation as a linking mechanism in the process of innovation. *Research Policy* **27**: 853–68
- Elzinga A 1997 The science-society contract in historical transformation: With special reference to ‘Epistemic Drift’. *Social Science Information* **36**: 411–45
- Funtovitz S O, Ravetz J R 1993 Science for the post-normal age. *Futures* **25**: 739–56
- Gibbons M, Limoges C, Nowotny H, Schwartzman S, Schott P, Trow M 1994 *The New Production of Knowledge*. Sage, London
- Gillespie C C 1960 *The Edge of Objectivity: An Essay in the History of Scientific Ideas*. Princeton University Press, Princeton, NJ
- Gross P R, Levitt N 1997 *Higher Superstition: The Academic Left and its Quarrels With Science*. Johns Hopkins University Press, Baltimore, MD
- Hacking I 1983 *Representing and Intervening: Introductory Topics in the Philosophy of Nature Science*. Cambridge University Press, Cambridge, UK
- Hanson N R 1965 *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science*. Cambridge University Press, Cambridge, UK
- Heisenberg W 1971 Der Begriff ‘abgeschlossene Theorie’ in der modernen Naturwissenschaft. In: Heisenberg W (ed.) *Schritte über Grenzen*. Piper, Munich, Germany, pp. 87–94
- Hohlfeld R 1983 Cancer research. A study of praxis-related theoretical developments in chemistry, the biosciences and medicine. In: Böhme G, van den Daele W, Hohlfeld R, Krohn W, Schäfer W (eds.) *Finalization in Science*. Reidel, Dordrecht, The Netherlands, pp. 93–126

- Knorr Cetina K 1995 Laboratory studies. The cultural approach to the study of science. In: Jasanoff S, Markle G E, Petersen J C, Pinch T (eds.) *Handbook of Science and Technology Studies*. Sage, Thousand Oaks, CA, pp. 140–66
- Knorr Cetina K 1999 *Epistemic Cultures*. Harvard University Press, Cambridge, MA
- Krohn W, Schäfer W 1983 Agricultural chemistry. The origin and structure of a finalized science. In: Böhme G, van den Daele W, Howfeld R, Krohn W, Schäfer W (eds.) *Finalization in Science. The Social Orientation of Scientific Progress*. Reidel, Dordrecht, The Netherlands, pp. 17–25
- Kuhn T 1962 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago
- Latour B 1993 *We Have Never Been Modern*. Harvester Wheatsheaf, New York
- Leeming W 1997 Revisiting finalization. *Social Science Information* 36: 387–410
- National Academy of Sciences 1972 *Physics in Perspective*, Vol. I. National Academy of Sciences, Washington, DC
- Pickering A 1995 *The Mangle of Practice*. University Press of Chicago, Chicago
- Rheinberger H-J 1997 *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Stanford University Press, Stanford, CA
- Shapin S 1992 Discipline and bounding: The history and sociology of science as seen through the externalism-internalism debate. *History of Science* 30: 333–69
- Webster C 1976 *The Great Instauration, Science, Medicine and Reform 1626–1660*. Holmes and Meyer, New York
- Weinberg A 1972 Science and transscience. *Minerva* 10: 209–22

W. van den Daele

Finance, Geography of

1. Money and Finance

The geography of money and finance is an important part of what money and finance actually are: the spatial pattern of automated teller machines and bank branches, the passions aroused by currency zones and credit ratings, the flow of financial information through fiber-optic cable networks and satellites, the hierarchy of international and offshore financial centers—all attest to this fact. Yet it is remarkable how little understanding of these different geographies existed until comparatively recently. With a few honorable exceptions, the geography of money and finance was neglected until the 1980s. Since then, however, geography has become a major element of the study of money and finance, for three main reasons.

First, the 1980s were a period of major changes in cultural perceptions of money and finance. Whereas previously finance had been seen as the preserve of stolid accountants, solid bank managers, and squalid dealers, in the 1980s, through the offices of film portrayals, television series, and print journalism, it became recast as a dynamic, even sexy, area of the economy.

Second, money and finance had become such a central element of the geography of modern economies that it was impossible to ignore. The ‘financialization’ of the economy, especially the tendency of economies to run to essentially short-term financial criteria such as shareholder value, had become so noticeable that it was no longer possible to envisage money and finance as a subsidiary activity whose role was simply to oil the wheels of production.

Third, geographers produced a series of works on money and finance which has proved the value of a distinctive geographic perspective that emphasizes how space creates money, as well as how money creates space. Foremost among these works were David Harvey’s *The Limits to Capital* (1982), a Marxian analysis of capitalism which included a seminal chapter on money; subsequent books by Leyshon and Thrift (1997a), McDowell (1997), and Clark (2000); and edited collections from Corbridge et al. (1994) and Martin (1999). In turn, this literature has been paralleled by work in politics, sociology, and even economics. In this article, each of four different but related geographies of money and finance is examined in turn.

2. International Finance

Geography first approached money and finance through the lens of *international finance*. Events like the developing-country debt crisis of the 1980s, the developed-country debt crisis of the early 1990s, and the establishment of the euro zone in the late 1990s demonstrated both the power of international finance and its essentially geographical nature. Geographers gravitated to four main areas of work. Not surprisingly, the first of these was the link between international finance and urbanization implicit in the hierarchy of world cities. Work taking place at a number of levels attempted to show the mechanics of this relationship; for example, specific financial markets were related to the success of particular cities. A further, natural extension of this work consisted of investigations into the global hierarchy of offshore financial centers, which, though smaller in scope, are also key features of the international financial system.

The second area of work concerned new financial instruments. In particular, the geography of new forms of options have proved of interest, not least because they have had direct impacts on the urban and regional fabric through various market imperfections and financial scandals (Tickell 1996, 2000, Pryke and Allen 2000).

One more area of research has been international debt. The debt crises of the 1980s galvanized geographers, and led to a continuing stream of work on finance and development. Then, finally, wrapping all of these areas of work together has been the issue of

governance. The shifting balance of power between the national state and financial markets, the growth of supranational regulatory institutions, and the status of world cities as centers of global governance in their own right are topics that underline just what an important area of work this has become (Thrift and Leyshon 1994, Gilbert and Helleiner 1999), one which overlaps, in particular, with the concerns of international political economy.

3. Retail Finance

Geographers gradually became interested in the national- and regional-scale financial systems through the issue of *consumer finance*. That interest has taken a number of different forms. To begin with, there was a general interest in the changing geographies of retail finance: for example, variations in employment in the industry between states in the USA, the hierarchy of domestic financial centers in the UK, the variable gender and ethnic composition of financial services labor forces, and so on. This work has produced a base from which it has been possible to inquire into the dramatic restructuring of financial services, covering subjects as diverse as the rise of call centers (and their urban and regional consequences) and the diffusion of new financial management ideas. In turn, it has also become possible to intervene in a number of policy domains. The most dramatic intervention has been in policies towards financial exclusion, through careful study of the large numbers of poorer people who do not have access to the formal domestic financial system and who are often further disadvantaged as a result (Leyshon and Thrift 1995, 1997b). While the number of those excluded has fallen in many Western economies, the fact remains that this is a pressing problem only made worse by some of the effects of restructuring, such as the closure of bank branches in poorer areas. The result has been both an interest in changing government policy to force the financial system to acknowledge its civic role and a widespread interest in financial institutions able to be used to construct alternatives: credit unions, micro-credit, alternative currencies (e.g., time dollars), small-firm financial schemes, local exchange trading schemes (LETS), and the like.

4. Electronic Finance

The impact of electronic telecommunications on money and finance has been critical. Historically, finance has always been one of the industries most likely to feel the impact of changes in telecommunications, given its emphasis on speed. However, of late, this impact seems to be so great that it is actually redefining what money is by producing new spatial and temporal frames, in rather the same way as in the

changeover from coin to paper money (Leyshon and Thrift 1997a, Pryke and Allen 2000). International finance has been in a process of change since at least the 1960s as a result of continual advances in telecommunications which are leading to nearly instantaneous 24-hour financial markets, trading globally. In turn, these developments are producing changes to the current hierarchy of world cities (e.g., Thrift 2000). An important finding made by geographers has been that the increasing amount of electronic information is not necessarily leading to a decline in face-to-face interaction. Especially in the larger international financial centers, which receive very large amounts of electronic information, face-to-face interaction in tightly defined social networks is vitally necessary to lessen the interpretive load in many markets, as well as to allow new stories about what is happening to be created and put into general circulation. Retail finance is also being changed by electronic telecommunications, and, in the long run, the changes here may be even more dramatic than in international finance. Currently a transition is under way from intermediate technologies such as call centers—which still include human elements—to much more heavily automated systems (for example, employing automated voice recognition to replace call center agents). In turn, this transition may well produce a new geography of web-based banking in which consumers will use intelligent agents to undertake much of their financial business (Leyshon and Thrift 1999, Thrift 2000).

Perhaps the most important outcome of electronic money and finance will turn out to be the growth of a combination of software and databases. The effect of the rise of automated decision systems, combined with large synthetic data sets, on the economic circumstances of most individuals has serious connotations, which are now being explored. For example, the concentration of data in the hands of a few private-sector firms raises obvious issues concerning civil liberties and regulation. Geographers specializing in geographical information systems have been intimately involved in the construction of many of these systems and must shoulder some of the responsibilities for the outcome.

In turn, the advent of this new electronic monetary space has contributed to financial exclusion. It seems likely that cities will 'splinter' even further, with serious consequences for social equality. In particular, a large minority of people will not only be excluded electronically but, given the changing status of money, financially as well.

5. Social and Cultural Issues

Finally, the social and cultural elements of the geography of money and finance have increasingly become visible. As an electronic site of exchange, money has increasingly become an evacuated signifier,

a kind of cultural degree zero. In such a situation, social and cultural aspects become more rather than less important, since they become the crucial identifiers of what counts as money in the absence of money as a tangible object. So, in certain senses, money has proliferated; all kinds of different monies, labeled according to their social and cultural usage, have sprung up.

Such developments have produced considerable geographical interest. To begin with, there has been a large amount of work on the historical iconography of money and finance, and especially the role of money in establishing national identity (see Gilbert and Helleiner 1999). Then there has been work on how money is read through a growing discursive apparatus of television programs, newspaper articles, and the like, which demands 'financial literacy' as a ticket of entry (Leyshon et al. 1998). And there are the social and cultural fault-lines created by money, both within its labor force (as, for example, in the matter of gender and sexuality; see McDowell 1997) and in its ability to produce social and cultural capital for certain elite segments of that labor force, which has enormous spread effects in cities like London and New York (Leyshon and Thrift 1997a).

6. Further Directions of Research

Further directions of research on money and finance are clear. The first of these is the increasing privatization of welfare, with welfare functions that were often produced or guaranteed by the state being run down, leading to a situation in which individuals increasingly have to make their own private arrangements for health and old age. Thus we are seeing the birth of what Clark (2000) has called '*pension fund capitalism*,' which, especially given the general aging of populations in many countries of the world, is becoming a pressing economic issue with profound consequences for the distribution of income, and for life chances. The second direction is *software*. Increasingly, the financial system will be run by machine codes, yet remarkably little is known about the agency of code or how to regulate its production. Here is a pressing theoretical and practical issue. The third direction is *speed*. Increasingly, electronic money is producing instantaneous exchange. There is much hyperbole about this state of affairs but precious little research informing us on how social skills and practices are adjusting and thereby changing what money is. Then, finally, there is the matter of financial exclusion. As already noted, this is a pressing issue which demands the design of different kinds of financial institutions. Geographers have been very active in this area and have been able to demonstrate the full range of competencies that they now have concerning money and finance, competencies that

as recently as the 1980s would perhaps have been regarded as surprising.

See also: Consumer Economics; Financial Institutions in Economic Development; Retail Trade

Bibliography

- Allen J, Pryke M 1994 The production of service space. *Environment and Planning, D. Society and Space* **12**: 453–75
- Clark G L 2000 *Pension Fund Capitalism*. Oxford University Press, Oxford, UK
- Corbridge S, Thrift N J, Martin R L (eds.) 1994 *Money, Power and Space*. Blackwell, Oxford, UK
- Gilbert E, Helleiner E (eds.) 1999 *Nation-States and Money*. Routledge, London
- Harvey D 1982 *The Limits to Capital*. Blackwell, Oxford, UK
- Leyshon A, Thrift N J 1995 Geographies of financial exclusion: Financial abandonment in Britain and the United States. *Transactions, Institute of British Geographers, NS* **20**: 312–41
- Leyshon A, Thrift N J 1997a *Money/Space: Geographies of Monetary Transformation*. Routledge, London
- Leyshon A, Thrift N J (eds.) 1997b Special issue on financial exclusion. *Environment and Planning, A* **29**(3)
- Leyshon A, Thrift N J 1999 Lists come alive: Electronic systems of knowledge and the rise of credit-scoring in retail banking. *Economy and Society* **28**: 434–64
- Leyshon A, Thrift N J, Pratt A 1998 Reading financial services: Texts, consumers and financial literacy. *Environment and Planning, D. Society and Space* **16**: 29–55
- Martin R L (ed.) 1999 *Money and the Space Economy*. Wiley, New York
- McDowell L 1997 *Capital Culture: Gender at Work in the City*. Blackwell, Oxford, UK
- Pryke M, Allen T 2000 Monetized time-space: derivatives: Money's new imaginary. *Economy and Society* **29**: 264–84
- Sidaway J, Pryke M 2000 The strange geographies of emerging markets. *Transactions of the Institute of British Geographers, NS* **25**: 187–201
- Thrift N J 2000 Less mystery, more imagination: The future of the City of London. *Environment and Planning, A* **32**: 381–90
- Tickell A 1996 Making a melodrama out of a crisis: Reinterpreting the collapse of Barings Bank. *Environment and Planning, D. Society and Space* **14**: 5–33
- Tickell A 2000 Dangerous derivatives: Controlling and creating risks in international money. *Geoforum* **31**: 87–99

N. Thrift

Financial Accounting and Auditing

After situating accounting and audit as social practices, this article examines research objectives and methodology in these fields.

Accounting and audit are closely linked social practices. By means of the information it produces, accounting assumes two principal functions: the ren-

dering of accounts and the taking of financial decisions. Standardization and audit then come into play, since accounting information must necessarily be a reliable and believable reflection of the events and conditions it portrays. The initial role of auditing is to assess the reliability and credibility of the accounting portrayal. Taken together and in their complementary roles, accounting and audit are major elements of social regulation.

The resulting importance of accounting and audit in contemporary societies, for better or for worse, has made them the object of active research. Depending on the research methodology employed, research may be conducted from a historical, sociological, or economic perspective.

1. Accounting and Auditing as Social Practices

The origins of accounting have been lost in the mists of time, predating even figures themselves. Indeed, we have been counting for a very long time (and continue, moreover, to do so) without the help of figures. However, generally it is agreed that modern accounting was born at the end of the Middle Ages in Northern Italy with the 'invention' of double-entry book-keeping. A chapter of the Franciscan monk Luca Pacioli's work on mathematics (the *Summa de Arithmetica, Geometria, Proportion e Proportionalita*), published for the first time in 1494, introduces this invention which was to enable accounting to fulfill its present-day functions.

1.1 The Functions of Accounting

In short, let us say that accounting fulfills, and has probably always fulfilled, two main functions: on the one hand, a rendering and control function, and on the other, an aid to decision-making function.

Its first function may be related to the old master-servant relationship. When a servant is given any sort of task or assignment by his master, he must, sooner or later, give a narrative or numerical account of the manner in which he has carried it out. Accounting is precisely one of the instruments which enables a servant to give an account to his master, and the master to check his work.

In the light of this, it is not surprising that modern accounting was 'invented' at the end of the Middle Ages in the big towns of northern Italy: Florence, Genoa, and Venice. Their bankers and merchants, at the head of veritable multinationals, operated all over the Mediterranean and northern Europe. To carry out business activities that were often far away from their home-towns successfully, they were obliged inevitably to rely on agents and representatives that their invention of double-entry records enabled them to control more effectively.

Similarly, it is easy to understand the importance accounting assumed in the nineteenth century when joint-stock companies mushroomed and the split between partner-capital contributors and directors appeared. Partners removed themselves, or were removed, from the running of companies, and directors were obliged legally to provide them with information on their administration at regular intervals by the production of balance sheets and income statements. What financial theory terms an 'agency relationship,' therefore, became a legal requirement. To use agency theory vocabulary, one might say that the partners played the role of principals, and the directors that of agents. Accounting is, therefore, the instrument that enables the principal to control what his agent has done, and the agent to give an account of what he has done. In this relationship, the accountant occupies a somewhat unusual position since he draws up the accounts—destined for the principal—of the agent who must give an account. English expresses this triangle very well since it has three words, to designate respectively he who has an account to give (the accountant), he to whom the account must be given (the accountee), and finally, he who does the accounts (accountant).

This 'accounting game' is obviously a complex social game to the extent that the person giving the account possesses more information than the person to whom the account must be given, and to the extent also that he may be tempted to manipulate the accounts and present them in a way favorable to himself. The accountant, who draws up the accounts, occupies, despite himself, the position of mediator, and so it is a good idea for him to belong to an organized profession with a strong sense of ethics.

Accounting's aid to decision-making function is as old as the preceding one and linked closely to it, moreover. The bankers and merchants of northern Italy, mentioned above, had to join forces to amass the capital needed to finance their businesses. They obviously had to share the profits, and to do so, they needed an instrument which would enable them to calculate what these were. The new double-entry accounting was also capable of responding to this need and naturally still exists in today's big companies. A company's annual accounts allow shareholders not only to keep track of the activities of their directors, but also to take the decision which is, as it were, the high point of their annual general meeting, the decision concerning the allocation of profits.

The two main functions of accounting have taken on a dimension that matches the responsibilities of today's companies. Today, companies must give an account not only to those who have entrusted them with their capital, but also to many other partners such as the Government, their suppliers, customers, employees, and in more general terms, the public. So their accounts (and in particular their balance sheet and income statement) have to satisfy various needs for

information which correspond to very different models and control and decision-making rationales. More specifically, they should, for example, allow:

(a) investors, as previously mentioned, to assess the way the directors administer the business and to take into account their decisions regarding purchasing, and the retention or sale of stocks and shares;

(b) banks and financial establishments to judge the financial solidarity and reimbursement capacity of the company;

(c) employees to analyze its social effectiveness and its durability;

(d) suppliers to understand its capacity to honor its debts;

(e) customers to measure its capacity to fulfill the orders it receives; and

(f) the Government to control its taxable profit.

At this stage, one might say that this type of accounting is the system of information through which a company, or any other organization (including nonprofit-making organizations), gives an account to its economic and social partners, and, in doing so, submits itself to their control and allows them to take decisions. It is because the information it produces is, for the most part, of a financial nature that it is described in this way. Moreover, financial accounting which responds to the need of third parties for information can be distinguished from the type of accounting known as management accounting which responds more specifically to a company director's need for information. If it is to assume its vocation of providing information for third parties effectively, financial accounting must obey recognized rules and the application of these rules must be controlled. This is the role of the audit.

1.2 Standardization and Auditing

In all major developed countries, accounting obeys standards fixed by the accounting profession and/or public authorities. In addition to purely national standards, there are now international standards produced by the *International Accounting Standards Committee* (IASC). The aim of standardizing corporate accounting is to make the financial information it produces intelligible and credible. However, to the extent that it is the companies themselves who do their own accounting, standardization is not an adequate condition for the credibility of their accounts. The accounts must be audited.

Auditing a company's accounts consists of checking that they have been drawn up in accordance with current procedures and standards and that they give a true and fair view of the company's financial situation. Obviously, the audit cannot consist of redoing the company's accounts. The auditor proceeds essentially by means of investigation and cross-checking in order

to identify anomalies likely to call into question the reliability and credibility of the accounts on which they must give an opinion. This opinion, which is the result of these investigations, serves to certify the accounts. Certification can be positive or negative, and accompanied by observations and reservations of varying degrees of importance. It should allow the recipients of the accounts, and in particular investors, to assess their validity.

In the course of the twentieth century, the accounting audit received legal consecration in very many countries. Joint-stock companies must have their accounts audited by independent professionals. In Anglo-Saxon countries, the audit assignment is undertaken by certified (US) or chartered (GB) accountants. These providers of credibility form what the sociologists call a 'profession' in the sense that they take on a task which is in the general interest and are, for this reason, subject to requirements regarding their competence and professional code of ethics. Auditors' codes of ethics place great emphasis on their independence.

The audit has expanded considerably its initial accounting vocation and in the second half of the twentieth century has undergone considerable, protean development, generating a major economic sector dominated at world level by several large consultancies such as Arthur Andersen, Price Waterhouse Coopers, Deloitte Touche Tohmatsu, Ernst and Young, KPMG, (known as the 'Big Five'). In the widest sense of the term, the audit now covers all 'listening' activities ('audit' comes from the Latin 'audire,' to hear), evaluation, supervision, and external control of organizations. To the accounting audit previously mentioned has been added the management audit, the quality audit, the environmental audit, etc. Moreover, all organizations, and not simply companies, make use of the audit.

Together, therefore, accounting and auditing form one of the mechanisms which contribute to maintaining links of collaboration and confidence between the various economic and social players, whether these be individuals or organizations. The multiplication and increasing complexity of the relationships between social players has created an enormous demand for accounting and auditing. One might even say that modern society is an 'audit society' (Power 1997).

2. Accounting and Auditing as Research Subjects

Accounting has never been a fixed, immobile technique. For five centuries at least, if one traces its modern origins back to the Middle Ages, it has adapted constantly to developments in companies and their environment. This adaptation is, in a sense, the expression of a form of applied research that one might qualify as spontaneous, carried out by the

practitioners themselves, heads of companies, and accountants. But it was not until the twentieth century that accounting and audit became objects of research for the social sciences, of interest in particular to historians, sociologists, and economists.

2.1 Accounting History

During the twentieth century, accounting has been the subject of very many historic investigations. All the major academic accounting reviews (*Abacus*, *Accounting and Business Research*, *Accounting-Organization and Society*, *The Accounting Review*) publish articles on the history of accounting. Entire reviews are even dedicated to it such as *The Accounting Historian's Journal* and *Accounting Business and Financial History*. Through these many publications, it would seem that the history of accounting can be approached in several different ways.

Certain historians are interested in accounting as a recording technique and pay particular attention to the genesis and evolution of the double-entry system that still is the most striking aspect of accounting. Other historians wish to see accounting as something more than pure technique and consider it to be an instrument of social and economic regulation. From this point of view, its history is not an autonomous history and is part of several histories: the history of the economic system, the history of companies, the history of techniques, and the general history of ideas.

One of the first to adopt this point of view was the economist Werner Sombart who postulated that there was a very strong link between the development of the capitalist system and that of accounting, going so far as to claim that one was not conceivable without the other, that double-entry accounting allowed the company to be set up as a place of economic calculation, to give a numerical expression to the profit objective, and thence to rationalize the allocation of resources. Although contested, Sombart's thesis has contributed to reviving the history of accounting.

2.2 Accounting Sociology

Since the past 1970s, and above all in Great Britain, the sociology of accounting has undergone major developments. These developments are sufficiently differentiated for two distinct schools of thought to be identifiable within this field of research: an interpretative, and a radical school.

The interpretative school, represented by authors such as Giddens, Morgan and Foucault, aims to interpret accounting practices, in particular national standardization practices, through reference to their technical, economic, social, cultural, and political

context. This type of research tries to highlight the various conditions which influence accounting and which mean that it cannot be considered as an isolated technique. Although it relies mainly on sociological methods, in particular the monograph, it also uses the historical method. Its main organ is the review *Accounting Organizations and Society*.

The radical school, which is Marxist or neo-Marxist in inspiration (with references to Habermas and Marcuse) considers that accounting and auditing are mechanisms designed and implemented by the dominant social groups to defend their interests and strengthen their domination. It is, therefore, a question of dispelling the mystery surrounding the neutrality and objectivity often attributed to them and of highlighting their ideological function. Researchers who position themselves within this trend willingly support the oppressed (poor countries, women, and in the United States, blacks) and lend them their expertise. From a methodological point of view, they make highly eclectic use of social science approaches. Intentionally provocative, and with a leaning towards intellectualism, they go so far as to call upon poetry to illustrate and defend their theses. Their main organ is *Critical Perspectives on Accounting*.

2.3 Accounting Economics

Economic research has given birth to what is commonly known as positive accounting theory. According to Jensen (1983), this is a theory capable of explaining '... why accounting is what it is, and what consequences accounting figures have on people and on the allocation of resources.' At the end of the 1970s, the theory was promoted by two professors from the University of Rochester, Ross L. Watts and Jerold L. Zimmerman.

The research conducted in the wake of these two pioneers was inspired principally by the agency theory and aimed to highlight the determinants of choices of accounting methods made by company directors, where standardization authorizes such choices. In agency theory, the company is a 'nexus' of contracts, formal and informal, concluded mainly between the directors and the stockholders. To the extent that accounting produces information that serves to draw up and control the execution of these contracts, company directors are said to pursue their personal interests and choose methods that give them an advantage.

Within the context of this problem, it is possible to formulate many hypotheses on the accounting behavior of company directors. One might suppose, for example, that directors linked to their company by a profit-sharing contract would choose accounting methods that tend to increase present profits at the expense of future profits. One might also suppose that

company directors whose debt ratios are bordering on the thresholds tolerated by the lenders would once again choose accounting methods that increase the profits and diminish the debt ratios. Such hypotheses may be subjected to the statistical tests common in economic research.

Finally, positive accounting explains the choices of accounting methods made by company directors by reference to various economic and financial characteristics of the companies: their size, level of debt, Stock Exchange listing, etc.

This research trend has been the subject of very many methodological and epistemological criticisms (see, e.g., Christenson 1983), but it remains the mainstream of accounting research in the United States. The resulting research papers are published mainly in the *Journal of Accounting Research* and in the review specially created by Watts and Zimmerman, the *Journal of Accounting and Economics*.

To these various types of research (historic, sociological, and economic), that are cognitive in nature, may be added applied research, the object of which is to define accounting concepts, principles, and procedures. Research aims to define a theoretical framework for accounting practice and is, to a greater or lesser extent, linked to the work of the standardization bodies. In contrast to the preceding types of research, it is described as normative.

See also: Competitive Strategies: Organizational; Conflict: Organizational; Corporate Culture; Corporate Finance: Financial Control; Corporate Governance; Corporate Law; Information and Knowledge: Organizational; Intelligence: Organizational; Intelligence, Prior Knowledge, and Learning; International Business; Monetary Policy; Rational Choice and Organization Theory; Rational Choice Explanation: Philosophical Aspects; Strategy: Organizational; Technology and Organization

Bibliography

- Burchell S, Clubb C, Hopwood A, Hughes A, Nahapiet J 1980 The roles of accounting in organisations and society. *Accounting Organizations and Society* 5(1): 5–27
- Christenson C 1983 The methodology of positive accounting. *Accounting Review* 58(1): 1–22
- Dunn J 1996 *Auditing: Theory and Practice*. Prentice Hall, New York
- Edwards J R 1989 *A History of Financial Accounting*. Routledge, London and New York
- Flint D 1988 *Philosophy and Principles of Auditing*. Macmillan, London
- Gray R, Owen D, Adams C 1996 *Accounting and Accountability*. Prentice Hall Europe, London
- Hopwood A G, Miller P (eds.) 1994 *Accounting as Social and Institutional Practice*. Cambridge University Press, Cambridge, UK
- Jensen M C 1976 Reflections on the state of accounting research and the regulation of accounting. In: *Stanford Lectures in*

Accounting, Stanford University Press, Stanford, CA, pp. 11–19

- Jones T C 1995 *Accounting and the Enterprise: A Social Analysis*. Routledge, London and New York
- Matthew M R, Perera M H 1991 *Accounting Theory and Development*. Chapman and Hall, London
- Power M 1997 *The Audit Society: Rituals of Verification*. Oxford University Press, New York
- Roslender R 1992 *Sociological Perspectives on Modern Accountancy*. Routledge, London
- Ryan B, Scapens W, Theobald M 1992 *Research Method and Methodology in Finance and Accounting*. Academic Press, London
- Watts R L, Zimmerman J L 1978 Towards a positive theory of the determination of accounting standards. *Accounting Review* 53(January): 112–34
- Watts R L, Zimmerman J L 1986 *Positive Accounting Theory*. Prentice Hall, Englewoods Cliffs, NJ

B. Colasse

Financial Institutions in Economic Development

The financial system can be considered as a composite of financial activities and the institutions that implement and regulate these activities. Financing of firms is one activity of the system, and capital (securities) markets and banks are their principal financing sources. The sociology of financing economic development examines the institutions that shape the development of firms and nations—particularly in emerging markets.

There have been several sociological analyses of development, including firm, industry, and country studies; but what remains largely unexplored is the role of banks and other financial institutions in this process, and how this relation is influenced by national and international contexts and events, to say nothing of local populations and their initiatives (cf. Evans 1995, Gereffi and Korzeniewicz 1993). After reviewing some theoretical approaches to financing growth and economic development (development) this entry focuses on banks, the principal institution in financing development in most emerging countries. It will identify studies that are rooted in social structure and that focus on such questions as how banks emerge, the sociopolitical reasons for their involvement, which forms are prevalent, and how development is financed. To conclude, some policy-related research challenges will be posed.

1. The Financial Context

Financing of households, businesses, and the government are the three main financial activities in any country. Financing is mainly achieved through capital

markets and/or banks as well as own savings by economic units. A firm can draw upon its savings to finance its production, it can issue stocks, bonds, or it can borrow from banks. Households usually borrow from banks, while governments issue bonds and/or borrow from banks. Cross-nationally and cross-temporally, there are significant differences in these activities. Businesses in the United States and the United Kingdom use capital markets (securities issues), while those in France, Germany, and Japan mainly use bank loans. Among these countries there are differences in bank-industry relations. German banks are inclined to intervene in industry because, in addition to being important shareholders, they control access to all sources of finance. In France, on the other hand, businesses have access to a wider array of public and private institutions whose lending terms are shaped by the government (Zysman 1983).

There is little consensus over which system contributes most to the overall development of an economy. Evidence indicates that having a robust banking and capital market is correlated with economic development (Demirguc-Kunt and Levine 1999). Levine (1997) reminds us of one classic controversy, i.e., whether Germany's development in the latter half of the nineteenth century was more rapid than Great Britain's because its economic structure was based on banks with close links to industry, whereas Great Britain's was more securities market-based and ties between banks and industry were less intimate. As countries mature and evolve into service-oriented economies, there is a growing belief that capital markets are increasingly more effective at intermediation than banks, as Ness (1974) and Haber (1997) illustrate for Brazilian development. Nonetheless, banks are essential to small and medium sized firms as they are less able to tap capital markets for funding. Similarly in emerging countries where there tends to be a lack of a tradition of investing, where capital markets are absent, or are only beginning to mature, banks are the principal financial intermediaries and have been the primary engines for financing economic development. They are heavily linked through equity and management to industry. In these countries the state may own a good portion of the banks, and share the market with foreign or large domestically owned banks. In China nearly 100 percent of the banks belong to the state; in India 70 percent are state owned. In contrast, in the United Kingdom more than 50 percent of the banks are foreign owned (*Economist* 1997). For an open-banking country in which foreign banks are treated the same as domestic banks, the United States has relatively little foreign ownership (less than 10 percent) of separately chartered banks (Barth et al. 2000). And as the twenty-first century begins, throughout the global economy banks are growing in size and becoming more multifaceted in their operations, often as a result of huge mergers and acquisitions which themselves are fueled by the advent

of new regional markets (Makler 1999). Bank growth is in response to an increasing and rapid demand for huge, often multilateral, financing coupled with banks' desire to provide all or one-stop financial services.

1.1 Banks in Emerging Economies: Some Anomalies

Economists view banks as the principal financial intermediary between savers and investors. Banks obtain deposits from the public and invest what they receive in loans and securities. But their role is broader. They are important centers in the formation of wealth, the distribution of economic power, and they influence the way societies change. Political scientists see banks as a powerful interest group attempting to influence public policy or elections. Anthropologists see banks as 'unorganized,' informal institutions that provide capital to small rural or urban businesses.

Sociologists have focused mainly on bank power and have elaborated Berle and Means' analysis of corporate ownership and control. Few have studied banks as institutions or have explored the meaning of their wide array of relations. Few have related form to performance. Even fewer have focused on emerging countries where banks have played a significant role in change and development, as Mizuchi and Stearns (1994) recognized. This would require comparisons among a diverse selection of countries in addition to those that have a similar standard of living such as Germany, Japan, the United Kingdom, and the United States (Levine 1997).

2. Development and Banks in the Social Sciences

While few studies of the role of banks in growth and development are found in sociology, a number of distinctive theoretical approaches have emerged in the general study of financial institutions that illuminate the conditions under which one or another type of bank will appear and explain what differences this makes for economic development.

In economics, one debate has to do with supply and demand. Some have characterized banks as demand-following, or as passive, permissive intermediaries that merely provide services or loans to creditworthy borrowers. Economists such as Goldsmith (1969), Gurley and Shaw (1967), and Patrick (1966) saw banks as supplying capital, encouraging efficient credit allocation, increasing accumulation, and sometimes assisting in the founding of new enterprises or in promoting mergers, but only upon the demands of their clients. Others viewed banks as supply leading. Among these was Schumpeter (1961), who brought credit creation by banks to the fore, identifying them as supply-leading and growth-inducing agents of development. In creating new sources of credit, banking systems were ranked slightly behind those of

entrepreneurship and innovation. For Schumpeter, well-managed banks fuel technological innovation by identifying and funding entrepreneurs who were identified as most likely to introduce and implement new products and production processes successfully. But in his discussion of commercial banks he simply mentioned the ability of banks to create purchasing power, without analyzing their method of capital accumulation or their relations with entrepreneurs.

It was Gerschenkron (1962) who introduced the importance of context and institutional form. In an often-cited explanation of the role of banks in development, he observed that the historical timing of a nation's industrialization determines which economic actors and social institutions will be involved. One of his principal hypotheses was that the more backward an economy relative to others in the same period, the greater the likelihood that financial institutions would be created to supply capital in advance of demand. In the transformation of once backward countries to industrialization, the role of the state, and the emergence of specialized private and state banks designed to increase the supply of capital and provide better-informed entrepreneurial guidance to nascent industries were also emphasized as important to development. As for form, he saw the pressure for industrialization in latecomers resulting in larger industries and banks. And because of the magnitude of bank lending to industry Gerschenkron predicted that banks would acquire 'a formidable degree of ascendancy over industrial enterprises which extended beyond the sphere of financial control and into that of entrepreneurial and managerial decisions' (Gerschenkron 1962, p. 14). Thus institutional dominance was introduced into his analysis.

Classical Marxist theorists also viewed banks as dominant. Lenin saw banking as concentrating into a 'single collective capitalism,' as predatory, and as controlling industry, causing capital concentration and conglomeration. To Hilferding (1981), finance capital (or that capital supplied to industrialists and controlled by financial intermediaries) was the means by which big banks forged concentration among non-financial corporations and created a new class 'under the direction of high finance.' And through investment in more and more basic sectors, banks propelled the state as their agent into the world market (Hilferding 1981, Harvey 1982). In applying parts of this theory to the United States, sociologists Mintz and Schwartz (1985) and Fennema (1982) found in some periods extensive interlocking between the directorates of major US corporations and banks, that industry was dependent on banks because of bank control over financial resources, capital flows and knowledge of money markets, and that American banks interlocked with important foreign banks to form a discrete consortium in world finance. But like Gerschenkron, neither Hilferding, Lenin, nor more recent scholars have demonstrated successfully that such a strategy

will result in finance capital's economic or political hegemony. To the contrary, in studying bank transformation in Brazil, Garman and his colleagues (1998) illustrate that while banking concentration resulted during privatization there was little evidence of finance capital's domination of the process.

The typical institutional form that emerged in Europe in the nineteenth century was the financial conglomerate or multi-sector company, transacting in different markets owned and controlled by a family-owned bank. In their quest for growth and development, governments have been bankers, have recruited domestic and foreign banks, and encouraged financial conglomeration and financial concentration. The classical cases in financing development are Germany and Japan where, in preparation for World War II, military governments nurtured large banks and encouraged their spread into and control of non-financial sectors in order to develop these sectors (Hadley 1970, Roe 1994). In Japan, the conglomerates were known as *zaibatsus* and today they are referred to as *keiretsus*. In South Korea they are *chaebols* and in Latin America they are usually known as *grupos* (Leff 1978, Granovetter 1994). In most emerging countries where the state is a formidable economic actor, state-owned banks share the market with large family-owned and foreign banks. These banks were recruited by the state, which provided subsidies to encourage the development of basic, agro-industrial, and export industries. As a result, the families achieved economic security and upward mobility, and the foreign investors became more embedded in the economy. The bank, usually at the core of the conglomerate, served primarily as a direct source of capital for member firms that were faced with increased capital requirements posed by industrialization and export markets. It also served as a source of employment for family members and close friends (cf. Gerlach 1992, Lamoreaux 1994). In exchange for the promise of investment, foreign banks also received 'benefices' such as authorization to open branches or to purchase domestic banks. During Brazil's industrialization in the 1970s foreign capital formed alliances with prominent family-owned firms in order to access local markets. In exchange these firms and their controllers were provided with technology, capital, and access to international markets (Evans 1979, Cardoso and Faletto 1979). What soon became known as Brazil's economic miracle was largely fostered by the banking sector, its conglomerates, and its alliances (Makler 2000). The state-banking sector relation in Brazil and other emerging countries is thus more symbiotic than in some advanced economies (cf. Hall 1991, Weber 1968).

Interlocking boards of directors and the resulting exclusivity favor large business growth over small. But this pattern seems to be tempering. Davis and Mizuchi (1999) found that the interlocking between the boards of American banks and major industrial

corporations declined significantly from 1982 to 1994, reflecting a strategic shift by banks away from lending to major corporations. While corporations are turning more to capital markets for their financing, interlocking and, as a consequence, lending is likely to be increasing between banks and smaller firms that are less able to secure funds from capital markets (cf. Barth et al. 2000). In relations between banks and smaller firms, commercial transactions are more likely to be embedded in social attachments that improve loan acquisition, offer favorable credit terms, and motivate smaller firms to achieve, as Uzzi (1999) has shown.

Another topic that has intrigued sociologists is the impact of foreign capital on social structure in emerging nations. Some nations, in their quest for rapid economic development, have encouraged foreign direct investment, including banks, because of the promise of technological and industrial growth that they represent. Foreign banks have been cast as stimulating competition in their sector, better able to withstand macroeconomic shocks because of their geographic diversity, and as relieving scarce regulatory resources because they are regulated by their home country (World Bank 2000). Yet some development economists as well as some sociologists have been dubious about foreign capital's contribution, while others have viewed it more favorably. In studying late nineteenth-century Chile, Zeitlin (1984) found that important family groups that had merged with foreign capital to develop the nitrate industry at first flourished but later were economically and politically weakened when the government, in an attempt to curb an export crisis, threatened nationalization and limitations on further foreign capital participation. However, other studies suggest that associations with foreign capital can foster development, local entrepreneurship, and provide mobility opportunities for local businesses. In comparing Brazil, India, and South Korea, Evans (1995) demonstrates that foreign capital and national governments can perform a nurturing or a midwifery role in the development of certain economies. In an earlier study of Brazil, Evans (1979) found that foreign capital formed alliances with prominent family-owned firms in order to access local markets. In exchange, these firms were promised technology and access to international markets.

There are few sociological studies of foreign banks in emerging countries, and those that exist are inconclusive as to whether banks contribute to development. In the 1990s foreign banks proliferated in Latin America. Guillén and Tschoegl (2000) have documented the dramatic rise of Spanish banks but advocate further research to understand their investment strategies. In Brazil, foreign banks have spread but it is unknown whether their lending, like many of the leading domestic banks, is mostly directed toward larger foreign or state-owned corporations, or whether it is becoming more widespread to include small and

medium sized businesses. In that country, banking secrecy provisions preclude detailed analysis of size and regional distribution of loans, and for their part banks resist studies of their lending, presumably fearful that such information will be used by their competitors (Ness 2000).

Values, politics, and laws underlie this proclivity to conglomerate, to concentrate finance capital, and to allow foreign or outside banks to practice in national markets. In some societies, centralization and concentration of economic and political power are tolerated while in others they are not. Because of their feudal heritage, a populist, anti-big-bank ethos has historically been absent in Japan and South Korea while instead large enterprises are celebrated (cf. Roe 1994, Orru et al. 1991). In several Latin American countries centralized power and large private and public institutions are also preferred. In Brazil, because of its patrimonial heritage, large financial conglomerates flourish, and Brazilian federalism leads development-prone administrations to decentralize financial power to them in exchange for their economic and political tribute (Makler 2000). In contrast, American ideology has distrusted large monopolies and banks and has favored smaller, encapsulated, locally accessible, and locally embedded institutions. Since the early nineteenth century the American banking system has been fragmented, consisting of numerous small banks without extensive branching and a few large banks that have been hampered legally from capitalizing in other financial services (Roe 1994). The mixing of banking and commerce is prohibited, and although banks can now engage in securities and insurance their spread into real estate activities is restricted. The state 'recruits' banks to invest in nonfinancial sectors such as community housing and urban development programs, not through continuing subsidies and dispensations as in patrimonial systems but through regulation.

This ideology is rooted in American federalism that fostered fragmented, decentralized, and 'separate but equal' banking where each state chartered and protected its own banks while excluding those from other states or foreign countries. National legislation echoes this ideology. The National Bank Acts of 1863 and 1864, and the Glass-Steagall Act of 1933 restricted banks from spreading geographically and sectorally into other financial and non-financial sectors, thus leading Roe (1994) to predict that the United States might have achieved greater growth earlier with stronger, less fragmented, money centered banks. Yet eleven Southeastern states closed their markets to outside banks. This gave resident banks a monopoly within the region (known as the Southeast Compact). As a result these banks registered greater profitability, made proportionately more property loans, and for a decade (1980–90) more economic growth occurred in this region than elsewhere in the country (Barth and Jahera 1994). Southeastern growth subsided with the

debut of national banking deregulation and banking liberalization in the 1990s. This legislation repealed six decades of constraining legislation and resulted in nationwide banking, mergers, acquisitions, and the spread of banks into securities, insurance, and other sectors. Concentration is increasing and the multi-sector financial conglomerate is likely to become the dominant organizational form, as in many European countries, Latin America, and other emerging countries. With the advent of electronic banking and nationwide conglomeration, the relation between local deposits and local financing will become increasingly tenuous.

3. Challenges

A growing apprehension with globalization and its effects on the development of national markets is casting an important challenge for sociologists as it did for economists who were concerned with post-World War II recovery half a century ago. Moving from studies of financial structure and power, sociological research at the beginning of the twenty-first century should focus on outcomes of financial intermediation, and these should be analyzed in ways that could inform public policy.

First, the ways that economic units, be they regional markets, nations, or large or small businesses, finance their starts and their growth should be clarified. For mature, more service-oriented economies, capital markets (securities) are viewed as more efficient but for emerging economies and for new businesses banks (loans) are the principal form. Is this the case, or will there be increasing multi-sourcing as access to financial markets is widened through deregulation, technological innovation, and new financial institutions?

Second, monitoring whether major changes in local and national economies are accompanied by changes in financing enables identification of new forms of financial intermediation that finance development. Granovetter (2001) and other scholars (cf. Kaplan 1999, Kenney and Florida 2000), for example, found that the rapid growth of California's Silicon Valley information technology industry depended upon the creation of a new form of financial intermediation: venture capital. Early venture capitalists were mainly successful engineers who invested their profits and often took equity and managerial positions, unlike most financiers who hardly associated with their investments.

Third, in studying emerging countries, recognition of the embeddedness of market segmentation or the duality of the financial market is challenging. Paralleling the two forms of financial intermediation (banks and securities) is informal financing. This is particularly prevalent in emerging markets among poorer populations that have little, if any, access to (or

knowledge of) formal financial institutions that serve large businesses (World Bank 1989, Carstens 1995). Biggart (2001) studied micro credit programs such as rotating savings and credit associations that lend modest amounts of money for entrepreneurial business activities to poor women who are organized in solidarity groups. By starting with the demand side, with credit seekers, pertinent financial institutions can best be identified and from there related to performance. Answering these two questions is a promising start: What is the relevance of one financial institution compared to another for development; and how, in turn, does development change and evolve to affect that institution and its organization? But more challenging is understanding the relation between financial development and social organization. Capital markets create dispersed savers (investors) who are interested primarily in financial return and risk, while banks have been more locally embedded and attentive to the special needs of their borrowers. Capital market development might well result in a depersonalized form of capitalism. The banking industry is becoming more concentrated in nationwide conglomerates rather than in locally oriented institutions, and globalization has brought foreign banks to many nations. For emerging economies these banks represent financial efficiency and economic stability, yet for all nations the sovereignty of their regulatory institutions and their economic development has been challenged.

See also: Development, Economics of; Development: Rural Development Strategies; Development: Socio-economic Aspects; Economic Sociology; Finance, Geography of; Foreign Investment: Direct; Spatial Pattern, Analysis of

Bibliography

- Barth J R, Jahera J 1994 Barriers to the integration of banking markets: The case of the United States. *CEPR Working Paper 440*. Center for Economic Policy Research, Stanford, CA
- Barth J R, Nolle D E, Root H L, Yago G 2000 *Choosing the Right Financial System for Growth*. Milken Institute, Santa Monica, CA
- Biggart N W 2001 Banking on each other: The situational logic of rotating savings and credit associations. In: Wagner J A III (ed.) *Advances in Qualitative Organizational Research*, Vol. 3, pp. 129–53
- Cardoso F H, Faletto E 1979 *Dependency and Development in Latin America*. University of California Press, Berkeley, CA
- Carstens C M 1995 *Las Finanzas Populares en México: El Redescubrimiento de un Sistema Financiero [Popular finance in Mexico: The Rediscovery of a Financial System]*. Centro de Estudios Monetarios Latinoamericanos, Editorial Milenio, Instituto Tecnológico Autónomo de México, Mexico City
- Davis G, Mizruchi M 1999 The money center cannot hold: Commercial banks in the US system of corporate governance. *Administrative Science Quarterly* 44: 215–39

- Demircug-Kunt A, Levine R 1999 Bank-based and market-based financial systems: Cross-country comparisons. *World Bank Policy Research Working Paper 2143*. World Bank, Washington, DC
- Economist 1997 Banking in emerging markets. *The Economist*, April 12
- Evans P 1979 *Dependent Development: The Alliance of Multi-national, State and Local Capital in Brazil*. Princeton University Press, Princeton, NJ
- Evans P 1995 *Embedded Autonomy: States and Industrial Transformation*. Princeton University Press, Princeton, NJ
- Fennema M 1982 *International Networks of Banks and Industry*. Nijhoff, The Hague, The Netherlands
- Garman C, Silva Leite C, Silva Marques M 1998 *Impactos das Relacoes Banco Central x Bancos Estaduais no Arranjo Federativo Pos-1994 [The Impacts of Central Bank-State Bank Relations within Brazil's Federalist Arrangements after 1994]*. Paper presented at the XXII Encontro Anual da ANPOCS (Associacao Nacional de Pos-Graduacao e Pesquisa em Ciencias Sociais), October
- Gereffi G, Korzeniewicz M 1993 *Commodity Chains and Global Capitalism*. Greenwood, Westport, CT
- Gerlach M 1992 *Alliance Capitalism: The Social Organization of Japanese Business*. University of California Press, Berkeley, CA
- Gerschenkron A 1962 *Economic Backwardness in Historical Perspective*. Harvard University Press, Cambridge, MA
- Goldsmith R 1969 *Financial Structure and Development*. Yale University Press, New Haven, CT
- Granovetter M 1994 Business groups. In: Smelser N, Swedberg R (eds.) *The Handbook of Economic Sociology*. Russell Sage Foundation and Princeton University Press, Princeton, NJ
- Granovetter M 2001 A theoretical agenda for economic sociology. In: Guillén M, Randall Collins R, England P, Meyer M (eds.) *Economic Sociology at the Millenium*. Russell Sage Foundation, New York
- Guillén M, Tschoegl A 2000 At last the internationalization of retail banking? The case of Spanish banks in Latin America. Working paper, Wharton School, University of Pennsylvania, PA
- Gurley J, Shaw E 1967 Financial structure and economic development. *Economic Development and Cultural Change* 15(3): 257–68
- Haber S 1997 Financial markets and industrial development: A comparative study of governmental regulation, financial innovation, and industrial structure in Brazil and Mexico, 1840–1930. In: Haber S (ed.) *How Latin America Fell Behind: Essays on the Economic Histories of Brazil and Mexico, 1800–1914*. Stanford University Press, Stanford, CA
- Hadley E 1970 *Antitrust in Japan*. Princeton University Press, Princeton, NJ
- Hall J R 1991 The patrimonial dynamic in colonial Brazil. In: Graham R (ed.) *Brazil and the World System*. University of Texas Press, Austin, TX
- Harvey D 1982 *The Limits of Capital*. Blackwell, Oxford, UK
- Hilferding R 1981 *Finance Capital: A Study of the Latest Phase of Capitalist Development*. Routledge & Kegan Paul, Boston
- Kaplan D A 1999 *The Silicon Boys and Their Valley of Dreams*. William Morrow, New York
- Kenney M, Florida R 2000 Venture capital in Silicon Valley: Fueling new firm formation. In: Kenney M (ed.) *Anatomy of Silicon Valley: Understanding an Entrepreneurial Region*. Stanford University Press, Stanford, CA
- Lamoreaux N 1994 *Insider Lending: Banks, Personal Connections, and Economic Development in Industrial New England*. Cambridge University Press, New York
- Leff N 1978 Industrial organization and entrepreneurship in developing countries: The economic groups. *Economic Development and Cultural Change* 26(4): 661–75
- Levine R 1997 Financial development and economic growth: Views and agenda. *Journal of Economic Literature* 35: 688–726
- Makler H 1999 Regional integration and trends in financial services. In: Haar J, Dandapani K (eds.) *Banking in North America: NAFTA and Beyond*. Pergamon, Oxford, UK
- Makler H 2000 Bank transformation and privatization in Brazil: Financial federalism and some lessons about privatization. *Quarterly Review of Economics and Finance* 40(1): 45–69
- Mintz B, Schwartz M 1985 *The Power Structure of American Business*. University of Chicago Press, Chicago
- Mizuchi M, Stearns L 1994 Money, banking, and financial markets. In: Smelser N, Swedberg R (eds.) *The Handbook of Economic Sociology*. Russell Sage Foundation and Princeton University Press, Princeton, NJ
- Ness W L 1974 Financial markets innovation as a development strategy: Initial results from the Brazilian experience. *Economic Development and Cultural Change* 22(3): 452–73
- Ness W L 2000 Foreign banks in Brazil: What do they do differently? Unpublished essay, Business School, Pontifical Catholic University of Rio de Janeiro, October
- Orri M, Biggart N W, Hamilton G 1991 Organizational isomorphism in East Asia. In: Powell W, DiMaggio P (eds.) *The New Institutionalism in Organizational Analysis*. University of Chicago Press, Chicago
- Patrick H 1966 Financial development and economic growth in underdeveloped countries. *Economic Development and Cultural Change* 14(2): 174–89
- Roe M 1994 *Strong Managers, Weak Owners*. Princeton University Press, Princeton, NJ
- Schumpeter J 1961 *Theory of Economic Development*. Oxford University Press, New York
- Uzzi B 1999 Social relations and networks in the making of financial capital. *American Sociological Review* 64(4): 481–505
- Weber M 1968 Patriarchalism and patrimonialism. In: Roth G, Wittich C (eds.) *Economy and Society*. Bedminster Press, New York, Vol. 3, pp. 1031–77
- World Bank 1989 *Financial Systems and Development. World Development Report*. Oxford University Press, New York
- World Bank 2000 *Entering the 21st Century. World Development Report 1999/2000*. Oxford University Press, New York
- Zeitlin M 1984 *The Civil Wars in Chile or the Bourgeois Revolutions that Never Were*. Princeton University Press, Princeton, NJ
- Zysman J 1983 *Government, Markets and Growth: Financial Systems and the Politics of Industrial Change*. Cornell University Press, Ithaca, NY

H. M. Makler

Fine Arts

Whereas the arts and other forms of creation may be the results of qualities inherent to the human species, the fine arts constitute a particular subset of this broad domain. For the most part a specifically European phenomenon, a social construction embedded in his-

torical, socioeconomic, and political developments in the context of the late Middle Ages, they reached their apogee with the rise of the nation state. Sustained by the development of a discourse that singled out certain art forms as unique and superior to other forms of creation, their high standing was achieved through association with prestigious supporters, penetrating esteemed milieus, and the establishment of selective procedures that became institutionalized over several centuries. Having become a well-entrenched cultural structure, even with the decline of the conditions and institutions that had sustained them, they continue to exert influence over other genres, some of which came into existence only much later.

1. Introduction

Neither timeless nor abstract, like all categories of thought, the fine arts came into being over time through a process of construction and subsequent reconstructions (Zolberg 1990). The term as used today cannot be assumed to possess the same connotations as when it first appeared. Moreover, although functional equivalents of the fine arts have been found in several geopolitical regions and at different historical moments (Alsop 1982), the fine arts represent a complex of meanings which were until relatively recently associated with European civilization. From a social scientific standpoint the fine arts need to be understood in relation to the arts and cultural creation more generally. The first part of this article deals, therefore, with the standing of the fine arts as a subset of aesthetic creation in relation to other forms.

Three developments supported the emergence and subsequent durability of the cultural structure of the fine arts:

(a) The constitution of its intellectual roots as an integral part of European discourse within the changing socioeconomic and political context of the Middle Ages and the Renaissance.

(b) Somewhat independently, professionalization of artistic creation in which individual achievements came to be favored over the communalism of the previously pre-eminent institution of creation, the guild system.

(c) Intersecting these trends, in the place of the previously dominant forms of patronage by the church, the nobility, and fairly autonomous cities, the emergence of new support structures in whose constitution creative artists played an active part: first, official patronage through academies; next, the dealer–critic system.

In the final section, this article analyzes processes of cultural change that, moving beyond anti-academicism, itself a part of the modernist movement, takes the far more radical stance of postmodernism.

Whereas modernism accepted the position of fine art as a hierarchically superior domain of creation, postmodernism contests the very idea of such a hierarchy. It disregards barriers between the fine arts and commercial art forms, professional artists and self-taught creators, art world members and outsiders to it. It rejects the separation of art from politics, science, nature, and even life itself. In trying to clarify the conditions under which certain forms, old and new, gain legitimacy as fine art, this article explores the possible consequences that expanding this construct to encompass existing fine art worlds, cultural politics, and commercial entertainments may have for the idea of fine art in the future.

2. The Making of the Fine Arts Core

2.1 Structuring Aesthetic Culture

Distinguished by a set of artistic categories that made up what has been authoritatively designated ‘the modern system of the arts,’ the fine arts owe their uniqueness to a particular cultural development. It encompasses the transformation of a body of ideas, largely derived from classical thought, thence reformulated in late medieval scholasticism and filtered through Renaissance conceptions. Culminating in the intellectual innovations of the Enlightenment, the fine arts became the established aesthetic structure with which all later arts and scholarship had to contend. Fully constituted by the eighteenth century, as Paul Kristeller noted in his classic formulation, it consisted of a core of five fine arts—painting, sculpture, architecture, music, along with the *belles lettres*—poetry and eloquence (Kristeller 1951, p. 497). While other forms migrated into and out of this cultural structure—dance and garden design, for example—these five became the core of what was meant by Art.

Kant, the first major philosopher to integrate aesthetics into his intellectual system, posited the enduring qualities of art, its universalism, the pleasure it gave, yet insisted that art should not provide too much ease of enjoyment. In this way Kant’s analysis attributed to fine art a seriousness equal to his theory of truth (epistemology or metaphysics) and his theory of goodness (ethics) (Kristeller 1952, p. 42). As later critics have noted, an important criterion in Kant’s formulation was to be art’s gratuitousness, its freedom from social necessity and usefulness. A striking consequence of this construction is that it can be seen as a bulwark for ideologies justifying social inequality (Bourdieu 1984, pp. 494–5).

Although it served as a foundation for philosophical writing throughout the nineteenth century, the system of the arts has been challenged by, among other things, ideas associated with the Romantic movement and by

new technologies. As Kristeller noted, aesthetic analysis based on the modern system of the arts is hard pressed to be appropriate to new mediums and art forms such as the modern concert grand piano and motion pictures (Kristeller 1952, p. 46). Thus, knowing the intellectual roots of the concept is insufficient for understanding the social significance of the fine arts. Without reducing one to the other, the orientations of intellectual and art history with their focus on autonomous genres and their structure, and influences emanating from within their particular domain need to be situated within the social, political, and economic changes with which they intersect. With aesthetic analysis and history of art as the intellectual foundation, the social history of the arts has provided an important perspective by emphasizing the changing social context within which the arts developed and with which they interacted (Hauser 1951)

2.2 Patronage of the Fine Arts: A Contribution of Social History

In their studies of artistic transformations, social historians of the arts stress the effects of changing patterns of patronage: shifts in support structures from the patronage; of the classical city states, imperial courts, churches, secular cities, and states to individual aristocrats, royal courts, or magnates (Hauser 1951, Castelnovo and Guinzburg 1981, Kempers 1992). They have observed that the social standing of art forms and artists is frequently commensurate with the status of their patronage source. Most social historians contend that patrons—those who ‘pay the piper’—call the tune. Further careful examination suggests that the nature of their influence is highly complex. The idea that art is a mere ‘reflection’ of its support system has been shown to be oversimplified (Balfe 1993, p. 4). Instead, it is important to recognize the diversity of intermediate institutions of culture and the expansion of varieties of possible sources of patronage, especially since the middle of the nineteenth century. While social historians provide insight into the impact of major social structural changes on the arts, much social history neglects the agency of artists, individually or in groups, to act on their own behalf. Recent studies indicate that artists played an active role, for example, in establishing the Royal Academy of Fine Arts in France (Heinich 1993) and, in the process, constituting what for over a century was considered ‘the canon.’

For painters who sought to gain the support of royalty and the nobility, the uncertainty of their social position appears to have been the main incentive underlying their strategies. With the establishment of the Academy in the seventeenth century (Heinich 1993), recognition by the highest status groups in society went a long way to providing that standing, at least for a lucky few. In conjunction with subsequent

developments, the fine arts (*beaux-arts*) came to be treated as inhabiting a separate sphere with its own intellectual legitimacy and, under certain conditions, a degree of autonomy from mundane social constraints. This new autonomy is implied in Max Weber’s observation that ‘art becomes a cosmos of more and more consciously grasped independent values which exist in their own right. Art takes over the function of a this-worldly salvation.’ Although Weber goes on to caution that ‘the refusal of modern men to assume responsibility for moral judgments tends to transform judgments of moral intent into judgments of taste (‘in poor taste’ instead of ‘reprehensible’)’ (Gerth and Mills 1946, pp. 941–2), he correctly notes that many Western artists since the late Middle Ages aspired to that autonomy. In that regard, the most durable supporting institution became the academic system (Pevsner 1940).

2.3 The French Academy: Institutional Support for Official Creativity

As an administrative structure encompassing educational institutions, competition, and rewards for talent, the French Academy became the principal gatekeeper in commissioning art for official public purposes. Following a rationale derived from the prestigious humanistic classical scholarship, with the purpose of symbolizing the glory of the monarch and (not very differentiated from him) the State, academicians selected aspirants, rigorously educated them in a certain kind of creativity, established and maintained standards of quality, and legitimated art forms. With the ruler’s accord, this entailed institutionalizing certain art forms and genres based on subject matter, materials, and style, elevating some hierarchically over others. In the process they established the ‘great culture’ that redefined Art (Corvisier 1978, Heinich 1993). The French academic system was widely admired, envied, and imitated in many nations. With variations in accord with local conditions, traditions, and diverse contingent factors (Pelles 1963), it became the successful rival to the previous leading patronage structure, the Church, and overwhelmed the provincial nobility.

Academic dominance of art careers had both symbolic and material consequences. Sustained by an aura of the sacred derived from the King himself, enjoying the acclaim of scholars and amateurs of high standing, fine art was at times treated as a quasi-religion of high culture. But the very fact of attaining high standing contributed to the ambiguity of the position of fine art. Intricately connected to leading institutions of religion, politics, education, and finance, its sacrality was constantly threatened by contamination from the worldliness associated with those domains. The penalty for artists who too obviously were concerned with financial success rather

than high-minded disinterestedness might be condemnation to separate development as part of a provincial 'little culture' to a weakened, marginalized position as a client of the Church (Corvisier 1978), or—worse yet—to the mercies of market forces.

Artists whose works were refused in official *Salons* where they could be viewed by high-status private collectors or clients and public officials with authority to commission and purchase art for public buildings, cathedrals, or churches with elite parishioners, had little recourse but to sell their paintings from their own studios or, worse, at public markets and open air fairs, to an indiscriminate clientele of lesser standing and with considerably lower financial rewards.

At the height of their pre-eminence as gatekeepers, academies and the choices they made were confronted by challenges emanating from artists and other art world participants, including dealers, critics, scholars, and collectors, who were outsiders to the official art world. In the context of eighteenth and nineteenth century shifts in political power away from an aristocracy that had been allied with a narrow stratum of wealthy financiers, an expanded and heterogeneous middle class came to prominence. It provided a diverse market that was no longer coterminous with the taste promoted by the official milieu. Instead, its members seemed to share an elective affinity with the entrepreneurship of the new dealer and critic system. This combination created conditions favorable to aesthetic innovations that were to challenge the officially recognized fine arts even more profoundly (White and White 1965, Boime 1971, Zolberg 1983).

2.4 From the Academy to Anti-academic Movements

Whereas under the Old Régime, fine art was what the monarch preferred, through much of the nineteenth century it was largely economic elites whose tastes, whether conventional or audacious, came to dominate (Boime 1971). Its demise as a result of the Great Revolution had left a void in the aesthetics that it had imposed on French cultural life. Remade after its virtual abolition and modified several times, the academic system flourished for the better part of the nineteenth century. Its eventual re-establishment as a set of state institutions was welcomed by many aspiring artists, who found that career avenues reopened to them (Boime 1971). But its success was soon diluted, overwhelmed by increasing numbers of aspirants (White and White 1965).

Through administrative control of admissions, the assignment by their juries to better or worse locations for art works on its walls, the prestigious biennial *Salons* might become opportunities for artists to be rewarded with prizes, government or private commissions, and future sales or, on the contrary, be quite unnoticed. Alienated from the traditional or official

artistic values of their time, convinced of its favoritism toward the most timid conformity, many artists became disenchanted with the academic institution that earlier artists had helped to launch. Highly innovative at its beginnings, the official academy, as is the case of many formal organizations, was accused of becoming increasingly rigid and oligarchic for excluding innovative ideas and creators. By the middle of the nineteenth century, academic administrators were being charged with behaving like crude bureaucrats, unreflectively enforcing old ideas, favoring the predictable and the safe over the new and audacious.

These accusations were sometimes exaggerated by critics and disgruntled aspirants. In fact, its officials were not resistant to change across the board. The hierarchical dominance of history painting, portraits of monarchs, nobles, and very wealthy officials and businessmen with which the century had begun was being overshadowed by the end of the nineteenth century with the elevation of what had previously been classed as lesser genres. They included landscape paintings and nominally allegorical or mythical themes, often in a style accentuating narrative as opposed to timelessness, sweetness rather than nobility (Bouguereau) (Boime 1971). But the ethos of tamedness imposed on art students and members conflicted with the increasingly prevalent conception of the artist as Romantic genius with a singular vision (Bourdieu 1995). The academy became synonymous with the overbearing institutions that trammelled their imaginations.

In trying to account for the eventual demise of academic authority, Harrison and Cynthia White have argued that the very success of the institution made it the goal of too many aspiring artists. Under the onslaught of swelling numbers of applicants, the academy was obliged to introduce increasingly selective mechanisms (White and White 1965). From an ethnographic perspective, the analytic categories developed by Howard Becker provide further insight into the social roles of artists in complex institutional settings or 'Art Worlds.' For Becker, the world of art is made up of several different types of artists. Certain artists are 'certified'—as 'integrated professionals,' or they may be nonconformists—'mavericks' (Becker 1982). In the exceedingly centralized French system especially, mavericks were viewed and often presented themselves as outsiders and innovators who knew the rules and *deliberately* violated them.

As Becker suggests, most were trained in the academic institutional system that rejected them, or studied with academically trained artists. But many were willing to risk exclusion from its sales institution, the official *Salon*, in order to be true to their ideas. Although, as later research indicates, in fact, relatively few of them were very innovative or audacious. This was even the case of the most celebrated event of the Second Empire, the *Salon des Refusés*, when Impressionists' works were displayed, as well as various

later dealer-organized exhibitions. But even the more conventional ones formed alliances and exhibited with independent groups of real mavericks (Huston 1989).

One of the most important results of their refusal to conform is the seemingly unending succession of new style after style. Associated with groups of artists who provided moral support for their efforts, each new stylistic form became a contender for official and public recognition. Its leaders were: Courbet and Naturalism; Corot and Millet and other Barbizon painters who led the movement for revalorizing landscape painting; Manet and numerous painters influenced by him who made up the Impressionist group, rejecting the foundation of classical drawing they had acquired in favor of color experiments that provided a direct relationship to nature; Neoimpressionists such as Signac and Seurat who undertook quasiscientific experimentation with the optics of color in their *pointillisme*; Van Gogh and his deliberate distortions of Expressionism; Gauguin and his combination of gothic line and form with eroticized exoticism; Cézanne and his uniquely difficult-to-characterize works (Rewald 1988). Most of these 'independent' artists found themselves barred from a place in official exhibitions for most of the nineteenth century.

For artists of modest means, the dealer-critic system that expanded and flowered in the nineteenth century was essential for permitting them to survive honorably as painters. As fine artists, they relished the informality of contractual arrangements that made agreements with dealers rest on a gentlemanly handshake (Vollard 1978, Moulin 1987). Even artists who were independently wealthy, such as Manet and Degas, sought to build support among new middle class collectors through the growing dealer system. Unexpected as it may have been to their critics, eventually a number of their aesthetic innovations came to be more highly prized than the more conventional art of their contemporaries. As controversial and willing to risk failure as they were, however, for all their innovations, they continued to work in the basic mediums and genres of art, rather than create entirely new forms of art.

3. How the Core Collapsed—On Artistic Change

3.1 The Tradition of the New

From approximately the middle of the nineteenth through nearly half of the twentieth century it became customary to speak of artistic change as a revolutionary succession of art movements or *avant-gardes* (Poggioli 1971). Most of them challenged stylistic strictures governing artistic form, some challenged the social structures of art worlds, and certain of them used their art to attack political and social arrangements beyond art worlds. Through these independent

artists of *avant-garde* or secession movements that challenged the academy's legitimacy, the academic fine arts fell into disrepute. Several of the new styles gained ascendancy, but even as anti-academic artists, they did not eradicate the idea of fine art itself, nor their aspiration to aesthetic autonomy.

The story of the apparently inexorable march of new styles encompassed elements that contributed to the 'tradition of the new,' as Harold Rosenberg termed it (1965). Its pattern of innovation, rejection, and acceptance has become familiar and conventional in the fine arts, and is acknowledged as part and parcel of the flow of modernism. In the post-World War II period, the United States produced what was then considered the most innovative artistic movement. Referred to variously as Abstract Expressionism, Action Painting, or simply the New York School, its members aspired to replace the figurative content and narrative of art, largely hegemonic at the time, by abstract representations of the artist's feelings and impulses. Viewed by many as incomprehensible or self-indulgent, as it gained acceptance among a few collectors and a place in adventurous museums, in line with the century-long pattern of the challenge by the new, it too was faced by rivals: even without an institution such as the academic system, *avant-gardism* itself had become institutionalized. Soon, they too found themselves and the aesthetics they had constructed challenged (Crane 1987).

As unexpected as they were, the new genres and styles—Pop Art, Minimalism, Pattern, Photo-Realism—could claim as much legitimacy as had previous newcomers. There was little basis for holding them off except through assertions of bad faith on the part of artists, decline of standards, lack of discrimination. On the one hand, these resemble the reproaches that have been invoked for a century and a half. On the other, the Pop Artists had crossed the barrier separating naked commercialism from disinterested art. Their popularity assailed what might be called the dominant canon of abstractionist modernity, leading the way to what has been termed the postmodern condition.

3.2 Postmodernism and Challenge from the Margins

Despite their confrontational stance, few artists of anti-academic movements actually repudiated 'the modern system of the fine arts' until well into the twentieth century (Danto 1986). Instead, the idea of 'fine art' has until very recently retained its lofty status. What has changed is that it may be freely attached to new genres, many of them nonexistent at the height of academic dominance, and most unrecognizable to nineteenth century observers. Whereas modernism accepted the position of fine art as a hierarchically superior domain of creation, postmodernism contests the very idea of such a hierarchy. It disregards barriers between the fine arts and commercial art forms,

professional artists and self-taught creators, art world members and outsiders to it. It rejects the separation of art from politics, science, nature, entertainment, and even life itself.

Maverick professionals have attempted to overthrow the existing canon that has been the foundation of their training (Becker 1982). Excluded groups on the fringes of established art worlds are a constantly renewed source of aesthetic change, whether through challenging aesthetic ideas or simply to gain entry without changing prevailing aesthetic forms or hierarchies. Most prominent are women, racial or ethnic minorities, or disadvantaged people more generally, who attempt, either on their own behalf or through the mediation of influential gatekeepers, to penetrate to the center (Lang and Lang 1990).

But over the years the margins of the art world have come to defy easy categorization. As has become startlingly evident, the barriers between high and low art, art and politics, art and religious rite, art and emotional expression have been repeatedly blurred and breached. One outcome of these assaults on conventions, in which the periphery has come to be valorized beyond all predictions, is that it is virtually impossible to find an aesthetic center. These developments have had their critics, many of them simply conservative either politically or aesthetically, or sometimes both.

How certain artists and art forms gained ascendancy has often involved political and social contexts that are intertwined with aesthetics. They may involve the construction of institutional support favoring particular stylistic forms while excluding others. Under authoritarian political regimes in modern times this has led to fierce repression of artists and their works (Nazi Germany, Stalinist Soviet Union, Maoist China, among others). Even in liberal regimes, artists may be restricted in the free expression of their creation. Whereas in the past the most important sources of support came from absolutist monarchs who were attempting to centralize their nation states at the expense of competing institutions and what they constructed as peripheral cultures, today the ground is shifting. This is indicated by growing interest in the construction of social identities that are not as co-extensive with a nation's boundaries as was the ideal expectation of the past. The survival or revival of cultural conflict around language, ethnicity, and religion provide a fertile ground for jockeying for pre-eminence by winners at the expense of dominated groups.

A familiar cliché has become that of imagining the artist facing a philistine world that misunderstands him, struggling for inspiration, and gaining recognition only belatedly. Applied to nineteenth-century composers such as Beethoven and Schubert; painters such as Gauguin and Van Gogh, in the twentieth century this model was moved to the lives of Jazz 'greats'—Scott Joplin, Billy Holliday, Thelonius

Monk, and many others. The insertion of Jazz and other popular music into the domain of the fine arts is far removed from academies and official legitimacy. But by now almost the entire body of popular commercialized culture and the constitution of stardom could be included in this analysis of forms that are outside of fine art canons.

Supposedly born as low-life brothel music, Jazz grew in the late 1910s and 1920s into a wickedly exciting, commercially powerful form of popular music. Taken up as an art form by European intellectuals, who incorporated it into ideas of an essentialist primitivism that were popular among artists and writers, and appealed to many amateurs, their attention provided the cachet that ordinary commercial musical forms lacked. Eventually, certain American intellectuals were drawn into defining Jazz as an expression of revolt against the conventional lives they were expected to live. By the 1940s, with the advent of be-bop and the regular forays of composers such as Duke Ellington into concert halls, Jazz began to take on the trappings and pretensions of Art, moves supported by patrons, official government bodies, foundations, while commercial entrepreneurs became less important. Its legitimacy functioned at two levels, one aesthetic and the other political. Rather than being content with the entry of the new musical forms into common usage through appropriation by industry agents, even at increased financial returns, African Americans became active in asserting the importance of their identity through its innovations.

The rise of Jazz to respectability and fine art standing is based, by now, on a long history reconstructed (or constructed). It includes a cataloged repertory and a body of admirers and practitioners. It has a closely perceived linkage with black cultural identity and hence has the support of related interest groups who advocate funding for minority cultural and political projects. Jazz (or 'black classical music' as some call it) is firmly embedded in the curricula of institutions of higher education and conservatories. It is deemed worthy of support by governmental agencies such as the National Endowment for the Arts and State Arts Councils. By now, Jazz has attained a level of respectability that would have been difficult to predict less than a century ago.

The very nature of art in the late twentieth century entails a constant process of absorbing, sampling, appropriating, amalgamating new and unusual forms, and incorporating a changing body of personnel. Currently, the field is much larger and the number of contenders has multiplied, as has the core itself. The line between classical and commercial entertainment has been crossed repeatedly. Beyond professionals and mavericks, creative work by folk and naïve artists has penetrated the museums and galleries that had been host to creators formed in the attachment to autonomous fine art of whatever form. No longer a single canon, the centralized academy and its in-

stitutional descendants have given way to competing groups who promote a broad range of genres or forms (Huyssen 1986, p. 218). It may be that, as Andreas Huyssen has suggested, the previous domination of the fringes by the West may be replaced by a healthy resistance of the dominated in the form of a productive tension between the political and the aesthetic (Huyssen 1986, p. 221).

See also: Avant-garde Art and Artists; Collective Identity and Expressive Forms; Culture, Production of; Nationalism and Expressive Forms; Popular Culture

Bibliography

- Alsop J 1982 *The Rare Art Traditions: The History of Collecting and Its Linked Phenomena Wherever They Have Appeared*. Harper and Row, New York
- Balfe J H (ed.) 1993 *Paying the Piper: Causes and Consequences of Art Patronage*. University of Illinois Press, Urbana, IL and Chicago
- Becker H S 1982 *Art Worlds*. University of California Press, Berkeley, CA
- Boime A 1971 *The Academy and French Painting in the 19th Century*. Phaidon Press, London
- Bourdieu P 1984 *Distinction: A Social Critique of the Judgement of Taste*. Trans. by Richard Nice. Harvard University Press, Cambridge, MA
- Bourdieu P 1995 *The Rules of Art: Genesis and Structure of the Literary Field* [trans. Emanuel S]. Stanford University Press, Stanford, CA [Orig. Éditions du Seuil 1992]
- Castelnuovo E, Guinzburg C 1981 Domination symbolique et géographie artistique dans l'histoire de l'art italien. *Actes de la Recherche en Sciences Sociales* 40(November): 51–72
- Corvisier A 1978 *Arts et sociétés dans l'Europe du XVIII^e Siècle*. Presses Universitaires de France, Paris
- Crane D 1987 *The Transformation of the Avant-Garde: The New York Art World 1940–1985*. University of Chicago Press, Chicago
- Danto A 1986 *The Philosophical Disenfranchisement of Art*. Columbia University Press, New York
- Gerth H, Mills C W 1946 *From Max Weber*. Oxford University Press, New York
- Graña C 1964 *Bohemian versus Bourgeois: French Society and the French Man of Letters in the 19th Century*. Basic Books, New York
- Hauser A 1951 *The Social History of Art*. Knopf, New York, 4 Vols.
- Heinich N 1993 *Du peintre à l'artiste: Artisans et Académiciens à l'âge classique*. Éditions de Minuit, Paris
- Huston L 1989 *L'Évolution des expositions d'art à Paris, et le pouvoir dans la vie artistique (1864–1914)*. Ph.D. thesis. Concordia University, Montréal, Canada
- Huyssen A 1986 *After the Great Divide: Modernism, Mass Culture, Postmodernism*. Indiana University Press, Bloomington, IN
- Kempers B 1992 *Painting, Power, and Patronage: The Rise of the Professional Artist in the Italian Renaissance* [Trans. from Dutch Jackson B]. Allen Lane/Penguin Press, London
- Kristeller P O 1951 The modern system of the arts. *Journal of the History of Ideas* 12: 496–527
- Kristeller P O 1952 The modern system of the arts II. *Journal of the History of Ideas* 13: 17–46

- Lang G E, Lang K 1990 *Etched in Memory: The Building and Survival of Artistic Reputation*. University of North Carolina Press, Chapel Hill, NC
- Moulin R 1987 *The French Art Market, A Sociological View* [Trans. Goldhammer A]. Rutgers University Press, New Brunswick, NJ [abridged from orig. Éditions de Minuit, 1967]
- Pelles G 1963 *Art, Artists and Society: Origins of a Modern Dilemma, Painting in England and France, 1750–1850*. Prentice-Hall, Englewood Cliffs, NJ
- Pevsner N 1940 *Academies of Art: Past and Present*. Cambridge University Press, Cambridge, UK
- Poggioli R 1971 *The Theory of the Avant-Garde*. Harvard University Press, Cambridge, MA
- Rewald J 1988 *Cézanne and America: Dealers, Collectors, Artists and Critics, 1891–1921*. Princeton University Press, Princeton, NJ
- Rosenberg H 1965 *The Tradition of the New*. McGraw-Hill, New York
- Vollard A 1978 *Recollections of a Picture Dealer* [Trans. MacDonald V M]. Hacker Art Books, New York [Enlarged from orig. Albin Michel, 1937]
- White H C, White C 1965 *Canvases and Careers, Institutional Change in the French Painting World*. Wiley, New York
- Zolberg V L 1983 Changing patterns of patronage in the arts. In: Kamerman J B, Martorella R (eds.) *Performers and Performances: The Sociology of Artistic Work*. Praeger, New York
- Zolberg V L 1990 *Constructing a Sociology of the Arts*. Cambridge University Press, Cambridge, UK

V. L. Zolberg

Fire, Human Use, and Consequences

1. Introduction

The ability to handle fire is, like language and the use of tools, a universal human attainment: no human society in historical times is known to have lived without it. The ability to handle fire is also exclusively human; while other animals have developed rudimentary signals and tools, only humans have learned to handle fire.

Today fire is continuously present in all human societies. It is used in many different guises, some highly visible, others largely hidden from public view and consciousness. Its control has played a far more important part in human history and social evolution than is generally acknowledged in the literature of the social sciences.

2. The Original Domestication of Fire

Fire is, according to dictionary definitions, a process of combustion, manifested by heat and light. It occurs relatively rarely in nature; when it does, it does not last long, with a few exceptions such as the burning coal-fields in Northern China which have been alight for

many centuries. Most natural fires have rapid and drastic effects. They cause a quick disintegration of organic matter, reducing it to ash and smoke in a process of transformation that is almost the prototype of 'destruction.'

During the Paleolithic, human groups have learned to turn this destructive and essentially purposeless natural force into a regularly available asset that could be used for a variety of 'productive' purposes. They 'domesticated' fire, they 'tamed' it and 'incorporated' it into their own societies. They were able to do so thanks to their general capacity for culture—for learning new forms of behavior, sharing those with others, and transmitting them to following generations. Adjusting their behavior in such a way that a fire could be kept burning continuously, they developed a fire regime which enabled them to expand their control over their natural environment, while at the same time affecting their own social arrangements and their personal habits and feelings (cf. Goudsblom 1992, pp. 12–41).

The uses of fire were diverse, ranging from the enjoyment of heat and light to clearing land for hunting and gathering and to cooking. Their overall tendency was to make life in human groups more secure and comfortable, but also more vulnerable and dangerous. *Fire increased at once people's destructive and productive potential.* Its control thus reinforced a basic trend in human history and social evolution: the increasing differentiation in behavior and power between humans and all biologically related species.

3. Fire and Agrarianization

The formation of the fire regime may be regarded as a paradigm for other regimes that were developed at later stages in human history. A paradigm in a double sense: first of all, substantively, the regime by which humans learned to extend their care and control over fire could serve as a model for subsequent forms of care and control over other forces in nonhuman nature such as plants and animals. Second, the domestication of fire may be seen as a model case in a methodological sense, since it exhibits some basic principles which may be observed again and again in human history and social evolution.

Toward the end of the Paleolithic, between 15,000 and 10,000 years ago, some human groups began to extend their care and control over selected plants and animals. This marked the beginning of what is often called 'the First Revolution' but is better viewed as the second great ecological transformation brought about by humans. That transformation could not have taken place without the firmly established control of fire. Once it was under way, it in turn deeply affected the further development of the part played by fire in human society.

The domestication of fire was a precondition for agrarianization first of all because in most areas using fire ('slash and burn') was the most efficient technique of opening up fields for cultivation by removing the existing vegetation. Moreover, most of the crops selected for cultivation needed cooking before consumption as a staple food. As a further, less tangible precondition, the solidly established 'species monopoly' over fire gave humans a dominance over other animals which enabled them not only to bring certain species under direct control but also kept most of the remaining 'wild' animals at a distance from their crops and herds. Finally, the long familiarity with a fire regime probably helped to prepare people for the strains of an agrarian regime demanding self-imposed 'deferred gratification.'

With agriculture and livestock, fire was no longer the only nonhuman source of energy under human control. Gradually it ceased to be the prominent focus of group life that it had been for many thousands of generations, becoming increasingly more dispersed over various sorts of specific 'containers,' such as hearths, ovens, and lamps, beyond which it was not allowed to spread. Its use was subjected to stricter regulation. Whereas during the initial stages of domestication, the main concern had been to keep the one fire burning, the main concern now became to prevent the many fires from causing a conflagration.

As the more advanced agrarian societies grew more populous, the number of fires in use increased proportionally. Those fires were increasingly concentrated within the 'fire-protected zones' of towns and villages, in special containers. The uses of fire became more specialized, with various crafts and professions such as smiths and potters developing their own pyrotechnic skills. With proliferation, concentration, and specialization came new forms of organization, if only to meet the growing need for fuel. Unavoidably, the use of fire also reflected the process of stratification, as some privileged people came to command huge supplies of fuel, and could use fire for impressive displays of power, whereas many others might never call any fire their own.

A common problem in all advanced agrarian societies was the prevention of uncontrolled fire in cities. For the growing urban populations the proliferation of fires increased the risks of blazes, while the accumulation of property meant that people had more to lose.

Of course, it was fire as a natural force with its aimless and self-generating destructiveness that was feared. In case of a military raid people had to reckon with the common practice of 'sack and burn.' But ordinarily they encountered fire only in its domesticated settings where the greatest danger lay in carelessness and where they had to rely on their own and their fellow citizens' incessant caution.

Once a fire did break out in a preindustrial town, there was little one could do. Usually the only way of

stopping it was by creating a 'fire break' and tearing down the buildings towards which the fire was advancing; deliberate destruction by human hands had to put an end to the blind destruction by fire. By far the most effective way of restricting the risk of conflagrations was then, as it still is, prevention.

There is a tendency today to conceive of problems of fire prevention primarily in terms of technology: of building materials and technical equipment. However, the problems have always been at least as much civilizational. The crucial issue was human behavior in its social context of informal and formal controls.

Little is known about the organization of fire brigades in preindustrial cities. The available evidence suggests that their effectiveness, and their very presence or absence, reflected important aspects of the social structure. In ancient Rome, fire brigades at first operated as private bands, and were later brought under imperial military command. A glimpse of the situation in the Roman provinces is shown in a letter from the emperor Trajan, written in 112 CE, in which he refused permission to the governor of Nicomedia to establish a fire brigade of citizens on the grounds that such an organization could easily 'turn into a political club' (Goudsblom 1992, p. 118). One of the social innovations in the towns of medieval Europe was that they had civil fire brigades.

4. Fire in Industrial Society

After the original domestication of fire and the rise of agriculture and animal husbandry, industrialization was the third major ecological transformation set in motion by humans. Again, natural forces originally outside the human realm were brought under human control. These natural forces consisted primarily of large, untapped sources of energy stored up in fossil fuels. Tapping the sources of coal, and later also of oil and gas, enabled people to exploit a great variety of other mineral resources as well, from iron to plutonium.

From the start, industrialization involved an enormous increase in the number, variety, and intensity of man-made fires. Just as the first stages of agrarianization were marked by extensive and conspicuous use of fire in the clearing of land by slash and burn, early industrialization dotted the landscape with smoke stacks and steam locomotives. In agriculture, the extensive use of fire gradually made way to more intensive forms of tillage with less conspicuous use of fire. Similarly, fire and smoke have become far less prominent in the industrial scene, as production and transport were powered increasingly by electricity and other forms of technologically processed energy (cf. Hughes 1983).

Yet modern industry as well as agriculture continue to be highly fuel-intensive. Most of the energy used is still derived from fossil fuels. Combustion processes

continue to play a central role, but they are relegated to special containers so that most people are not confronted with any of the features that can make the presence of a fire annoying and dangerous. The caprices of the flames are fully controlled. Soot, smoke, and risks of burning are reduced to a minimum. The furnaces and combustion chambers in which great heat is concentrated remain cool on the outside.

Typical products of modern fuel-intensive industry are cars, with engines designed to be propelled by finely tuned and minutely controllable combustion processes. Cars are made of steel, plastic, and glass—materials produced and processed at high temperatures. Yet when people get into their vehicle and turn on the electrical ignition to start the engine they are unlikely to be consciously aware of using fire and the products of fire. Nor are they used to stopping and thinking of the unintended contribution they are making to CO₂ emissions and, probably, to global warming (cf. Fleming 1998).

The highly fuel-intensive character of modern technology is matched by the advanced standards of fire prevention. Every branch of industrial production, from the building trade to the manufacture of clothes and furniture, is subject to strict measures with respect to fire. In such branches as the aircraft or the hotel industry, regulation has become international.

Along with productivity the social potential to destroy, or 'destructivity,' has increased greatly. Not only have the means of preventing and fighting fires become more effective, but so have the means of starting fires, while cities as targets for arson have also increased in number and size. Consequently, the hostilities of World War II have caused some of the largest urban fires in history, as in Hamburg, Dresden, Tokyo, Hiroshima, and Nagasaki. Forest fires caused directly or indirectly by human action have also been more numerous and larger than in any previous century (cf. Pyne 1995).

Forms of fire use that were common and relatively inconsequential before the industrial age have become far more consequential under today's socioecological regimes. Thus, in many parts of Asia and Africa the need for firewood is now significantly contributing to deforestation. The high concentrations of wood stoves and ovens in the crowded metropolises in China and elsewhere have detrimental effects on human health.

5. Symbolic and Religious Aspects

From all over the world myths have been collected relating how humanity first came to master fire, often in a feat of cunning and deceit (Frazer 1930). Those myths convey a sense of the preciousness of fire which is represented as a unique possession, separating humanity from the animal world. Many people told that neighboring groups living further away in the

bush did not have fire and were unable to cook their food; those stories have all been proven slanderous.

Fire is generally believed to have purifying effects. The domain of culture is seen as coinciding with the range of domesticated fire. Thus, cooked meat is regarded as edible for humans, raw meat as only fit for beasts. Similarly, land cleared by burning is trusted as belonging to the human domain, whereas uncleared bush and forest are feared as a dangerous hideout for predators, serpents, insects, and evil spirits.

The sense of preciousness has also found expression in rites. The communal fire would be the center of group life, and people would treat it as sacred, surrounding it with ceremony and taking care that it would never cease burning. The Roman cults of the hearth and of Vesta, the fire goddess, show how such ancient rites could persist in an urban setting (cf. Pyne 1997).

Yet, with urbanization, the sense of the preciousness of fire was often superseded by a general anxiety of the dangers it posed. Thus while Hindu and Zoroaster fire cults kept alive the age-old worship of fire, they were surrounded by ceremonial care. In Christianity and Islam, some old fire traditions persist, as in the burning of candles and incense. Fire no longer plays any part in beliefs about the origin of human society, however. Instead, it is strongly associated with evil. Sinners and unbelievers are threatened with a life after death in which they will be tormented by eternal burning.

Fire festivals, such as those held all over Europe on Saint John's day, when things considered impure would be committed to the flames, were occasions of combined joy and horror. Modern literature and art abound with evocations of the terrors of fire. Most novelists who have described a blaze attribute the disaster, quite realistically, not to natural forces such as lightning but to human action—either to personal negligence, madness, or malice, or to collective aggression, in war.

A more dispassionate way of regarding fire, not in terms of its good or bad effects upon human beings but as a natural force in its own right, was first developed by the philosophers of ancient Greece. Out of various antecedents a generally accepted doctrine emerged according to which the universe consisted of four elements: earth, water, air, and fire. In China and India similar systems of thought developed. With some variations the view in which fire constituted one of the major elements remained the dominant cosmology for natural scientists and physicians well into the modern era.

Alchemists and chemists, as the direct successors to the metallurgists and 'fire masters' of a previous age, made the study of fire into their central concern. Just as modern mechanical industry in its initial stages owed a great deal to the steam engine, the rise of modern physical science would have been inconceivable without the laborious efforts of earlier investigators experimenting with, and theorizing about, fire.

6. Conclusions

The control over fire has been a human species monopoly since the Paleolithic. It constitutes an integral part of the apparatus with which humans have established and continue to maintain their dominance over other species.

Ever since the earliest stages of its domestication, the same sequence of events has recurred again and again in human relations with fire: deliberately sought advances in control are followed by unintended increases in dependency, and as the dependency sinks deeper into the social and economic infrastructure, it tends to become increasingly less clearly perceptible. Like agrarianization, industrialization began with conspicuously huge applications of fire. As industrial production became more specialized and highly organized, so did the use of fire. In everyday life in highly industrialized contemporary societies, flames are visibly present only in such highly domesticated guises as cigarette lighters, candles for ceremonial use, or wood fires intended to create a sphere of comfort and relaxation. When, on the other hand, a fire is shown on television or in the newspapers, it almost always spells war and disaster. The regularly controlled combustion processes upon which industrial production largely rests are now mostly relegated away from public view.

In accordance with this apparent eclipse of fire, it has also ceased to be regarded in the natural sciences and philosophy as one of the elements out of which the world is composed. The very concept of fire has disappeared from scientific discourse, to be replaced by the more abstract concept of 'energy'—something that cannot be directly seen, heard, smelled, or felt. Similarly, many a day may pass in the life of the average citizen of a modern industrial state during which he or she does not witness any physically observable fire. In the social sciences, too, fire has become almost completely unnoticed. It can be concluded from the foregoing that this is unwarranted.

See also: Agricultural Sciences and Technology; Climate, History of; Environmental and Resource Management; Environmental Challenges in Organizations; Food Production, Origins of; History of Technology; Human–Environment Relationships; Land Use and Cover Change

Bibliography

- Bachelard G 1938 *La psychanalyse du feu*. Librairie Gallimard, Paris [1964 *The Psychoanalysis of Fire*. Beacon Press, Boston]
- Canter D (ed.) 1980 *Fires and Human Behaviour*. Wiley, New York
- Fleming J R 1998 *Historical Perspectives on Climate Change*. Oxford University Press, Oxford, UK
- Frazer J G 1930 *Myths of the Origin of Fire*. Macmillan, London

- Goudsblom J 1992 *Fire and Civilization*. Penguin, London
- Hughes T P 1983 *Networks of Power. Electrification in Western Society, 1880–1930*. Johns Hopkins University Press, Baltimore, MD
- Jones E L, Porter S, Turner M 1984 A Gazetteer of English urban fire disasters. *Historical Geography Research Series 13*. Geo Books, Norwich, UK
- Pyne S J 1982 *Fire in America. A Cultural History of Wildland and Rural Fire*, 1st edn. Princeton University Press, Princeton, NJ
- Pyne S J 1995 *World Fire. The Culture of Fire on Earth*. Henry Holt and Company, New York
- Pyne S J 1997 *Vestal Fire. An Environmental History, Told through Fire, of Europe and Europe's Encounter with the World*. University of Washington Press, Seattle, WA

J. Goudsblom

Firm Behavior

At the heart of the supply side in the theory of how markets function is the theory of the firm. This article surveys the received economic wisdom on how firms make their decisions as to the types and quantities of products they will supply, the prices at which they sell, and the investments they make toward future operations.

That economists have adopted the convention of talking and writing about the behavior of 'the firm' is semantically curious. Supply decisions are made by entities bearing a host of names—corporations, companies, enterprises, businesses, ventures, farms, factories, establishments, and, of course, firms. The word 'firm' is derived from the Latin verb *firmare*, which evolved over time to describe the confirmation of a transaction by signature. The notion of a 'firm' metamorphosed during the nineteenth century to define houses of business organized through written agreements, notably, partnerships, and in modern lay parlance, it is used most frequently to describe partnerships in such fields as law and accounting.

1. Complex Organizational Forms

Meanwhile, the forms of business organization were changing in legal structure and scope. Corporations replaced partnerships, and multi-unit business enterprises operating both within a nation and multi-nationally came to supply the lion's share of services and (especially) products. Already in the sixteenth century, the Venetian merchant Antonio was said by Shakespeare to have 'an argosy bound to Tripolis, another to the Indies ... a third at Mexico, a fourth for England, and other ventures he hath squandered abroad.' In 1995, US business corporations reported

sales of roughly US\$14 trillion, partnerships US\$0.85 trillion, and non-farm individual proprietorships US\$0.81 trillion. In 1982, 7,030 companies with 500 or more employees surveyed by the US Bureau of the Census (1986) generated 41.3 percent of all business employment. On average, those relatively large enterprises operated 64 separate domestic business establishments or plants each, not including foreign branch units. Thus, much of the industrialized world's business is done by organizations of considerable geographic scope and complexity.

Within this array of organizational forms, the theory of the firm is ambiguous, or one might say flexible, in isolating the locus of supply decision-making. The decisions encompassed under the theory may be made at the individual branch or factory level, at the headquarters level, at some intermediate management level, or across a composite of levels.

Few 'firms,' however they may be bounded, carry out internally all the activities needed to provide the goods or services they ultimately supply. Many intermediate inputs, supplies, machines, and specialized production and marketing services are procured through arms-length market transactions from other entities. The characteristic defining the bounds of a 'firm,' Ronald Coase argued in a seminal article (1937), is 'suppression of the price mechanism.' According to Coase, activities are collected within what we call a firm when transaction costs incurred in using the price mechanism for external purchases or sales exceed the cost of organizing those activities through direct managerial controls. Transaction costs may be high when work specifications are complex or rapidly changing, contract negotiations and renegotiations are burdensome, and /or there is a substantial risk of opportunistic deviations by one or another party once contracts have been fixed. The higher the transaction costs, the more a firm is likely to be vertically integrated to carry out internally the activities needed to implement its ultimate supply objectives.

2. The Goals of the Firm

If an organization is to act purposefully, it must have more or less clearly defined goals. In the standard static theory of the firm, the goal is simple: the firm maximizes profits. On the output side, this entails expanding production until marginal cost rises into equality with marginal revenue (which, for a purely competitive firm, is the product's parametric price). On the input side, profits are maximized by hiring the diverse inputs until the value of their marginal product, or in monopoly cases their marginal revenue product, is equal (when inputs are competitively supplied) to the inputs' prices. In addition, the firm must operate on its production function rather than off it; that is, it must avoid the manifold risks of X-

inefficiency (see, for example, Leibenstein 1966). This will be examined further later in this document.

Matters become more complicated when time and uncertainty enter the picture. On the time dimension, two general problems can be distinguished. For one, firms are faced with a continuing menu of opportunities to invest in longlived production facilities and new products. The standard theory says that they should accept all mutually independent investment projects whose internal rate of return exceeds the firm's marginal cost of investible funds. (Interdependence or mutual exclusivity of projects requires the satisfaction of additional conditions.) Additionally, when a firm has monopoly power and hence a choice of alternative output price levels, setting a price that maximizes current profits encourages the entry and expansion of competitors who erode future profit-earning opportunities. The firm must then make a tradeoff in its pricing between present and future profits. In the most elegant control-theoretic formulation of this problem (Gaskins 1971), the firm holds its price below the short-run profit-maximizing level so as to slow the rate of competitors' expansion and hence the erosion of future profits, appropriately discounted to present value, by the optimal amount.

One crucial variable in these calculations is the marginal cost of investible funds, also known as the firm's discount rate. Standard treatises on corporate finance suggest that the firm's discount rate be measured as the weighted average cost, taking into account income tax implications, of the various external sources of funds, i.e., common stock, long-term debt, and short-term loans; and the extent to which external financing, as compared to the reinvestment of current earnings rather than paying dividends, is utilized. Two complications arise here. First, the optimal mix or portfolio of financing methods must be chosen. There is disagreement in the literature over the extent to which external finance costs should and do vary with those portfolio choices. Second, debt costs vary over time with changes in macroeconomic money market conditions; and common stock financing costs fluctuate with both general macroeconomic conditions and with investors' volatile 'animal spirits,' to use Keynes' (1936, p. 161) metaphor. There is disagreement among scholars over whether firms should accept current market signals unequivocally in setting their investment discount rates or whether, recognizing that investments are for the long term, they should try to see beyond current market trends and 'smooth' anticipated external finance market costs in computing discount rates. If they fail to heed short-run equity market trends, they may be subject to hostile takeover by another enterprise more closely attuned to maximizing current shareholder value. Compare Marris (1963), Jensen (1988), and Scherer (1988).

A related question is how a firm's decision makers should treat the intrinsic uncertainty of investment payoffs. In principle, external financial investors eval-

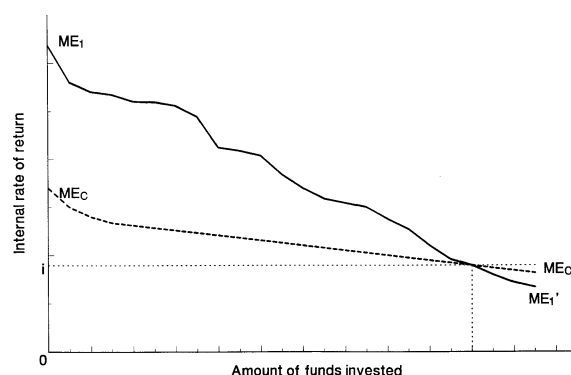


Figure 1
Marginal efficiency of investment schedule for the firm

uate the overall risks of a firm's investments and, by their reactions, introduce appropriate risk premiums into the firm's cost of capital. Using these, internal decision makers should presumably choose investments so as to maximize the discounted expected value of their investments. Often, however, risk aversion creeps into investment evaluations when decision makers or their staff adjust discount rates upward, shade estimated payoffs downward, or avoid through sensitivity analyses projects with substantial downside risk. Again, the firm that internally introduces more risk aversion than external investors prefer thereby increases its vulnerability to a hostile takeover. A further perplexity has been largely ignored in the relevant literature. Under standard practice, firms array their prospective investment projects in descending order of internal rates of return (IRRs), accepting all projects with IRRs in excess of the cost of external capital. This descending array of projects is the microeconomic analogue of what Keynes (1936, pp. 135–6) called 'the schedule of the marginal efficiency of capital.' It is illustrated by the line ME_1ME_1' in Fig. 1.

Assuming this procedure to be unbiased, all projects but the marginal one (at point Z in Fig. 1) yield a return in excess of the rate O_i required to attract outside capital; that is, they yield a supranormal return. Received theory states that in long-run equilibrium, competitive firms do not realize supranormal returns, although monopoly firms may do so. But if the representative competitive firm faces a schedule of investment opportunities such as ME_1ME_1' , how can it not enjoy supranormal returns, contradicting the predictions of the theory? There are two possible answers.

(a) Competitive firms, unlike monopoly firms, may systematically overestimate the profitability of their investments because they fail to take into account the likelihood that product prices and hence returns on investment will be reduced when many other firms,

each an atom relative to the market, make similar investments. This answer can be defended theoretically, but seems inconsistent with the premise that investors try to factor into their decisions all the expected environmental changes that could affect the returns on their investments. Compare Richardson (1959).

(b) The difference between competitive and monopoly firms may lie as much in the shapes of their marginal efficiency functions as in the expected output and price behavior of rival firms. In particular, the monopoly firm's MEI function may resemble ME_1ME_1 in Fig. 1, whereas the typical competitive firm, faced with competition in exploiting investment opportunities as well as in pricing its output, confronts a function more like the dash-dash schedule ME_cME_c , with much less (and in the limit no) possibility of rationally expecting supranormal returns.

3. Getting Things Done

Even when it is agreed that business firms should maximize either their current profits or the discounted present value of their long-term profits, there is many a slip between cup and lip. Difficulties intrude because complex and difficult decisions must be taken in the midst of what von Clausewitz called the fog of war and because decisions must be implemented within complex organizational frameworks. Organizational complexity in turn engenders communication breakdowns, coordination failures, goal conflicts, and motivational problems.

3.1 Bounded Rationality

Running anything but the most pedestrian business enterprise is difficult. Firms' decision makers seldom have a clear view of how consumers will respond to price changes outside the range of those with which experience has been accumulated; that is, demand curves are imperfectly visualized. In oligopolistic markets, the reactions of rivals are crucial to profitability but difficult to predict. Whether technological innovations will function as hoped is uncertain. Even less certain is how consumers will respond to product design changes, minor or radical. Decisions that might turn out brilliantly if rival firms remain passive could fail if rivals respond with aggressive countervailing strategies.

In a debate that raged during the mid twentieth century, it was asserted that business decision makers were not only subjected to great uncertainty about the environment in which they operated, but didn't understand the basic logic of profit maximization as taught by economists. In a famous reply, Machlup (1946) insisted that the variables used in firm decision making

are only 'subjective estimates, guesses, and hunches' and that, just as automobile drivers can solve the problem of passing another car on a two-lane, two-way road without comprehending the differential equations of acceleration and convergence, so also business decision makers (or at least, the successful ones) intuitively solve the equations of profit maximization. Since that time, hundreds of thousands of young people trained in graduate business administration programs have entered the real world of business. It is surely true, even if it once was not, that business decision makers at least understand what marginal analysis means.

The more serious problem is undoubtedly what Herbert Simon (1959) called 'bounded rationality'—that is, the impossibility of having all the information about present and future conditions needed to formulate and solve decision problems with the rigor economic theorists might desire. Seeking to explain how business decision makers function under these conditions, scholars have proposed and extended a 'behavioral theory of the firm' that recognizes real-world information availability and processing limitations (see, for example, Cyert and March 1963). Rather than assuming profit maximization as a well-focused objective, the behavioral theory of the firm postulates that decision makers adopt more limited proximate goals—e.g., setting plausible short-term sales and profit targets—and then 'satisfice' in striving to attain them. The targets may then be ratcheted upward over time to converge, at least under favorable conditions, toward a true profit maximum. Similarly, decision-making processes are simplified to emphasize 'rules of thumb' and 'routines' whose pursuit leads in the general direction of profit maximization and whose modification, under the stress of adverse outcomes or the recognition that environmental variables have changed, tends in a rough way to emulate the outcomes that would emerge from a perfectly rational but informationally unattainable profit maximization exercise (see, for example, Nelson and Winter 1982).

3.2 Decentralization in Complex Organizations

Some suppliers of goods and services such as traditional family farms and 'Mom and Pop' retail outlets have only primitive organizational structures. However, much of the world's business is done by geographically dispersed multi-unit, multi-product enterprises. For them, complex organizational problems cannot be avoided. Hardly anything could be achieved by a large organization, whether it be a business enterprise, a church, or an army, without resort to decentralization. In the simplest models, goals and means are chosen by the organization's head, called the 'principal,' and the work is carried out by lower-level employees, called the 'agents' in the large body of

literature on principal-agent relationships. In more realistic multi-unit enterprises, strategic decisions on the markets to be served, the units responsible for serving them, and the financial and physical means at those units' disposal are set by top management; more detailed decisions on product specifications, prices, outputs, and marketing methods are made by unit managers; and the work is done by individuals within the various subordinate organizational layers. The paradigm for such decentralization, surveyed historically by Chandler (1962) and said by Williamson (1970, p. 175) to be 'American capitalism's most important single innovation in the twentieth century' is the so-called multidivisional or M-form organization. Under it, corporate leaders supported by a small central staff provide broad strategic direction and assign goals to virtually autonomous operating division heads, whose performance in meeting their goals is monitored by the corporate staff, with rewards and punishments being dealt out by the corporate leadership in proportion to the divisions' goal attainment. Or in still more complex variants, the corporate office sets overall policy and strategy; group executives are responsible for achieving broad product line or geographic region goals; and individual product division heads must operationalize the efforts needed to meet the goals (or bear the consequences if they fail). Although no one-to-one correspondence is possible, it is the individual product division within complex corporate organizations that is most like the 'firm' as it has been traditionally construed in microeconomic theory.

3.3 Information Flow Occlusions

With decentralization come several problems. One is the decentralization of knowledge. Central office executives usually know more than individuals lower in the corporate hierarchy about overall corporate goals (entailing *inter alia* understanding what stockholders want) and sources of finance. Production foremen have the most detailed knowledge of what can and cannot be done in cranking out products. Field sales representatives often know best what kinds of products their customers want and what rival firms are doing. R&D staff members have expertise in determining what product features can be achieved physically—insights that are sometimes sharpened by interaction with customers, but that for more radical product changes lie outside the range of customers' perception. Division managers serve as information collection points and are best positioned to see the 'big picture' in their lines of responsibility. Information on customer needs and production possibilities must flow upward from the shop or field to division managers and corporate officials; information on policies and decisions must flow downward for implementation. But information flows are commonly attenuated and distorted. Lower-level employees are loath to admit

their known shortcomings, problems they anticipate but hope to avoid, and the organizational 'slack' they could turn into higher output if pressed. Through hubris, superiors often speak without listening or hear only what they want to hear. Revolutions in information transmission, processing, and storage technology have mitigated some of these problems, but they have by no means eliminated them. As Andrew Grove, one of the most thoughtful information technology business leaders of the twentieth century, observed (Byrne 2000):

Brains don't speed up. The exchange of ideas doesn't really speed up, only the overhead that slowed down the exchange ... You can reach people around the clock, but they won't think any better or any faster just because you've reached them faster.

3.4 Suboptimization Failures

Decentralization implies pushing decisions down to operating levels at which hands-on knowledge and responsibility for implementation are present. But when this is done, suboptimization failures can occur. Division A may supply a key input to Division B. But to meet its financial performance targets, Division A needs a generous markup of prices above costs which could adversely affect the financial performance of Division B or lead it to substitute a less suitable input for the high-priced input. Timely intervention by higher management, with a possibly adverse impact on the motivation of the sacrificing entity, is required to solve such 'transfer pricing' problems. Or functional group X may have the special skills needed to help functional group Y solve a problem, but the effort required from group X may come at the expense of meeting X's direct performance targets, so the desired cooperation fails. Again, subtle intervention by higher-level management can solve the problem, but only by deviating from the decentralization principle.

3.5 Goal Conflicts

Perhaps the greatest challenge is ensuring that all members of a business organization conform to higher-level goals. The problem begins at the top. In principle, it was said earlier, firms maximize profits, long-run or short-run, to serve the interests of their owners—in the case of corporations, the equity shareholders. But there are many reasons why this may not happen. A pathbreaking book by Berle and Means (1932) observed that as the ownership of modern business corporations became dispersed away from founder-owners toward increasing numbers of shareholders, each with too small a stake to warrant intervening actively in their investment target's management, the ownership function became separated

from the control function. Top managers, Berle and Means continued, exercised effective control over their organizations and enjoyed within broad limits discretion as to the objectives they would seek to advance. They might opt to maximize the value of shareholders' ownership interests, but they might equally well pursue alternative and conflicting goals—for example, maximizing the size of the business empires they control by acquiring other companies on terms that reduce original shareholders' stock values; paying themselves lavish salaries; enjoying perquisites such as richly furnished offices and trips in corporate jets to posh golf and skiing locales; contributing enthusiastically to local community cultural establishments; or simply seeking a 'quiet life' devoid of entrepreneurial risk-taking.

Corporate boards of directors are—in principle—supposed to represent stockholder interests. Reality is more complex. During much of the twentieth century, directors of American business corporations were chosen and appointed by company managers, owed their continuing presence on the board to the managers, and adopted codes of behavior that discouraged sharp questioning and opposition to management positions except in times of crisis (see Mace 1971). As a result, they provided at best a weak check against managerial actions that reduced shareholder value. In other nations, widely differing board of directors structures exist, but again, board members have often served as less than diligent stewards of common shareholder interests.

Gradually, this has changed. During the last two decades of the twentieth century there were pressures from institutional investors (such as pension funds), government regulators, and the stock exchanges throughout much of the world to improve 'corporate governance,' rendering corporate directors less beholden to inside management and more dedicated to defending stockholder interests. In the USA, for example, corporations have been induced to restructure their boards so that 'outside' members comprise a majority, to create board member nomination committees dominated by 'outside' directors, and to form compensation committees in which only outsiders determine the salaries and bonuses of inside managers. Much more emphasis has been placed on bonuses and stock options rather than straight pay as means of motivating top managers to perform well on behalf of stockholders. As a result, boards have come to intervene more aggressively in company decision-making and, when performance has been less than satisfactory, to fire the responsible top executives.

Nevertheless, problems remain. The compensation of top corporate managers in the USA during the 1990s rose to unprecedentedly high multiples of average company salaries. One reason was a booming stock market that enhanced the value of stock options without requiring extraordinary performance on the part of managers. Also, with superior information on

company prospects, inside managers could 'game' the timing of stock option awards and bonus target-setting to maximize the chances of a favorable outcome. And to help them in the task of top manager compensation-setting, board committees hired outside management compensation consultants who instigated emulatory competition among companies to reward their executives—few of them leadership superstars—at least as liberally as peers and in the process raised the average for all.

At the end of the day, leadership and the ability to create effectively functioning teams are crucial to organizational success. Even among those who have reached the top ranks in business organizations, the quality of leadership varies widely. Graduate schools of management have found it much harder to teach leadership than any of the other requisite arts and skills (but see Heifetz 1994). Economists have few qualifications for saying anything sensible about how superior leadership emerges, so the subject is left to other disciplines contributing to this *Encyclopedia*.

The goal conflicts observed between shareholders and top managers are replicated manyfold at the lower levels of corporate organizations. Operating unit managers 'game' against their corporate superiors to set performance targets and bonus schemes that are biased on the side of yielding generous compensation. Knowing that personnel turnover is high and corporate staff memories short, operating unit managers take actions that sacrifice durable profitability for the sake of higher short-run profits and bonuses. Production and sales force supervisors recognize that their tasks could be carried out with smaller staffs, but refrain from profit-enhancing cutbacks to hold reserves against future contingencies and to maintain a 'nice guy' reputation. These incentive conflicts might be reduced in companies with 'superincentive' compensation schemes; for example, in high-technology startup companies that award potentially lucrative stock options to all employees in responsible positions. The phenomenon is a relatively new one on which clarifying research is needed.

4. Natural Selection

Achieving good economic performance at the operating levels of complex organizations is not easy. Communication and incentive failures abound. There is considerable X-inefficiency. But goods and services do get delivered at levels that on average are nothing short of remarkable. One reason is that in national economies with firmly implanted capitalist institutions, most business firms must strive to improve their internal incentive systems and hew to something approximating a profit maximization line lest they be displaced by competitors who do a better job of satisfying consumer wants. In vigorously competitive

industries, firms that stray too far from the path of profit maximization are unlikely to withstand the forces of natural selection (see Alchian 1950, Nelson and Winter 1982). When competitive pressures are weak, however, product market forces may be insufficient to jeopardize non-maximizers' survival. In such cases—more the exception than the rule in well-functioning economies—the disciplinary job is left to the still-imperfect instruments of corporate governance.

See also: Bounded Rationality; Business History; Corporate Culture; Corporate Governance; Information and Knowledge: Organizational; Information, Economics of; Organizational Decision Making; Organizations, Sociology of; Population Ecology

Bibliography

- Alchian A A 1950 Uncertainty, evolution, and economic theory. *Journal of Political Economy* **58**: 211–21
- Berle A A, Means G 1932 *The Modern Corporation and Private Property*. Macmillan, New York
- Byrne J A 2000 Back to the future: Visionary vs. visionary. *Business Week*, August 28: 210–14
- Chandler A D Jr 1962 *Strategy and Structure: Chapters in the History of American Industrial Enterprise*. MIT Press, Cambridge, MA
- Coase R H 1937 The nature of the firm. *Economica* **4**: 386–405
- Cyert R M, March J G 1963 *A Behavioral Theory of the Firm*. Prentice-Hall, Englewood Cliffs, NJ
- Gaskins D W Jr 1971 Dynamic limit pricing: Optimal pricing under threat of entry. *Journal of Economic Theory* **3**: 306–22
- Heifetz R A 1994 *Leadership Without Easy Answers*. Harvard University Press, Cambridge, MA
- Jensen M C 1988 Takeovers: Their causes and consequences. *Journal of Economic Perspectives* **2**: 21–48
- Keynes J M 1936 *The General Theory of Employment Interest and Money*. Harcourt, New York
- Leibenstein H 1966 Allocative efficiency vs. “x-efficiency”. *American Economic Review* **56**: 392–415
- Mace M L 1971 *Directors: Myth and Reality*. Harvard Business School Division of Research, Boston
- Machlup F 1946 Marginal analysis and empirical research. *American Economic Review* **36**: 519–54
- Marris R 1963 A model of the managerial enterprise. *Quarterly Journal of Economics* **77**: 185–209
- Nelson R R, Winter S G 1982 *An Evolutionary Theory of Economic Change*. Harvard University Press, Cambridge, MA
- Richardson G B 1959 Equilibrium, expectations, and information. *Economic Journal* **69**: 223–37
- Scherer F M 1988 Corporate takeovers: The efficiency arguments. *Journal of Economic Perspectives* **2**: 69–82
- Simon H 1959 Theories of decision-making in economics and behavioral science. *American Economic Review* **49**: 253–83
- Simon H 1979 Rational decision-making in business organizations. *American Economic Review* **69**: 493–513
- US Bureau of the Census 1986 *1982 Enterprise Statistics: General Report on Industrial Organization*. ES-82-1, Washington, DC

Williamson O E 1970 *Corporate Control and Business Behavior*. Prentice-Hall, Englewood Cliffs, NJ

F. M. Scherer

First Language Acquisition: Cross-linguistic

In this article, motivations for cross-linguistic research into children's acquisition of their first language(s) are noted with a brief historical outline of the field (Sect. 1). This is followed by a review of trends which appear to be universally shared across children acquiring different languages (Sect. 2), compared with the impact of typological properties of particular target languages on processes of acquisition (Sect. 3). The article concludes with a note on desiderata (Sect. 4).

1. Approaches to the Field

The major motivation for the cross-linguistic study of language acquisition, as of cross-linguistic research in general, is to compare what is universally shared across children irrespective of the specific target language or speech community to which they belong, on the one hand, with interlanguage variation and the impact of properties peculiar to a particular language or language type on the other (Slobin 1997b). A related, psycholinguistically motivated question is how the specifics of target-language typology interact with acquisitional principles and strategies deriving from more general extralinguistic factors of perception, cognition, and social interaction (Slobin 1985b).

The first studies of children's acquisition of their native language(s) which can be characterized as cross-linguistic in orientation date from the nineteenth and early twentieth centuries. These were generally diary studies of childhood bilingualism, among children raised in more than one language, mainly in Europe. Such studies provided potentially rich data for cross-linguistic comparisons, as did a range of studies of children acquiring languages other than English (Ferguson and Slobin 1973). However, these were by and large not systematically directed to the testing of psycholinguistically motivated hypotheses, and they tended to focus on phonology, inflectional morphology, and the lexicon, with little concern for syntax or semantics.

The rise of developmental psycholinguistics in the USA in the 1960s created a field of inquiry which aimed to integrate insights from linguistic theory and description with developmental patternings in first language acquisition. Such studies tended to be anglo-centric, but provided a basis for research in other

languages. For example, the general developmental stages delineated in Brown (1973) for the early years of grammatical development appear to apply across languages (see Sect. 2 below).

This anglocentricism was countered by the Berkeley project on four typologically distinct languages (Slobin 1982). Since then, data have been collected from a wide range of languages, both across and outside Europe. These include surveys of several dozen typologically distinct languages in the five volumes edited by Slobin (1985–1997). Other collections focus on particular domains including: early clause structure (Bowerman and Perdue 1990), the lexicon–syntax interface in verb acquisition (Lieven 1998), and narrative construction (Berman and Slobin 1994, Hickmann et al. 1996). Generative grammars gave rise to a parameter-based analysis of acquisition in different languages and domains, for example, Hyams (1986) and Weissenborn (1992) on null subjects, Meisel (1992) on functional categories and the ordering of constituents (see *Generative Grammar*).

Research in child language since the 1980s has thus yielded a large body of data on acquisition of numerous languages, including ones with non-European typology. Many of the latter studies are anchored in ethnographic-type fieldwork, for example, for Inuktitut by Shanley Allen, for Mayan Quiché by Clifton Pye, and for Sesotho by Katherine Demuth; or in research conducted by native speakers, as in Soonja Choi's work on Korean. This work combines to demonstrate that (a) children early on adapt to the particular constraints of their native-language typology, as verb-initial or verb-final, accusative or ergative, as synthetic or agglutinating; (b) they do so on the basis of the particular cultural norms for caretaker–child interaction in their speech community; (c) they do so with equal rapidity and ease across languages; and (d) within languages, the acquisition of certain subsystems may be delayed for language particular reasons of relative structural and/or semantic complexity, on the one hand, and because they belong to restricted registers of usage not accessible to young children, on the other.

2. Universal Trends in Language Development

Different points of departure underlie claims about universals in language acquisition. A nativist, Chomskyan-based view will specify these in terms of the formal principles characterizing a generative model of Universal Grammar (UG) with cross-linguistic variation accounted for by differential setting of structural parameters according to the target language. A Piagetian orientation will explain them in terms of general cognitive principles and shared conceptual underpinnings which provide a universal epistemological basis for acquisition of linguistic

knowledge (Slobin 1985). The view espoused by this author is that children start out the task of acquisition equipped with categorial and semantic distinctions of the type termed by Chomsky 'substantive' as against formal universals, akin to Keenan's (1979) idea of 'naive' or primitive universals (Berman 1993b). For example, children do not need to learn that there are two main kinds of sound segments, consonants and vowels; that propositions distinguish between predicates and arguments; or that there are intransitive and transitive clauses, with one-, two-, or three-place predicates.

Whatever the explanation they propose, researchers agree that there is a clear biological underpinning to the overall developmental stages manifested by children in command of language production irrespective of mother tongue. Just as children sit before they stand, and walk before they run, so children across the world babble before they produce single words, rely on single words before they start combining words, and produce isolated clauses before they join these together by increasingly specific lexical means and complex syntactic devices.

Overall progression thus involves an evolution in construction types. Language production typically starts out with utterances, as behavioral elements which provide the initial basis for expressing relations between semantic content and phonetic form. These typically take the form of isolated lexical items or unanalyzed strings of words, starting around the end of the first year of life. Structure dependence emerges when these lexical units are combined in simple-clause structure, realized by morphosyntactic markings of verb–argument relations. Next, adjacent clauses are combined, first by coordination, subsequently by various kinds of subordinate constructions—initially mainly finite and subsequently nonfinite—leading up to complex clause-linkage by derived nominalizations (Berman 1993a). This needs to be qualified for a verb-final language like Turkish, where the unmarked means of clause linkage is by nonfinite, nominalized constructions. These emerge relatively early in Turkish compared with European or Semitic languages.

With respect to acquisition of morphology, the domain in which languages tend to differ most, children generally acquire the bulk of the inflectional system of their language before they master its derivational morphology. This is understandable since inflection is more regular, productive, and grammatically obligatory than derivation. Besides, derivational processes depend on a level of vocabulary expansion not achieved by young preschool age. In fact, in some instances, like derived nominals or the high-register Latinate lexicon of English, children will master derivational processes only well into school age. In contrast, they by and large command the inflectional systems of their language by the age of three, irrespective of how morphologically opaque or synthetic these happen to be. Cross-linguistic differences affect

acquisition of inflection along two main dimensions. Psychological processing principles like perceptual saliency coupled with regularity and transparency of form-meaning mappings account for the fact that, for example, Turkish-speaking children gain command of the bulk of their rich system of agglutinating, inflectional suffixes by a much earlier age than children acquiring languages with more synthetic or syncretic morphologies like Hebrew or Icelandic. And experiential factors of frequency of use in the everyday colloquial speech that constitutes the linguistic input and output of young children explain why certain grammatical categories are acquired relatively late in languages where they constitute high-register or literate constructions.

In the domain of syntax, children typically gain command of simple clause structure in their language between ages two to three years. Basic word order emerges very early indeed, first between phrasal constituents, only later within phrases. Constituent ordering is followed by grammatical marking of verb-argument relations, around the same time or somewhat before inflectional marking of grammatical agreement. The overall progression is thus much as Brown (1973) defined for English: from single words to early word combinations, thence to simple clause structure, followed by a range of devices for clause linkage.

Another commonality is the general development observed across children from rote-learning of unanalyzed items to partial productivity followed by command of structure-dependent rules and generalizations, which are subsequently constrained by conventions of lexical and discourse usage. This has been demonstrated, for example, for how children acquire the semantics, morphology, and syntax of causative verbs in three typologically distinct languages: English at the more analytic or syntactic end of the scale, Hebrew with relatively synthetic bound morphology, and Inuktitut as a polysynthetic language which marks causativity by both morphological and lexical means (Allen, in Lieven 1998).

General language development thus proceeds across different phases, from pregrammatical to structure-dependent to appropriate use. And it becomes both more elaborated and varied in structure and use, on the one hand, and more restricted or constrained by target-language specifics, on the other. Thus, in acquiring the sound system, babies proceed from universal crying to uttering a range of speech sounds, followed by babbling, which is increasingly restricted to elements of the native language to a point where eventually the system will be so constrained as to impose a non-native pronunciation in other languages. Similarly, in morphology, across languages children proceed from nonalternation between forms to immature strategies for initial markings of grammatical categories, thence to 'creative errors' of over-regularization, until eventually the system becomes fully

constrained by lexical convention. And in the acquisition of the lexicon, again across languages, children move from a restricted vocabulary of nursery terms and over or underextensions of conventional items to a period of creative innovation; eventually the conventional lexicon takes over, while at the same time vocabulary growth is both constrained and elaborated by socially and culturally appropriate distinctions of register and usage, lasting into adolescence.

3. Impact of Target-language Properties

As against these cross-language commonalities and claims for universal constraints, whether formal-linguistic or perceptual-cognitive, current research reveals a marked effect of target-language typology from a very early age. Such more language-specific claims have been explained in cross-linguistic perspective as reflecting a modified, neo-Whorfian characterization of the relation between language and thought (Slobin 1996) and as demonstrating the critical impact of caretaker input from the very start (Bowerman 1995). This very early sensitivity to target-language properties has been observed in a variety of linguistic domains. For example, language-specific effects exert an influence on speech perception and babbling during the first year of life (Juczyk 1997). And well before they have command of grammatical inflection, children will adjust their speech output to the prosodic character of their language, as intonation or tone-based (Demuth 1993), and they will show rapid command of vowel-harmony where required, as in Turkish (Aksu-Koç and Slobin 1985).

Children's encoding of the semantics of space and which distinctions they mark linguistically (e.g., in shape or location of figure and ground respectively) is also markedly influenced by target-language specifics. Thus, the universality of the early preference among children acquiring European languages (and also Hebrew) for the prepositions *in*, *on*, and then *under*, which was convincingly explained in terms of perceptual properties of space and shared cognitive preferences, has now been queried, since it turns out that quite different factors govern the encoding of spatial distinctions in languages like Korean or Tzeltal compared with English and Dutch (Bowerman 1995). Children appear to relate to the categories relevant to their own native language without initially manifesting universal patternings in the expression of spatial distinctions.

Null subjects are another domain in which universalist claims have been made. Thus, Hyams (1986) suggested that children first set the prodrop parameter to the unmarked value of plus, irrespective of target language, since English-speaking children in the early stage of grammar acquisition produce numerous

subjectless sentences, as do their Italian peers. Weissenborn (1992) showed, however, that target language input has a marked influence from very early on in this respect, since children acquiring French produce fewer subjectless sentences than those acquiring German, and the latter do so less than children acquiring strongly pro-drop languages like Italian and Spanish.

The very early impact of target language is evident not only in which categories are distinguished, but also how. This is clearly shown in Clark's (1993) cross-linguistic comparisons in the domain of word formation. When young children express the notion of 'undoing' or reversal of actions, they do so from early on in ways peculiar to their native tongue: by the prefixes *un-* in English and French *dé-*, and by particles like *ab* and *aus* in German, *af* and *uit* in Dutch. Hebrew-speaking children, in contrast, must do this lexically or periphrastically, since their language lacks the morphological means for this semantic category. Similarly, in order to express causativity, children will start by using syntactic means with verbs like *make* in English and *faire* in French, a construction type used only rarely by Hebrew-speaking children for this purpose. Subsequently, English-speaking children will overextend verbs like *fall* or *sing* to transitive contexts; in Inuktitut, children will first use morphological and then lexical means of causativity while in Hebrew, they will first neutralize and then alternate morphological patterns to mark verb transitivity (Allen in Lieven 1998, Berman 1993b).

This research converges to suggest that apparently universal acquisitional principles of structural simplicity and semantic transparency in form-meaning mappings are affected and perhaps even superseded by typologically pervasive properties of the target language. For example, it has been argued that children will find it easier to juxtapose words than to affix bound elements to a stem or root, and that these processes are perceptually more salient and so more accessible to young children than stem-internal vowel alternation. However, in a language like Hebrew, which uses vowel alternation to express basic grammatical distinctions such as verb-tense, children mark past vs. present tense at much the same time as their counterparts acquiring languages which use more linear morphological means. Moreover, although their language allows both noun compounding and external, linear affixation to a stem for new word formation, Hebrew-speaking children typically innovate nouns, verbs, and adjectives by the more typically Semitic device of attaching syllabic affixes to a consonantal root: and they rarely use the option favored by English-speaking children for creating novel verbs by zero derivation or syntactic conversion (Berman 1999).

Children's acquisition of word classes and the composition of their early lexicon are other areas which suggest that claims for universal underpinnings

of acquisition may need to be modified. Prior research on children's lexical repertoire at the one-word stage in several languages showed that more nouns were acquired earlier than verbs, while the vocabulary spurt in the period before early word combinations also consisted mainly of noun like terms. Gentner (1982) explained this in terms of general cognitive principles, since verbs as relational terms are holistic rather than coherently packaged entities, while nouns (at least those in the input and output of one- to two-year-olds) tend to encode semantically 'natural' categories that refer to concrete objects which are perceptually identifiable. Recent research has shown, however, that this noun bias may be an artifact of the nature of the lexicon and lexical input in the target languages, and that in some cases, verbs may be acquired earlier and in greater quantities than nouns (for example, the studies by Brown and by Choi in Lieven 1998).

At more advanced stages, developments in form-meaning mappings in extended discourse provide further evidence for the impact of 'typological imperatives.' A five-language cross-linguistic sample of picturebook narratives produced by children aged three to five and nine-year-olds compared with adults (Berman and Slobin 1994) showed that the expression of textual temporality was critically affected by native-language tense/aspect systems even among our youngest subjects. In Hebrew, children marked only grammatical tense, in Spanish they distinguished both simple from progressive and perfective from imperfective aspect, while in Turkish they gave grammatical expression to modality. Moreover, there was no evidence for 'compensation' in the sense that children attempted to give lexical expression to notions not grammatically encoded in their language (e.g., by using periphrastic adverbs to mark aspectual distinctions in languages like German or Hebrew). On the contrary, not only 'do speakers tend to limit their attention to those semantic distinctions marked in the grammar, they may even overattend to such distinctions in early phases of development' (Berman and Slobin 1994, p. 619). Slobin (1996a, 1996b, 1985-97) has since extended these observations to the expression of locative trajectories. Proceeding from Talmy's (1985) distinction between verb-framed languages (Hebrew, Spanish, and Turkish in our sample) and satellite-framed languages like English and German that rely heavily on locative particles, he shows that the way young children—like adult speakers of the same language—talk about motion, paths, and directionality depends critically on how these notions are grammatically and lexically encoded in their language. Here, too, the underpinnings of such distinctions are evident very early: at the one-word stage, Germanic-speaking children use particles like *off* or *aus* to express direction of motion in ways encoded by verb-stems by children acquiring French, Spanish, or Hebrew. For example, at the age of 18 months, my Hebrew-English bilingual daughter would say either English *down* or a

truncated stem form of the Hebrew verb *la-redet* 'to get/go down,' when she wanted to be taken out of her high chair.

In another cross-linguistic project comparing children's narrative texts produced by children aged from five to 15 years on the basis of two picture-series, Hickmann et al. (1996, p. 391) focused on the devices for introducing, maintaining, and shifting reference to characters in a story. They found that the English, French, German, and Chinese 'rely differentially on local vs. global devices to mark newness: postverbal position is obligatory in Chinese ..., indefinite determiners in the other languages; across languages: newness markings emerge late (age seven), local markings emerge before global, and local and global markings are strongly related.' Their findings point to an important interaction between shared developmental trends and linguistic specificity. Relatedly, in the Berman and Slobin (1994) study, across languages, three-year-olds tended to relate to each picture in isolation, four- and five-year-olds were able to encode linearly expressed local relations between events in terms of local temporal relations of sequentiality and subsequently of causality, school-age children constructed a globally organized hierarchical narrative, and adults tended to adopt individual rhetorical styles with rich interpretation and a personal narrative stance.

Ongoing cross-linguistic research on the production of narrative and expository texts by schoolchildren and adults reaffirms such interaction between generally shared developmental trends and typologically determined expression of semantic distinctions and selection of rhetorical options. Nine-year-olds can all construct wellformed personal experience narratives, but are unable to produce coherent expository texts. Further, only from high-school age will one find consistent, deliberate, and appropriate differentiation between language usage across spoken and written texts, marking highly language-particular register distinctions. In contrast, overall genre distinctions appear similar across languages. For example, narrative texts are couched in past tense and mainly perfective aspect, and they have light subjects (mainly pronominal in English and French or null subjects in Hebrew and Spanish), while expository texts rely more on generic or extended present tense and have more lexical and heavier, more noun-phrase subjects and objects. This, too, interacts with target-language options and preferences. For example, in order to express a detached, impersonal perspective in English, passive is quite widely used even by grade-school children, increasingly by adults; and passive is also quite common in the Icelandic texts. In contrast, passive is used far less in other languages, even though it occurs as a structural option: Swedish subjects rely heavily on the impersonal pronominal *man* corresponding to the very common use of *on* in French; Spanish and Hebrew achieve the same rhetorical effect by means of sub-

jectless or second person impersonal constructions. Similar patterns emerged for differential use of pronominal compared with null subjects and also for relative clauses as devices for syntactic packaging in the picturebook narratives.

Discourse-based cross-linguistic comparisons thus suggest that reliance on certain linguistic forms may depend not only on structural factors of relative morpho-syntactic complexity or productivity, but also on the extent to which a given rhetorical option is favored for specific discourse purposes in current language usage.

4. Conclusion

In sum, on the basis of shared cognitive and linguistic universals, even very young children are attentive to key typological properties of their language. From an early age, they recognize 'where the action is at' in the input language, not only which categories are formally distinguished, but also how. *When* this type of sensitivity finds expression will depend on general developmental, largely extralinguistic factors. For example, the spatial distinctions studied by Bowerman (1995) and her colleagues will precede command of derivational marking of linguistic subcategories, and these will emerge earlier than rhetorical mastery of linguistic forms in the context of extended narratives. But in each case, the specific ways in which children encode form-meaning relations accord with how this is done by adult speakers of the same target language rather than by children of the same age in other languages.

As noted in the introduction, a large body of data is currently available for acquisition of different domains in a rich range of languages from Europe and elsewhere. What now seems necessary is to extend systematic and deliberately comparable research designs across a range of languages and linguistic domains—along lines conducted for morpho-syntax by Slobin (1982), for word formation and derivational morphology by the structured elicitation and diary studies reported in Clark (1993), for semantic encoding of space by Bowerman (1995), for temporality by Weist and co-workers (Weist et al. 1993), as well as for narratives by Hickmann (1996) and by Berman and Slobin (1994). Such studies might profitably focus on two contrasting directions of cross-linguistic comparison: (a) languages which are typologically very distinct, for example, ergative or accusative, verb-initial or verb-final, analytic versus synthetic or agglutinating; and (b) languages which are typologically close, like Dutch and German, or Arabic and Hebrew, but which differ along certain restricted dimensions. Such work should aid in specifying the relative weight of the various factors touched on here. It would help to delineate the interaction between language-universal and language-particular, between intralinguistic,

cross-linguistic, and extralinguistic factors in acquisition. And it could provide psycholinguistic motivation for the distinction between language-specific, typologically shared, and cross-linguistic or universal features of language structure and use in developmental perspective.

See also: Language Acquisition; Language Development, Neural Basis of; Linguistic Typology; Second Language Acquisition; Vocabulary Acquisition, Psychology of

Bibliography

- Aksu-Koç A A, Slobin D I 1985 The acquisition of Turkish. In: Slobin D I (ed.) *The Cross-linguistic Study of Language Acquisition: Vol. 1*. Erlbaum, Hillsdale, NJ, pp. 839–78
- Berman R A 1993a Crosslinguistic perspectives on native language acquisition. In: K Hyltenstam, A Viberg (eds.) *Progression and Regression in Language*. Cambridge University Press, Cambridge, UK, pp. 245–66
- Berman R A 1993b Developmental perspectives on transitivity: A confluence of cues. In: Levy Y (ed.) *Other Children, Other Languages: Issues in the Theory of Language Acquisition*. Erlbaum, Hillsdale, NJ, pp. 189–241
- Berman R A 1999 Children's innovative verbs versus nouns. In: Menn L, Bernstein-Ratner N (eds.) *Methods for Studying Language Production*. Erlbaum, Mahwah, NJ, pp. 69–93
- Berman R A, Slobin D I 1994 *Relating Events in Narrative: A Crosslinguistic Developmental Study*. Erlbaum, Hillsdale, NJ
- Bowerman M 1995 Learning how to structure space for language: A crosslinguistic perspective. In: Bloom P, Peterson M, Nadel L, Garrett M (eds.) *Language and Space*. MIT Press, Cambridge, MA, pp. 385–486
- Bowerman M, Perdue C (eds.) 1990 *Linguistics Special Issue: The Structure of the Simple Clause in Acquisition*, pp. 28–36
- Brown R 1973 *A First Language: The Early Stages*. Harvard University Press, Cambridge, MA
- Brown P 1998 Children's first verbs in Tzeltal: Evidence for an early verb category. In: Lieven E (ed.) *Linguistics Special Issue: Developing a Verb-category: Cross-Linguistic Perspectives*, pp. 713–54
- Clark E V 1993 *The Lexicon in Acquisition*. Cambridge University Press, Cambridge, UK
- Demuth K 1993 Issues in the acquisition of the Sesotho tonal system. *Journal of Child Language* 20: 275–301
- Ferguson C A, Slobin D I (eds.) 1973 *Studies of Child Language Development*. Holt, Rinehart and Winston, New York
- Gentner D 1982 Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In: Kuczaj S A II (ed.) *Language Development, Volume II: Language, Thought, and Culture*. Erlbaum, Hillsdale, NJ, pp. 301–34
- Hickmann M 1998 Form, function, and context in narrative development. *Journal of Pragmatics* 29: 35–56
- Hickmann M, Hendriks H, Roland F, Liang J 1996 The marking of new information in children's narratives: A comparison of English, French, German, and Mandarin Chinese. *Journal of Child Language* 23: 591–620
- Hyams N M 1986 *Language Acquisition and the Theory of Parameters*. Reidel, Dordrecht, The Netherlands

- Jusczyk P W 1997 *The Discovery of Spoken Language*. Bradford Books, Cambridge, MA
- Keenan E L 1979 On surface form and logical form. *Studies in Linguistic Sciences* 8: 2
- Lieven E 1998 (ed.) *Linguistics Special Issue: Developing a Verb-category: Cross-Linguistic Perspectives*, pp. 36–4
- Meisel J M (ed.) 1992 *The Acquisition of Verb Placement*. Kluwer Academic Publishers, Dordrecht, The Netherlands
- Pye C 1994 A crosslinguistic approach to the causative alternation. In: Levy Y (ed.) *Other Children, Other Languages: Issues in the Theory of Language Acquisition*. Erlbaum, Hillsdale, NJ, pp. 243–64
- Slobin D I 1982 Universal and particular in the acquisition of language. In: Wanner E, Gleitman L R (eds.) *Language Acquisition: The State of the Art*. Cambridge University Press, Cambridge, UK, pp. 128–70
- Slobin D I 1985–1997 (ed.) *The Crosslinguistic Study of Language Acquisition, Volume 1* (1985a), *Volume 2* (1985b), *Volume 3* (1993), *Volume 4* (1997a), *Volume 5* (1997b). Erlbaum, Hillsdale, NJ
- Slobin D I 1996 From 'thought and language' to 'thinking for speaking'. In: Gumperz J J, Levinson S C (eds.) *Rethinking Linguistic Relativity*. Cambridge University Press, Cambridge, UK, pp. 70–96
- Talmy L 1985 Lexicalization patterns: Semantic structure in lexical forms. In: Shopen T (ed.) *Language Typology and Syntactic Description, Vol. 3: Grammatical Categories and the Lexicon*. Cambridge University Press, Cambridge, UK, pp. 36–149
- Weissenborn J 1992 Null subjects in early grammars: Implications for parameter-setting theory. In: Weissenborn J, Goodluck H, Roeper T (eds.) *Theoretical Issues in Language Acquisition: Continuity and Change in Development*. Erlbaum, Hillsdale, NJ, pp. 269–99
- Weist R M, Kaczmarek A, Wysocka J 1993 The function of aspectual configurations in the conversational and narrative discourse of Finnish, Polish, and American children. *Papers and Studies in Contrastive Linguistics* 27: 76–109

R. A. Berman

First Past the Post, in Politics

1. First Past the Post, in Politics

First past the post (FPTP) is the name given to the simplest possible system of electing persons to office. Also known as *single member plurality* system, FPTP refers to contests where the candidate with the most votes, and only that candidate, is elected from a given electoral district. FPTP is used in legislative elections in 68 of the 211 independent states and related territories identified in the *International IDEA Handbook of Electoral System Design* (1997, p. 20). Even though FPTP is used in about a third of the states and territories, countries with FPTP constitute nearly half of the populations included in these units, because

they include populous countries such as Bangladesh, India, Nigeria, Pakistan, the UK, and the USA. Many of the countries with FPTP are former British colonies.

Two features distinguish FPTP from other electoral systems: the number of elected seats per district (one) and the rules by which those seats are allocated among candidates (simple plurality). Because both of these factors contribute to the unique character of FPTP, it is worth considering them in sequence.

1.1 Single-member Districts

Under FPTP, only one person is elected per district. In the US House of Representatives, for instance, although there are 435 districts, each district is represented by a single person (hence the term 'single-member district'). The same is true for the British House of Commons. In contrast to FPTP, in many national legislatures, the top two, three, or n vote-getters are elected to represent the district. In the extreme case, the entire country is a single multi-member district, and all legislators can be said to represent the same district. The Israeli Knesset, for example, has a solitary district comprising the entire country, and all 120 members of Knesset represent it. The same is true of the Dutch legislature, which has 150 members from a single national district.

In the systematic comparison of types of electoral systems, the number of elected seats within a district is termed the *district magnitude*. In FPTP, where the district magnitude is one, voters usually vote directly for candidates. If a party's candidate wins in a district, that party has won essentially 100 percent of the representation for that district, despite receiving only a portion, and not necessarily more than half of the total votes. For example, suppose that there are three parties evenly distributed over a hundred districts. If the vote in each district is 40 percent for party A and 30 percent each for parties B and C, party A will have 100 percent of the seats with 40 percent of the votes. This pattern (with more realistic numbers) often gives the largest single party a proportion of seats in the legislature that is higher than its proportion of votes in the electorate. Thus, leading or majority parties in FPTP tend to be over represented.

This exaggeration of the seat shares of the parties that are largest in votes is not unique to FPTP. Virtually all electoral systems do so, but FPTP does so more than proportional representation systems (Rae 1971). However, it is also possible in FPTP systems for a party to receive a majority of votes, but only a minority of seats, for example, when such a party's supporters are heavily concentrated in a few districts.

In systems where the district magnitude is high (and therefore, by definition, not FPTP), voters typically cast their ballots for parties who are then allotted legislative seats more or less in proportion to their

share of the electorate's vote. In general, higher district magnitudes imply some kind of proportional representation (see *Proportional Representation*).

1.2 Plurality Rule

Within single-member district systems, a winner can be picked via one of several alternative rules. Plurality, the simplest rule, implies that the one with the most votes wins. Other options include absolute (more than 50 percent) majority requirements, with runoff elections between the top two finishers if necessary; approval voting, wherein voters give one vote to each candidate they approve of; the Condorcet procedure, in which the winner of a series of paired comparisons (if there is one) is designated the winner; and the Borda count, wherein points are assigned by voters to candidates in order of their preference.

When there are only two candidates, simple plurality rule is equivalent to majority rule and has four desirable properties: decisiveness, anonymity, neutrality, and positive responsiveness (May 1952). When there are more than two candidates, neither plurality rule nor any other system fulfils another (weaker) handful of simple desiderata: universal admissibility of preference orderings, the Pareto criterion, independence of irrelevant alternatives, and nondictatorship (Arrow 1963).

1.3 Consequences of FPTP

The great virtue of FPTP is its simplicity. Voters make a single choice and the candidate with the most votes wins. However, FPTP has fewer desirable properties than many other methods in multicandidate elections. Compared to other rules, plurality is the least likely to pick a Condorcet winner, i.e. a candidate who is able to beat all other candidates in pairwise contests. Also, simulations show that the plurality rule results in average social utilities that are lower than those under other rules (Merrill 1988). Such drawbacks may not be very consequential when there are only two candidates. Furthermore, when there are only two options, as in a two-party system, there is no need for more complex electoral systems.

When three or more candidates compete in a FPTP election, some of the candidates may divide the votes of some segment that might otherwise agree on a winning candidate. This could result in a victorious candidate who is less preferred by the electorate than one of the losers. If voters are aware of this eventuality they may be able to coordinate in order to elect a more preferred candidate. FPTP encourages voters to cast their ballots strategically and consider the behavior of other voters. If a voter's first-choice candidate has little chance of getting a plurality of votes, the voter

may find it advantageous to vote for a less preferred but more viable candidate instead of 'wasting' his or her vote on a candidate with little chance. In contrast, multi-member districts and even single-member districts with decision rules other than plurality may reduce the incentives for voters to deviate from their first choice.

Strategic behavior and an aversion to wasting one's vote means that FPTP elections are typified by contests between two major candidates, whereas alternative systems often exhibit several viable candidates. Extending this result to parties, FPTP systems routinely have only two major parties, a phenomenon that has been coined *Duverger's law*, named after the French scholar who extensively described the relationship (Duverger 1954). Proportional representation systems, however, usually have more than two parties, a result that has come to be known as *Duverger's hypothesis* (Riker 1982, Cox 1997).

When there are only two viable candidates in an FPTP election, no abstentions, and the candidates are choosing their positions to maximize their electoral support, the candidate ideologically closest to the median voter is advantaged (Downs 1957). FPTP systems thus tend to produce more moderate outcomes than alternative systems. The incentive for candidates to take moderate positions can produce candidates who are only minimally differentiated, leading some to conclude that the choices in FPTP are often of little consequence. One potential advantage to such a system, however, is that the outcomes from one election to the next tend to be more ideologically consistent, resulting in small, incremental policy changes over time.

See also: Electoral Systems; Political Representation; Voting, Sociology of; Voting: Tactical

Bibliography

- Arrow K J 1963 *Social Choice and Individual Values*, 2nd edn. Wiley, New York
- Cox G W 1997 *Making Votes Count: Strategic Coordination in the World's Electoral Systems*. Cambridge University Press, Cambridge, UK
- Downs A 1957 *An Economic Theory of Democracy*. Harper, New York
- Duverger M 1954 *Political Parties*. Wiley, New York
- International IDEA Handbook of Electoral System Design 1997. International Institute for Democracy and Electoral Assistance, Stockholm
- May K O 1952 A set of independent necessary and sufficient conditions for simple majority decision. *Econometrica* **20**: 680–4
- Merrill S III 1988 *Making Multicandidate Elections More Democratic*. Princeton University Press, Princeton, NJ
- Rae D W 1971 *The Political Consequences of Electoral Laws*, rev. edn. Yale University Press, New Haven, CT

Riker W H 1982 The two-party system and Duverger's law: an essay on the history of political science. *American Political Science Review* **76**: 753–66

G. D. Adams and W. R. Keech

First World War, The

1. Causes

The outbreak of the First World War in the summer of 1914 has been attributed to accident, design and confusion. The first position suggests that no one intended that armed conflict would break out in 1914; the second, that German and Austrian military elites planned the conversion of a Balkan diplomatic crisis into an armed confrontation; the third, that political and military leaders throughout Europe began to maneuver with some notion that war was possible, but found that they could not resist the momentum of confrontation (Joll 1992).

What was the sequence of events subject to these interpretations, and which has commanded the support of most historians who have addressed the issue of the descent into war in 1914? On 28 June 1914, the Archduke Franz-Ferdinand of Austria–Hungary, heir to the throne, was assassinated in Sarajevo. The crime was carried out by a group of Serb student patriots in neighboring Bosnia, then a province of the Austro–Hungarian empire. Their aim was to demonstrate the force of Slavic nationalism, and to challenge the domination of these south Slavic lands by imperialists in Vienna and Budapest.

In this aim, they succeeded completely. The government in Vienna decided to make this crime a matter of state, and to lay the blame for this political murder on the independent Kingdom of Serbia, and its intelligence services. On 5–6 July, Austrian officials went to Germany to seek support for their policy of making Serbia 'pay' for the assassination. German officials gave their approval, offering a 'blank check' to their Austrian allies. With this backing, on 23 July the Austrians presented Serbia with a list of demands which constituted a challenge to the status quo in the Balkans.

This is where Russia entered the conflict. A humiliation for Serbia would humiliate her Slavic ally, Russia. With Russian backing, Serbia met most of the Austrian demands, but refused to capitulate completely. In support of Serbia, Russia ordered partial mobilization of her armies on 29 July. In support of Austria, and in response to the numerical superiority of Russian forces, the German army mobilized too.

Once German mobilization was in progress, then the crucial link appears between a conflict in the Balkans and an all-European war. German military

planning had for years established a set of priorities to safeguard the security of the German empire. The primary point here was the need to avoid fighting a two-front war. Thus, it was necessary for Germany to eliminate a western threat on the part of Russia's ally, France, before facing the huge armies Russia could put in the field. To do so, a military plan—the Schlieffen plan—was developed through which German forces could move in a huge arc from Cologne in southwestern Germany westward through Belgium and then southward through France. The destination, after 42 days, was in the vicinity of Paris, where the French army would be destroyed, just as had occurred in 1870.

The Schlieffen plan was thus a German response to the threat of a two-front war. That threat had materialized out of the confrontation of Germany's ally Austria-Hungary with Russia's ally Serbia. But the German plan to destroy the French army by an arcing move through Belgium threatened to bring in the UK, guarantor of Belgian independence (Ritter 1958).

And that is precisely what happened. The German invasion of Belgium precipitated a state of war among the five great powers: Germany and Austria-Hungary on one side, and France, Russia and the UK on the other. By 4 August 1914, the First World War had begun.

Had anyone intended this moment to occur? Probably not. While the chief catalyst of the war crisis was the decision by Austria and Germany to press Serbia to pay a political price for the assassination, the outbreak of war was less a conspiracy than a complex mixture of arrogance, a sense of national 'honor,' ignorance, and confusion. All the actors in this drama misjudged the other side. Each move forward in the crisis was followed by a further heightening of tension. First Austrian honor was at stake; then Russian honor; then German fears of encirclement led her to respond to a perceived threat from Russia by threatening Russia's ally, France; and when confronted by a likely invasion of France, the UK hesitated, and then in defense of her honor, she declared war on Germany.

Two elements in the crisis were crucial in the prevention of a diplomatic solution. The first was temporal; the second, structural. First, timing. The war crisis lasted for one month. But a closer view of these events presents us with an even shorter timetable in the slide towards war. One week separated the Austrian ultimatum to Serbia—on 23 July—from Russian and German mobilization—29–31 July. In that time period, it was simply impossible for alternatives to be explored, or for the weight of antiwar and pacifist opinion, by no means negligible, to be brought to bear on political leaders. The crisis moved too fast to be stopped.

The second determinant of a general European war was, broadly speaking, structural. That is, the chronic instability of southeastern Europe was insufficient in

and of itself to bring the Great Powers to war. What turned a minor conflict into a major crisis was the linkage between the Balkan conflict and the longer-term balance of power between Germany and the UK. Once the UK and Germany were aligned on opposite sides, then the crisis of 1914 became a contest for control over northwestern Europe. Should Germany defeat France in 1914, as she had done in 1870, the German navy would occupy the Channel ports. British shipping lanes, necessary to feed her home population, would be open at the pleasure of the German navy. No British government could accept that prospect (Kennedy 1980).

It is true that the British Foreign Secretary, Lord Grey, did not make this issue clear in his diplomatic handling of the war crisis. The German Chancellor, Bethmann-Hollweg hoped that the UK would stay out of the war. But geographical and strategic imperatives overshadowed diplomatic miscalculations. An invasion of Belgium and France posed a real and present danger to British vital interests. And when the UK entered the war, so did the British empire. What had started in Sarajevo, echoed in Sydney, Capetown and Calcutta. By 4 August 1914, the world was indeed at war.

2. *Conduct*

The 1914–18 war spanned the globe, but its outcome was determined by what happened on the battlefields of Western Europe. This article therefore offers an outline of the major military encounters of the war, encounters which left lasting traces both on the landscape of Europe and on the contours of the rest of the twentieth century.

The battlefield in 1914 was Victorian in character. Artillery provided more fire power than ever before, but most of it was horse-drawn. Four years later, the battlefield was transformed. By the end of the war, it took on contours recognizable to most soldiers who would come to fight in later conflicts. Infantry, artillery, tanks, and air power were all coordinated through communications systems and supply systems of great complexity and sophistication. Supplying animal power meant relatively little; supplying machines meant everything (van Creveld 1977). The 'storm of steel' in the words of the German infantryman Ernst Jünger, had come to stay (Jünger 1929).

2.1 1914

The two major military encounters of 1914 produced very different outcomes. In Belgium and France, the German army, one million strong, pushed forward in a swing westward and then southward. It met stiff resistance from Belgian forts, and then was harried by stubborn French and British defensive action. As the

German army moved west into Belgium, the French army moved east towards Alsace and Lorraine, provinces lost to Germany in 1871. German defenses held and inflicted massive casualties on the French. On one day, 24 August 1914, over 24,000 French soldiers died in action. This 'Battle of the Frontiers' was a massive defeat for the French. But the strength of their east-west rail links enabled the bulk of French forces to move west to meet the real threat of the German advance towards Paris. There on the river Marne, the German thrust was stopped. The German army retreated northward and dug in on the river Aisne. When French and British forces pursuing the German army reached the Aisne, they were unable to move the invaders. There the 'Western Front' was born. It soon stretched from Belgium to Switzerland.

There is little doubt that the Battle of the Marne was a German catastrophe. The whole purpose of German strategy had failed. There would be a two-front war. By the end of 1914 the chief of staff of the German army, von Moltke, had been replaced by Erich von Falkenhayn. But what redressed the strategic balance was a massive German military victory on the Eastern front. Two columns of Russian troops advanced into East Prussia at the same time as the German army moved through Belgium and France. These columns were separated by a string of lakes. A German staff officer, Max Hoffmann, took advantage of the geography and of the existence of a north-south rail link. Under the cover of darkness, the German army in the north moved south. The Russian column in the north had no idea they were facing an empty landscape; the Russian column in the south had no idea they were facing a numerically superior force, that is, until they were annihilated at the Battle of Tannenberg. The commander of Russian forces in the south, Samsonov, committed suicide. The two commanders of the German forces, who had not planned the operation, but who reaped its rewards, became national heroes. They were Paul von Hindenburg and Eric Ludendorff (Showalter 1991).

2.2 1915

The war spread in 1915, but the rough stalemate of 1914 remained unbroken. One major attempt to break it was in Turkey, then an ally of Germany. To 'stop our troops from chewing on barbed wire' on the Western front, as Winston Churchill, Britain's First Lord of the Admiralty, put it, a plan was hatched to knock Turkey out of the war by a naval operation. The aim was to use British naval power to force the straits that connected the Mediterranean Sea to the Black Sea. Once done, then the Turkish capital of Constantinople would be indefensible. The defeat of Germany's ally Turkey would serve many purposes. It would open supply lines to Russia; it would convince all and sundry, especially in the Islamic world, of the error of siding with Germany in the world war; it

might even convince Germany to withdraw from France and Belgium.

Nothing of the sort happened. Here too, a mixture of arrogance, ignorance and confusion ruled. The enemy was Oriental, and all the distortions attending Western attitudes towards the Orient flourished. The enemy was underestimated. Attempts to push a naval force through the Dardanelles failed in March 1915. The Allies decide to mount a joint land-sea operation, but no one had a clear idea of the terrain on which soldiers would land. When they went ashore on 24-25 April, they faced steep cliffs and stiff Turkish resistance. They never got off the beaches. A six-month stalemate ensued, after which Allied forces withdrew.

2.3 1916

Two further attempts were made to shift the balance of power on the Western front. The first was at Verdun, in eastern France. There on 21 February 1916, more than one million shells fell on French positions in the hills north of the garrison city of Verdun. The idea was to force the French into total defense of this relatively unimportant position. The French, Falkenhayn reasoned, valued Verdun as a symbol of French freedom and power (Afflerbach 1994). He was right. The French committed 259 out of its full complement of 330 infantry regiments to the struggle for Verdun. They moved into Verdun along a vast conveyor belt, which the French called *la voie sacrée*, the sacred way. It certainly was a field of sacrifice. Falkenhayn succeeded in bleeding the French army white, but he did so at a massive cost to his own troops. Perhaps half a million men died at Verdun, where, by November, the French army had recaptured all the ground lost in the early days of the battle. Lasting ten months, the Battle of Verdun was the longest battle in history. The front remained about where it had been before the bloodbath (Horne 1962).

The same failure to break the stalemate occurred further to the north, where the British and French armies opened a major offensive on 1 July 1916. Here the fault lay in an underestimate of the enemy's fortifications and an overestimate of the power of artillery to destroy them. A one-week barrage by British guns was supposed to obliterate German defenses between Amiens and St Quentin in the lowlands of the Somme. But the deep dugouts under the German lines withstood the barrage, and when British forces moved out early in the morning of 1 July, they were mown down by German machine guns. Of the 100,000 men who went over the top, 60,000 were casualties, of whom 20,000 died in that single day. Thereafter, British and German forces regrouped. A further advance on 14 July was somewhat more successful, but by September, it was apparent that no breakthrough would take place. German lines held, but at a high price. German casualties numbered approximately 450,000; together

French and British casualties surpassed 600,000. One million men fell, and the lines remained roughly where they had been at the beginning of 1916 (Keegan 1976).

Stalemate was also the rule at sea, where the British fleet kept the German fleet penned up to the east of Denmark. One attempt to shift the war to the west at Jutland resulted in mixed fortunes for both navies. But the ultimate result was that the German high seas fleet was kept away from Britain.

2.4 1917

In 1917, the balance of power remained unchanged on the Western front. But in the east and in Italy, Germany and Austria were in the ascendant. First came the crisis in Russia. After the February revolution which toppled the Czar and put a provisional government in power, a provisional government vowed to carry on the war. This was its worst mistake, made evident by a catastrophic offensive launched on 1 July 1917. When the Russian offensive failed, a counter-offensive produced a massive Russian retreat which effectively destroyed the Russian army as a fighting force. When the Bolsheviks came to power in October, they saw what had to be done. They sued for peace, and signed an armistice on 3 December 1917. Germany and her allies had won the war on the Eastern front.

In Italy, the Allied cause suffered a series of alarming reverses. At Caporetto, the Italian army collapsed, and fled westward 100 km where a new defensive line was hastily formed. In France, a spring attempt to pierce German lines at the Chemin des Dames failed, and when the French commander Nivelle persisted in trying to do the 'un-doable,' his army mutinied (Smith 1994). Nivelle was replaced by Pétain; the offensive was called off; the anger of frontline troops abated. British attempts to shift the balance of power in Flanders were also unsuccessful, in part through the unhappy coincidence of their attack at Ypres occurring in the wettest summer on record. The British offensive, which ended in the rubble of the village of Passchendaele on 10 November 1917, literally drowned in mud (Prior and Wilson 1996).

The German navy attempted to shift the balance of power by launching unrestricted submarine warfare in February 1917. This was a colossal miscalculation. It did not starve the UK into negotiating an armistice. Despite massive losses to British shipping, domestic food supplies remained adequate (Offer 1991). What was worse from the German point of view, was the provocation posed by naval attacks to the USA, which entered the war in April 1917.

2.5 1918

In January 1918, it would have been difficult for a neutral observer to predict Allied victory by the end of

that year. How did it happen? First, the political and economic balance of power shifted with US entry into the war in 1917. Given the time needed to raise and then move an army of two million men across the Atlantic, it was apparent that US troops and US supplies would directly enter the balance of power only in 1918. This gave the Allies a new and deep reservoir of men and materiel which the Central Powers could not match. Thus, it was not US firepower that turned the tide, but US reserves.

This became apparent after German army launched its last major military gamble of the war. On 21 March, 47 divisions of the German army broke through at the point where British and French forces met on the Western front. Finally, the Western front moved. Within a week, German forces had moved 40 miles (65 km) to the west. In April they attacked at Ypres; in May and June to the south, reaching the Marne. But there the offensive broke down. The gamble had failed. The German army had lost one million men in a last, desperate and futile effort to win the war (Keegan 1999).

When German forces saw that they could not win, the balance of power shifted radically towards the Allies. Starting in early August, German troops began to surrender en masse. High Command was getting reports of low morale among troops, who were convinced that further sacrifices were in vain. At this point, unrest on the home front merged with despair over the military situation. The High Command saw that the game was up (Deist 1991). At the same time, Austrian and Bulgarian positions crumbled. Ludendorff urged the Kaiser to turn to the US President, Woodrow Wilson, to broker an Armistice. Ludendorff resigned. The Kaiser fled to the Netherlands, leaving it to others to face a disaster he and his military had engineered. The Armistice came on 11 November 1918. The Great War was over.

3. Consequences

3.1 Demographic

Total casualties and losses as a proportion of those who served passed a threshold beyond previous experience. Wherever the threshold is, the total of roughly nine million dead soldiers (according to varying estimates) is beyond it: this constitutes roughly one in eight of the men who served. Adding statistics on other casualties, it is apparent that roughly 50 percent of the men who served were either prisoners of war, wounded or killed.

The most murderous theatre of operations was the Eastern front, where disease and enemy action described the course of a nineteenth century war waged with twentieth century weapons. Of all Serbs who served in the war, 37 percent were killed; roughly one in four Rumanians, Turks and Bulgarians also perished. On the Western Front, where the war was won and

lost, combat was about half as lethal: German and French losses were about one in six of those who served; British losses were one in eight.

Initially casualties among social elites were higher than among the rest of the population. The longer the war lasted, the greater was the democratization of loss. The reason is that officer casualties were higher than those in the ranks, and the social selection of the officer corps mirrored inequalities in prewar life. Consequently in its initial phases, the higher up in the social scale a man was, the greater were his chances of becoming a casualty of war. By 1917, elites were sufficiently decimated to require the armies to draw junior officers from wider social groups which in their turn suffered disproportionately higher casualties in the last two years of the war.

Among the poor and the underprivileged, the story is different. Prewar deprivation saved the lives of millions of working class men and poor peasants, whose stunted stature and diseases made it impossible for them to pass even the rudimentary standards of medical fitness for military service during the war. In the British case, roughly 35 percent of the men

examined for military service were either unfit for combat or unfit to wear a uniform at all (Winter 1985).

Numerically, workers and peasants formed the vast majority of the army of the dead. The same was true of disabled men, visible reminders of the war on street corners and in farming villages throughout the world.

More lethal than combat was a visitation of epidemic influenza, colloquially known as 'the Spanish flu.' In 1918 and 1919 this mutant virus killed young, healthy adults throughout the world. The war may have helped spread the disease, though it did not cause it.

3.2 Political

The Treaty of Paris, the peace settlement ending the Great War, was signed on 28 June 1919, five years to the day after the assassination of Franz Ferdinand in Sarajevo. Its major effects were to ratify the break-up of the Austro-Hungarian and Turkish empires, and to establish a new state system in Europe, wherein Germany was stripped of Alsace and Lorraine in the west and much of East Prussia in the east. Primary

Table 1

Some estimates of military losses among combatant countries in the 1914–18 war

	Total killed or died	Total mobilized	Prewar male population (aged 15–49)	Total prewar population	Total killed		
Country	(in thousands)				per 1,000 mobilized	per 1,000 males (aged 15–49)	per 1,000 people
Britain and Ireland	723	6,147	11,540	45,221	118	63	16
Canada	61	629	2,320	8,100	97	26	8
Australia	60	413	1,370	4,900	145	44	12
New Zealand	16	129	320	1,100	124	50	15
South Africa	7	136	1,700	6,300	51	4	1
India	54	953	82,600	321,800	57	1	0
France	1,327	7,891	9,981	39,600	168	133	34
French colonies	71	449	13,200	52,700	158	5	1
Belgium	38	365	1,924	7,600	104	20	5
Italy	578	5,615	7,767	35,900	103	75	16
Portugal	7	100	1,315	6,100	70	5	1
Greece	26	353	1,235	4,900	73	21	5
Serbia	278	750	1,225	4,900	371	227	57
Rumania	250	1,000	1,900	7,600	250	132	33
Russia	1,811	15,798	40,080	167,000	115	45	11
USA	114	4,273	25,541	98,800	27	4	1
Allied total	5,421	45,001	204,018	812,521	120	27	7
Germany	2,037	13,200	16,316	67,800	154	125	30
Australia–Hungary	1,100	9,000	12,176	58,600	122	90	19
Turkey	804	2,998	5,425	21,700	268	148	37
Bulgaria	88	400	1,100	4,700	220	80	19
Central Powers' total	4,029	25,598	35,017	152,800	157	115	26
Grand total	9,450	70,599	239,035	965,321	134	40	10

Source: Winter (1985).

responsibility for the war, and for making material reparations for it, were laid at the feet of the new German Republic.

Through a League of Nations, international conflict would be regulated, if not resolved. An International Labour Organization was established to help address problems of working class life, a clear response to the Bolshevik revolution. The new Soviet Union was not a party to the peace negotiations or to the settlement. Instead a kind of quarantine was applied to the Soviet Union, and tacit support was given by the victorious powers to modest military intervention in Russia on the side of the anti-Bolshevik forces. This intervention was a total failure.

While the principle of self-determination for all was established, colonial and imperial issues were resolved through bilateral talks among imperial powers. Thus spheres of influence in the Middle East were divided between France and the UK, and contradictory promises were given to Arabs and Jews as to the future postcolonial disposition of the area.

In short, the Great War set the terms of the international order for the rest of the twentieth century (Boemeke et al. 1998). Domestically, the political consequences of war were mixed. Socialist and labor parties had participated in government. Right wing nationalist groups also grew. Among them were fascist parties, whose elevation of the virtues of armed struggle was mixed with a hatred of the political left. The center of politics weakened, evidenced by the waning of support for Liberal parties. The polarized left-right divide took on violent form in Italy, Germany, and throughout Eastern Europe, mixing ideological and territorial conflicts in such a way as to ensure the outbreak of a new and more lethal world war in just 20 years.

The war opened a phase of domestic strife in which the upheaval of war was mixed with revolutionary struggles. Civil war in Russia lasted for three years after the Armistice; armed conflict continued in the new states of Poland and Hungary. The victorious Allies quarantined the new Soviet Union, and began a phase of cold warfare which lasted for 70 years. For this reason alone, some historians see the war as inaugurating the 'short twentieth century,' from 1914-91 (Hobsbawm 1995).

3.3 Economic and Social

The Great War was both the apogee and the beginning of the end of European domination of the world economy. The big loser was the UK, whose export economy suffered from import substitution in countries throughout the world who developed finished goods during the conflict, when British exports were unavailable. Chronic financial instability weakened London as a banking center. The era of free trade was over (Milward 1984).

The losers paid for the war in a host of ways.

Reparations were forced on Germany, though she managed to avoid paying most of the bill. Worse was the continuation and acceleration of wartime inflation into the postwar years, wiping out savings and deepening domestic social and political conflict in Germany and Austria (Feldman 1993). This drastic reduction in the value of money had serious effects on the social structure and political outlook of middle class groups, threatened by proletarianization (Kocka 1984).

Domestically, economies of scale were introduced during the war which proved important after it. Hence military conflict helped accelerate the movement away from the small, family firm towards corporate economic life. State interest in and intervention in scientific research and education also expanded during and after war, in the light of evident strategic considerations (Hardach 1987).

After the war, the state continued to play a greater role in economic life than ever before. This was not through the ownership of industry, but through the management of debt. In most industrialized countries, taxation levels doubled over the war decade, and much of it was spent servicing war debt. The USA was the exception here, turning because of the war from a debtor to a creditor nation, though the chronic weakness of countries indebted to it after the war, like the UK, helped turn the Wall Street crash of 1929 into the world depression.

3.4 Cultural

Every combatant country mourned its war dead in public. The outcome is a vast array of war cemeteries near the fields of battle and of war memorials dotted in villages and towns throughout Europe, North America, Africa, Asia and the Antipodes. Annual commemorative ceremonies were (and still are) held on 11 November (Winter 1995) (Table 1).

The monumental character of the war has indelibly marked the literature, poetry, and visual arts of the twentieth century (Cork 1994, Fussell 1977, Hynes 1991). The Great War is the point of reference for much subsequent writing about catastrophe, and has introduced terminology—like shell shock—which has entered colloquial language. In many countries, though not in the former Soviet Union or the USA, its shadow is longer than that of the Second World War.

See also: Contemporary History; Imperialism, History of; Military History; Russian Revolution, The; Second World War, The; War: Anthropological Aspects; War: Causes and Patterns; Warfare in History

Bibliography

Afflerbach H 1994 *Falkenhayn: Politisches Denken und Handeln im Kaiserreich*. Oldenbourg, Munich, Germany

- Boemeke M F, Feldman G D, Glaser E (eds.) 1998 *The Treaty of Versailles: a Reassessment after 75 Years*. Cambridge University Press, Cambridge, UK
- Cork R 1994 *A Bitter Truth: Avant Garde Art and The Great War*. Yale University Press, New Haven, CT
- Deist W 1991 *Militär, Staat und Gesellschaft: Studien zur Preussisch—Deutschen Militärgeschichte*. R. Oldenbourg, Munich, Germany
- Feldman G D 1993 *The Great Disorder: Politics, Economics, and Society in the German Inflation, 1914–1924*. Oxford University Press, New York
- Fussell P 1977 *The Great War and Modern Memory*. Oxford University Press, London
- Hardach G 1987 *The First World War, 1914–1918*. Penguin, Harmondsworth, UK
- Hobsbawm E J 1995 *Age of Extremes: The Short Twentieth Century, 1914–91*. Abacus, London
- Horne A 1962 *The Price of Glory: Verdun 1916*. St. Martin's Press, New York
- Hynes S 1990 *A War Imagined: the First World War and English Culture*. Bodley Head, London
- Joll J 1992 *The Origins of the First World War*, 2nd edn. Longman, New York
- Jünger E (ed.) 1929 *Storm of Steel: From the Diary of a Storm Troop Officer on the Western Front*. Chatto and Windus, London
- Keegan J 1976 *The Face of Battle*. Cape, London
- Keegan J 1999 *The First World War*. A Knopf, New York
- Kennedy P M 1980 *The Rise of the Anglo—German Antagonism 1860–1914*. Allen and Unwin, London
- Kocka J 1984 *Facing Total War: German Society 1914–1918*. Berg, Leamington Spa, UK
- Milward A S 1984 *The Economic Effects of the Two World Wars on Britain*, 2nd edn. Macmillan, London
- Offer A 1991 *The First World War: An Agrarian Interpretation*. Clarendon Press, Oxford, UK
- Prior R, Wilson T 1992 *Command on the Western Front: the Military Career of Sir Henry Rawlinson, 1914–18*. Blackwell, Oxford, UK
- Prior R, Wilson T 1996 *Passchendaele: The Untold Story*. Yale University Press, New Haven, CT
- Ritter G 1958 *The Schlieffen Plan: Critique of a Myth* [Wilson A, Wilson E trans.]. Praeger, New York
- Showalter D E 1991 *Tannenberg, Clash of Empires*. Archon Books, Hamden, CT
- Smith L V 1994 *Between Mutiny and Obedience: The Case of the French Fifth Infantry Division during World War I*. Princeton University Press, Princeton, NJ
- van Creveld M 1977 *Supplying War*. Cambridge University Press, Cambridge, UK
- Winter J M 1985 *The Great War and the British People*. Macmillan, Basingstoke, UK
- Winter J M 1995 *Sites of Memory, Sites of Mourning: The Great War in European Cultural History*. Cambridge University Press, Cambridge, UK

J. Winter

Fisher, Irving (1867–1947)

Irving Fisher is widely recognized as one of the greatest American scientific economists. He has been credited

with being a primary architect of the ‘pillars and arches’ of modern economics—to borrow a phrase from J. A. Schumpeter—through his contributions to the theory of capital and interest, to monetary theory, and to the construction of index numbers. He was also an activist in the promotion of economic and social reforms. His reputation suffered in the 1930s when his approach to the discipline was eclipsed by Keynesianism. His accomplishments were then neglected and had to be rediscovered.

1. Beginnings

Irving Fisher was born on February 27, 1867 in Saugerties, New York and was the elder son of a Congregational clergyman. His was a home grown talent, nurtured in a single American institution: Yale University. Thus, unlike most American economists reaching professional maturity in the last decades of the nineteenth century, he had no direct exposure to the teachings of the German historical school. His intellectual development was shaped by two forceful members of Yale's faculty: Willard Gibbs, a mathematical physicist, and William Graham Sumner, a political economist who was also a champion of Social Darwinism. Their joint influence was observable in his choice of topic for a Ph.D. dissertation—*Mathematical Investigations in the Theory of Value and Prices*. With the publication of this exercise in general equilibrium theorizing in 1892, Fisher was hailed as a path breaker in mathematical economics. Fisher joined the Yale faculty in 1892 as an instructor in mathematics. In the following year, he married Margaret Hazard, a daughter of a wealthy Rhode Island textile manufacturer; their union was to produce two daughters and a son. In 1895, his appointment at Yale was shifted to the Department of Political Economy, an affiliation he retained until retirement in 1935. In 1898, he was promoted to full professor, but was soon diagnosed thereafter with a potentially fatal case of tuberculosis that idled him for the next three years. Upon recovery, he decided that the life of an ivory-tower academician was no longer sufficient to satisfy him. Indeed he was convinced that scientific professionals had an obligation to engage with the world by promoting economic and social uplift.

2. Contributions to the Theory of Capital, Income, and Interest

Fisher had undertaken preliminary inquiries into the nature of capital before the interruption occasioned by illness. He returned to this theme in 1904 when his energies approached full restoration. Results emerged quickly in two major treatises.

In *The Nature of Capital and Income* (1906), Fisher set out a rigorous formulation of key concepts that had formerly been clouded by confusion. Capital (including human capital) was identified as the stock of assets that yield services, whereas income referred to the flow of services that generated satisfactions. He further demonstrated that the value of capital was determined by discounting the future income stream. This finding marked a distinct analytic advance. Though now generally accepted, Fisher's insistence that asset values were oriented toward the future, rather than to historical cost, was controversial when first presented. One feature of the book's argument has continued to be controversial, i.e., his claim that the correct concept of income should identify it with consumption spending and should exclude saving. *The Rate of Interest* (1907) was a logical sequel to *The Nature of Capital and Income*.

The earlier book had established the pivotal role of the interest rate in capitalizing the value of an anticipated income stream. It thus remained to demonstrate how the rate of interest was itself determined. When Fisher first addressed this issue, debate was polarized between two rival positions: one explanation of interest focused on increments in output arising from the use of capital (the technological school); a second emphasized the necessity to reward savers for deferring consumption (the psychological school). Fisher's initial analysis was centered on 'impatience'—or an unwillingness to forgo today's consumption without the prospect of greater consumption possibilities tomorrow—as the principal explanatory factor. When he returned to this topic in *The Theory of Interest* (1930), he argued that the technological and the psychological arguments were complementary, that is, that the interest rate was determined by the interaction between 'impatience to spend income and the opportunity to invest it.' The real interest rate was thus unambiguously a real phenomenon, not a monetary one. This line of analysis would stand in sharp contrast to the monetary theory of interest that Keynes was to offer in his *General Theory of Employment, Interest and Money* (1936). For Fisher, money still mattered because of its impact on the price level. It could thus affect the nominal interest rate, which Fisher treated as the real interest rate plus the inflation rate.

3. Early Work in Monetary Theory

The Purchasing Power of Money (1911) was conceived as an exercise in establishing the validity and usefulness of the quantity theory of money, a doctrine that had been politically contaminated in the polemics over 'free silver' in the 1890s. In Fisher's formulation, 'the equation of exchange' was written as $MV + M'V' = PT$. (In this expression, M represented the quantity of cash and V the velocity of its circulation; M' stood

for total deposits subject to check and V' for the average velocity of their circulation; P was defined as an average price and T as the volume of trade.) This book was a vigorously monetarist document in which Fisher maintained that changes in the general price level were linked to proportionate changes in the money supply. This was true in the 'normal period,' defined as a state of equilibrium in which the values of T and the V 's were constant. His choice of terminology on this point bred misunderstandings: he recognized readily that economic reality was typified by 'transition periods'—when adjustments to disturbances were being worked out—in which the constancies did not hold.

Fisher was not content to offer solely an analytic explanation for variations in the purchasing power of money. Convinced as he was that monetary instability produced distributive injustices by distorting the relative positions of creditors and debtors, he felt obliged to suggest a corrective. His proposed remedy prescribed that changes in the general price level should be offset by variations in the gold content of the dollar (i.e., the 'compensated dollar' plan). But this policy recommendation did not mesh with the analysis presented in *The Purchasing Power of Money*. The 'compensated dollar' scheme rested essentially on a 'commodity theory of money' in which the real value of money was determined by the value of its gold equivalent. This conclusion did not follow from the 'equation of exchange' that Fisher had set out as the basis for his 'quantity theory of money.' Schumpeter (1948) and Patinkin (1993) have suggested that this oddity can be explained because Fisher, the reformer, got the better of Fisher, the scholar.

4. Analytic Innovations in the 1920s

Given Fisher's concern for stabilization of the general price level, his interest in perfecting the instrument for measuring its behavior—namely, the index number—was readily comprehensible. He addressed this in *The Making of Index Numbers* (1922), a work which has continued to be influential. Shortly thereafter, he created an Index Number Institute (INI) that provided weekly reports on leading economic indicators, along with commentary on current events. These analyses were distributed nationally via a syndicated news service. At its peak, Fisher's INI column had a potential to reach about seven million readers.

By the mid-1920s, Fisher's studies had convinced him that the stakes in price level instability were higher than he had earlier thought them to be. When writing *The Purchasing Power of Money*, he regarded distributive injustices as the major evil associated with unstable prices. His later analyses indicated that instability of the general price level was responsible for another major problem—fluctuation in the volume of

trade and, with it, variations in aggregate income and employment. This conclusion, he maintained, was supported by statistical evidence indicating that the rate of change in the general price level was highly correlated with variation in the volume of trade. This relationship, however, was not instantaneous.

His empirical investigations suggested a lag of approximately seven months between the turning point in the rate of change in prices and change in aggregate economic activity. In this research, Fisher introduced a novel statistical technique—the distributed lag—which has become a standard tool in econometrics. With its aid, his results appeared to show conclusively that a rising price level stimulated trade, while a falling price level depressed it. It thus followed that the way to eliminate macroeconomic fluctuations was to maintain a stable price level. In that spirit, the Fisher of the mid-1920s wrote that the ‘so-called business cycle’ was really ‘a dance of the dollar’ and that there was no reason why such disturbances should be tolerated.

5. *Crusading for Causes*

Fisher, the recovered TB patient, was an indefatigable champion of causes. This was evident in his tireless promotion of monetary reforms: in the first instance, in lobbying for a ‘compensated dollar’ that promised to stabilize the dollar’s purchasing power. In the 1920s, he lobbied vigorously for legislation that would compel the Federal Reserve to maintain price level stability. His views by then had changed on how this should be accomplished. In 1923, with the discovery of the potency of ‘open market operations’—a technique for altering commercial bank reserves through the purchase or sale of government securities by the Federal Reserve system—Fisher readily embraced this procedure as the major instrument for central bank control of the money supply. He was a frequent visitor to Washington as a star witness before Congressional committees with jurisdiction over monetary affairs. He also poured considerable energy into mobilizing public opinion behind reform proposals through the creation of organizations such as the Stable Money League.

Fisher’s reformist agenda was not restricted to matters directly linked to his expertise as an economist. He was an enthusiast for measures to improve the health and longevity of the population. On his initiative, a group of leading citizens was organized in 1907 to lobby—unsuccessfully as it turned out—for the creation of a Federal Department of Health that would establish national standards for public health practices and monitor their enforcement. He enjoyed greater success in the promotion of a Life Extension Institute to persuade life insurance companies that it would be profitable to require and pay for annual

physical examinations for policy holders. To reinforce its activities, he co-authored a first-of-its-kind manual of good health practices, entitled *How to Live*, that went through some 21 editions.

Fisher was one of the nation’s staunchest advocates of national prohibition of the consumption and sale of alcoholic beverages. Once the 18th amendment to the Constitution (1920) had been enacted for that purpose, he was a leading voice in arguing against its repeal. Even though enforcement of prohibition left much to be desired, he maintained that this legal ban had still contributed positively to American economic expansion in the ‘new era’ of the 1920s through its effect on improved worker efficiency and on reduced workplace absenteeism. This line of thinking was responsible, in part, for his disastrous misreading of economic conditions in the autumn of 1929 when he predicted that the American economy was on the threshold of permanent prosperity. In the 1920s, he was also an inveterate campaigner for an international organization to promote world peace and for membership of the US in the League of Nations. He was a prominent figure as well in the eugenics movement that he hoped would encourage state governments to enact legislation permitting sterilization of the mentally defective and the criminally insane.

Fisher contributed significantly from his own pocket to support these causes. His ability to do so was enhanced by the sale of one of his inventions—an Index Visible filing system—to Rand–Kardex (shortly to become Remington Rand) for a handsome block of stock and for a position on the board of directors. When informing his son about this turn of fortune, he wrote: ‘This added income will enable me to further the four chief causes which we have at heart, the abolition of war, disease, degeneracy, and instability of money (Fisher to Irving Norton Fisher, June 17, 1925).

6. *Coping with the Disaster of Depression in the 1930s*

Fisher was totally caught off-guard by the collapse of the stock market in the autumn of 1929 and by the events that followed. His personal financial circumstances were left in shambles: he managed to avert personal bankruptcy only by borrowing heavily from his wife’s unmarried sister. The Great Depression was also a professional embarrassment. It did stimulate him to rethink where his earlier line of analysis had failed him. The result—reported in *Booms and Depressions* (1932)—was the debt-deflation theory of Great Depressions. The problem, as he then saw it, stemmed from a situation of overindebtedness in the late 1920s. Once a break in the price level had occurred, the real burden of debt increased. This led to distress selling, to contraction in deposit currency, and to slackening in

the velocity of monetary circulation. Deflation—far from being self-correcting—simply bred more deflation. The remedy was implicit in the diagnosis: ‘reflation’ of the price level to its elevation at the time when most outstanding debts were contracted. When that point had been reached, the general price level should be stabilized.

To accomplish this, Fisher urged a number of proposals on the Roosevelt administration. For example, he endorsed ‘stamped scrip’—a scheme which called for a special currency issue that would lose 2 percent of its value each week unless a two-cent stamp were affixed to it. This device (which he had borrowed from Gesell) was intended to accelerate the velocity of monetary circulation. Though the idea was not original with him, he later put his full weight behind advocacy of a plan that would require commercial banks to hold ‘100 percent reserves’ against demand deposits, thereby eliminating their capacity to create (or destroy) check-book money by extending (or calling) loans. The money supply would instead be controlled by a public authority (e.g., a Currency Commission) that would expand the money supply as needed to achieve a prescribed ‘reflation’ target and thereafter to stabilize. Fisher’s conception of a ‘monetary rule’ was tied to outcomes, as opposed to the ‘rule’ recommended in Milton Friedman’s brand of monetarism that stipulates a fixed rate of growth in the money supply.

Fisher’s thinking about recovery from the Great Depression always had a monetary orientation. He was unsympathetic toward fiscal stimulants—such as Keynesian-style programs of public works—and he made no secret about that. He was also unsympathetic to the analytic message of Keynes’s *General Theory*. Even so, he and Keynes came close to convergence in their recommendations for tax policy in wartime: both championed a tax targeting consumption. The analytic props to their conclusions, however, had nothing in common. Fisher’s views—which he spelled out in his final book, *Constructive Income Taxation* (1942)—rested on the analysis he had worked out in 1906 in *The Nature of Capital and Income*. This argument maintained that income should properly be defined as expenditure on consumption: thus, a sound tax system, in all circumstances, should take aim at consumption spending. Keynes’s position, on the other hand, was informed by his conception of the management of the components of aggregate demand, as formulated in *The General Theory*.

Despite the fact that the Roosevelt administration took little heed of his unsolicited advice, Fisher still believed that something hopeful could be salvaged from the wreckage of depression in the form of lasting improvement in scientific investigation in economics. It had long been his aspiration to build an international organization to bring economists interested in linking economic theory with mathematics and statistics into closer collaboration. Circumstances of

the early 1930s permitted this dream to become a reality. The presence at Yale in 1930 of Ragnar Frisch of the University of Oslo made it easier for three who shared the dream—Fisher, Frisch, and C. F. Roos—to lay the groundwork for the Econometric Society. The obvious shortcomings of economic forecasting in the early depression years—and an apparent need for improvement—convinced a well-to-do businessman to underwrite this new organization.

7. A Word about the Legacy

Fisher was accorded numerous honors by professional colleagues: among them, election to the presidencies of the American Economic Association, the American Statistical Association, the Econometric Society, the American Association for Labor Legislation. But he lacked a committed cadre of disciples: there was no ‘Yale circus’ to match Keynes’s ‘Cambridge circus.’ In the last decade of his life, his approach was largely overshadowed by Keynesianism. To younger members of the profession, he appeared to be out-of-step with the times, with the consequence that his work tended to be ignored.

With the waning of the Keynesian ascendancy, Fisher’s achievements have been revisited. His pioneering analyses of capital and interest have now approached the status of an orthodoxy in the graduate training of prospective professionals. A ‘rediscovery’ in the 1970s documented that an article he wrote in 1926—entitled *A Statistical Relation between Unemployment and Price Changes*—anticipated the essential insight of the Phillips Curve by more than three decades. It is interesting to speculate on how the principals in an intense controversy that divided the profession in the 1960s and 1970s would have behaved had they known that Fisher had identified the ‘reswitching problem’ in 1907.

In the 1990s, some of Fisher’s practical proposals have enjoyed a renewed lease on life. In 1997, the US Treasury issued a ‘stabilized bond,’ the return on which would be adjusted to maintain constancy in its purchasing power. Fisher had advocated the creation of this type of financial instrument some seven decades earlier. In 1995, the US Department of Commerce resurrected another Fisher conception, dating from 1922, when adopting his formula for an ‘ideal index number’ to calculate the GNP deflator.

Fisher died on April 29, 1947 in New York City.

See also: Economics, History of; Health Economics; Keynes, John Maynard (1883–1946); Monetary Policy

Bibliography

Allen R L 1993 *Irving Fisher: A Biography*. Blackwell, Cambridge, MA

- Fisher I 1997 *The Works of Irving Fisher*, 14 Vols. Pickering and Chatto, London
- Patinkin D 1993 Irving Fisher and his compensated dollar plan. *Federal Reserve Bank of Richmond Economic Quarterly* 79: 1–33
- Schumpeter J A 1948 Irving Fisher's econometrics. *Econometrica* 16: 219–31

W. J. Barber

Fisher, Ronald A (1890–1962)

Sir Ronald Fisher, the father of modern statistics, was for most of his working life a professor of genetics, first in London and then at Cambridge. He made lasting contributions to mathematical and evolutionary genetics as well as to statistical theory. In statistics his contributions stand alongside those of his great predecessors Gauss and Laplace, while in evolutionary theory he is acknowledged as the greatest of Darwin's successors.

Fisher was born in London on February 17, 1890, the son of George Fisher, a fine-art auctioneer, and his wife Katie. His twin brother was stillborn. At Harrow School, which he entered in 1904 as a scholar, he distinguished himself in mathematics despite being handicapped by poor eyesight which prevented him working by artificial light. His teachers used to instruct him by ear, and Fisher developed a remarkable capacity for pursuing complex mathematical arguments in his head. This manifested itself later in life in an ability to reach a conclusion whilst forgetting the argument, to handle complex geometrical trains of thought, and to develop and report essentially mathematical arguments in English (only for students to have to reconstruct the mathematics later). Fisher's early interest in natural history was reflected in the books chosen for special school prizes at Harrow, culminating in his last year in the choice of the complete works of Charles Darwin in 13 volumes. Fisher entered Gonville and Caius College, Cambridge, as a scholar in 1909, graduating BA in mathematics in 1912. At college he instigated the formation of a Cambridge University Eugenics Society through which he met Major Leonard Darwin, Charles's fourth son and president of the Eugenics Education Society of London, who was to become his mentor and friend. He spent a postgraduate year in the Cavendish Laboratory, Cambridge, studying the theory of errors under F. J. M. Stratton and statistical mechanics and quantum theory under J. H. Jeans (later Sir James Jeans).

Prevented from entering war service in 1914 by his poor eyesight, Fisher taught physics and mathematics in schools for the duration of the war and in 1919 was appointed Statistician to Rothamsted Experimental

Station, an agricultural station at Harpenden north of London. In 1933 he was elected to succeed Karl Pearson as Galton Professor of Eugenics (i.e., of Human Genetics, as it later became) at University College, London, and in 1943 he was elected Arthur Balfour Professor of Genetics at Cambridge and a Fellow of Gonville and Caius College, succeeding R. C. Punnett. After he retired in 1957, Fisher traveled widely, spending his last few years in Adelaide, Australia, as an honorary research fellow of the C. S. I. R. O. Division of Mathematical Statistics. He died there of a postoperative embolism on July 29, 1962. His ashes lie under a plaque in a side aisle of Adelaide Cathedral.

Fisher married Ruth Eileen Guinness in 1917 and they had two sons and six daughters, and a baby girl who died young. He was elected a Fellow of the Royal Society of London in 1929 (as a mathematician) and was created Knight Bachelor by Queen Elizabeth II in 1952 for services to science. He was the founding President of the Biometric Society (now the International Biometric Society) in 1947, and served as President of the Royal Statistical Society (of the UK), of the International Statistical Institute, of the Genetical Society of Great Britain, and of his Cambridge college, Gonville and Caius. He received many honorary degrees and accepted the honorary membership of many academies at home and abroad, and was awarded all the principal medals of the Royal Society, the Royal (1938), Darwin (1948), and Copley (1956).

Fisher's first publication, whilst still an undergraduate, proposed the method of maximum likelihood for statistical estimation (though not in those words) and before graduating he had already given a paper remarking on the surprisingly good fit of Mendel's data. In 1916, without any academic appointment but encouraged by Leonard Darwin, he completed the founding paper of biometrical genetics 'The correlation between relatives on the supposition of Mendelian inheritance' (1918), eventually published in 1918 after it had been rejected for the *Philosophical Transactions of the Royal Society* on the advice of the referees, Karl Pearson and R. C. Punnett (both of whom he later succeeded, as he was fond of remarking). In this paper, Fisher introduced the word 'variance' and established the basis of the analysis of variance.

From his post of statistician at Rothamsted, Fisher made advances which revolutionized statistics, whilst his advances in genetics and evolution were hardly less pathbreaking. In a single publication in 1922 he proved that heterozygotic advantage gives rise to a stable gene-frequency equilibrium, introduced the first stochastic model into genetics (a branching process), and initiated the study of gene-frequency distributions by means of the diffusion approximation, and in another paper he applied the method of maximum likelihood for the first time, to the estimation of linkage.

Also in 1922, building on his method of maximum likelihood, Fisher published 'On the mathematical

foundations of theoretical statistics' in the *Philosophical Transactions* (1922a), a revolution in statistical thought. This time the Royal Society's referees, A. S. Eddington and G. U. Yule, were more complimentary. Amongst the new words and phrases Fisher introduced in the paper in order to clarify the basis of his approaches to statistical inference are likelihood, consistency, degrees of freedom, efficiency, sufficiency, information, significance level, test of significance, sampling distribution, percentage point, statistic, and others of a more technical nature still. Information in Fisher's statistical sense (1925) antedated Shannon's (1948) use of the word for a somewhat different purpose: Fisher information measures the content of a message, whereas Shannon's measures the ability of a channel to transmit a message, independently of its meaning. In 1930, Fisher published the first satisfactory account of what J. Jerzy Neyman was later to call the 'confidence' property of interval estimation.

In distribution theory Fisher built on his 1915 paper deriving the exact distribution of the correlation coefficient (for whose solution he introduced the concept of a sample-space) and used his considerable mathematical ability to solve many further distributional problems, including that of the multiple correlation coefficient and the variance ratio (which G. W. Snedecor of Iowa State University, where Fisher was visiting professor in 1931 and 1936, called 'F' in his honor). His work at Rothamsted on agricultural statistics soon led to the creation of the science of experimental design with the appreciation of the fundamental role of randomization, the development of the principles of blocking and confounding, and the introduction of Latin squares and factorial analysis and similar designs chosen for their statistical efficiency. The methods which Fisher invented or improved were brought together in 1925 in his book *Statistical Methods for Research Workers* (1925b), which he updated continually over the next 23 years until the thirteenth edition was published in 1958 (a posthumous fourteenth edition was prepared from notes written before his death).

In addition to the statistical work associated with his employment at Rothamsted, Fisher pursued his interest in evolutionary theory, guided and encouraged by Leonard Darwin. Putting his deep knowledge of Mendelian theory and Charles Darwin's writings to good use he built on the success of his earlier investigations of 1918 and 1922 to write *The Genetical Theory of Natural Selection* (1930), published in 1930. This book is now seen as the chief document of the 'Neo-Darwinian revolution,' that revival of Darwinism in the 1930s and 1940s from the neglect it had suffered in the early years of the century following the rediscovery of Mendel's paper in 1900. The book contains a wealth of new evolutionary arguments, from the fundamental theorem of natural selection to ideas about sexual selection, inclusive fitness and parental expenditure. More than any other work *The*

Genetical Theory established a firm basis for the modern view that evolution by natural selection is primarily a within-species phenomenon.

Taking up his appointment as Galton Professor of Eugenics at University College London in 1933, Fisher's pace did not slacken. Now able to pursue experimental work, which previously he had only undertaken in a private capacity, he demonstrated his ability as a natural historian. Experimental organisms included mice, poultry, and the purple loosestrife, and even dogs (under the auspices of the Genetical Society). In a long collaboration with E. B. Ford of Oxford he studied natural populations for the purpose of estimating the power of natural selection, thus contributing to the foundation of a school of ecological genetics. But it is in human genetics that he made his most lasting contribution of this period. In 1935 he secured funds from the Rockefeller Foundation to establish a Blood-Group Serum Unit at the Galton Laboratory with the express purpose of initiating the construction of a linkage map of man, for he had already seen the connection between 'Linkage studies and the prognosis of hereditary ailments' (to use the title of his lecture to the International Congress on Life Assurance Medicine in that year). Here is to be found the intellectual origin of the Human Genome Project. At the same time Fisher, with J. B. S. Haldane and L. S. Penrose, was advancing the special statistical theory required in the estimation of human linkage.

Fisher (1935) also saw the publication of *The Design of Experiments* in which his pioneering work in that subject at Rothamsted was expounded fully for the benefit of workers in agriculture and many other fields. Like *Statistical Methods for Research Workers*, the book was an outstanding success, running to seven editions by 1960 (an eighth was later prepared from notes left by Fisher). It opened with a famous description, often quoted, of an experimental design for the problem of testing the assertion of a lady who maintained that she could tell by tasting a cup of tea whether the milk or the tea had been the first to be poured into the cup. In the course of this discussion Fisher coined the phrase 'null hypothesis' for the hypothesis under test.

In 1936, Fisher subjected Mendel's published data on segregation in the garden pea to close statistical scrutiny, concluding that in some respects the data seemed to follow too closely the expectations on Mendel's genetic hypothesis, a discovery that worried him greatly, and left him speculating that perhaps Mendel's assistant, knowing full well what his master hoped to find, was tempted to adjust some of the counts accordingly.

In 1943, Fisher moved to Cambridge as Arthur Balfour Professor of Genetics, where he was reunited with his colleagues from the Blood-Group Unit which had been evacuated there during the war. An immediate consequence was his brilliant solution of the Rhesus blood-group puzzle, involving three closely

linked loci which between them explained the array of serological reactions which to everyone else had appeared chaotic: Fisher did for Rhesus what Mendel did for round and wrinkled.

After the second world war ended in 1945 Fisher attempted to establish bacterial genetics in his Cambridge department and to retain for Cambridge the Blood-Group Unit, but without success. Work in his small department revolved around linkage in the mouse and studies on purple loosestrife, wood sorrel, and primroses, always with a strong background of mathematics and statistics. His *Theory of Inbreeding* (1949) was published in 1949, and in 1950 he published the first paper applying a computer to a biological problem.

In the subject of statistics, developments in the USA during and after the war had tended to obscure the great advances which Fisher had made in the subject, and he found himself having to defend many of his views against critics brought up in the more mathematical and operational climate then dominant. His repeated insistence that statistical inference was a mode of reasoning separate from, though related to, approaches based on decision theory and other behavioral considerations won him few friends, and his continued (if somewhat puzzled) defense of his troublesome 1930 concept of fiducial probability was cited as evidence that his judgment was fallible. He defended his position in a fourth book, *Statistical Methods and Scientific Inference* (1956), published in 1956 with a second edition in 1959 and a posthumous edition with many additions and clarifications. In this work he particularly emphasized the direct role of likelihood in statistical inference, in addition to the method of maximum likelihood, and continued his life-long criticism of the improper use, as he saw it, of prior probabilities and Bayes's theorem in statistical inference, a methodology previously called 'inverse probability' for which he had himself coined the word 'Bayesian' in 1950.

As well as writing his books, Fisher collaborated with Frank Yates, also of Rothamsted, to produce the most widely-used compilation of statistical tables (before computers rendered such volumes redundant). First published in 1938, *Statistical Tables for Biological, Agricultural and Medical Research* (1938) went through many editions.

In retirement Fisher continued his academic work, but also became embroiled in a controversy as to the extent to which the published data supported the hypothesis, then novel, that cigarette-smoking was a cause of lung cancer. Perhaps not fully appreciating the strong feelings which the issue was to generate, he focused on the finer points of the statistical analysis which interested him, on the need always to disentangle causation and correlation, and on the fact that the evidence did not seem to him to disallow the possibility that genetical variability might be a contributory factor in the observed association.

Fisher's papers number nearly 300, and he also wrote many reviews, particularly in *The Eugenics Review* between 1915 and 1935, and letters to journals. He was an accomplished formal lecturer as may be seen from his many presidential and similar addresses, and an occasional broadcaster on scientific topics. His interests were wide, as may be instanced from two examples from his publications: in 1934 he discovered and published the idea of a randomized or 'mixed' strategy independently of von Neumann, and in 1958 he published the first suggestion that the theory of games might be applied profitably to problems in evolutionary biology, the possibility having been drawn to his attention by Dr. Cavalli (later Professor L. L. Cavalli-Sforza).

Small of stature, with thick glasses and a beard, Fisher did not suffer fools gladly. He was a skilled controversialist in conversation, but his quick temper sometimes rendered further discussion impossible. His contemporaries divided rather cleanly into those who regarded him with awe and affection and gratitude for the generosity with which he offered his ideas to them, and those who found him tetchy, difficult, and remote. Especially as a professor at Cambridge, he showed great interest in the few students who passed through his small department, and he always enjoyed the company of young people. Of the many thumb-nail sketches which his greatness has inspired, perhaps the following comment on Fisher by the Cambridge cosmologist Sir Fred Hoyle contains the closest likeness in the smallest span: 'I am genuinely sorry for scientists of the younger generation who never knew Fisher personally. So long as you avoided a handful of subjects like inverse probability that would turn Fisher in the briefest possible moment from extreme urbanity into a boiling cauldron of wrath, you got by with little worse than a thick head from the port which he, like the Cambridge mathematician J. E. Littlewood, loved to drink in the evening. And on the credit side you gained a cherished memory of English spoken in a Shakespearean style and delivered in the manner of a Spanish grandee' (Hoyle 1999).

Fisher was one of the great intellects of the twentieth century. His contributions to evolutionary theory and mathematical genetics ensure him a place in the Darwinian pantheon whilst he is the undisputed father of modern statistical theory.

The statistical methodology he developed has had a far-reaching impact in every field of science and social science. Whether one considers the mathematical theory of distributions, the design and analysis of experiments, the theory of estimation, the theory of significance testing, or the logic of statistical inference, Fisher's is the primary influence. As a teacher of research workers through his statistical books he had no peer.

In genetics he was the principal architect of the application of statistics, much of his theory having been inspired by practical problems, especially in

human genetics. With J. B. S. Haldane and others he initiated the study of linkage in man specifically for the purpose of throwing light on hereditary disease risks, thus foreshadowing the human genome project, and he encouraged the study of associations between blood groups and disease. His studies on natural and artificial populations of both plants and animals would alone have given him an honored place in the development of genetics in the twentieth century. His '1918 paper' is accepted as the starting point for the entire development of biometrical genetics, with its ramifications in plant and animal breeding.

In evolutionary biology, his book *The Genetical Theory of Natural Selection* (1930) hides a brilliant series of evolutionary insights behind a density of writing that requires more concentration than most readers have been prepared to give it, but is now acclaimed as the cornerstone of the neo-Darwinian revival of the mid-twentieth century.

See also: Distributions, Statistical: Approximations; Estimation: Point and Interval; Hotelling, Harold (1895–1973); Natural Selection; Significance, Tests of; Statistical Methods, History of: Post-1900

Bibliography

- Bennett J H (ed.) 1971–74 *Collected Papers of R. A. Fisher*, 5 Vols. University of Adelaide, Adelaide, Australia
- Bennett J H (ed.) 1983 *Natural Selection, Heredity, and Eugenics, Including Selected Correspondence of R. A. Fisher with Leonard Darwin and Others*. Clarendon Press, Oxford, UK
- Bennett J H (ed.) 1990 *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher*. Clarendon Press, Oxford, UK
- Box J F 1978 *R. A. Fisher: The Life of a Scientist*. Wiley, New York
- Fisher R A 1918 The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**: 399–433
- Fisher R A 1922a On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A* **222**: 309–68
- Fisher R A 1922b On the dominance ratio. *Proceedings of the Royal Society of Edinburgh* **42**: 321–41
- Fisher R A 1925a Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society* **22**: 700–25
- Fisher R A 1925b *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh
- Fisher R A 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, UK
- Fisher R A 1935 *The Design of Experiments*. Oliver and Boyd, Edinburgh
- Fisher R A 1949 *The Theory of Inbreeding*. Oliver and Boyd, Edinburgh
- Fisher R A 1950 *Contributions to Mathematical Statistics*. Wiley, New York
- Fisher R A 1956 *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh
- Fisher R A, Yates F 1938 *Statistical Tables for Biological, Agricultural and Medical Research*. Oliver and Boyd, Edinburgh

Hoyle F 1999 *Mathematics of Evolution*. Acorn Enterprises, Memphis, TN

A. W. F. Edwards

Fishing Societies

The theme of this article is the place of fishing activities and coastal economies in social scientific thought and some of the major issues developed in the relevant literature; the category of fishing, natural and social models of fishing activities and coastal communities, gender relations, ecological knowledge, and resource management.

1. The Category of 'Fishing'

According to some of the evolutionary theories of the nineteenth century, fishing represented a separate and early stage in the history of humans. Such an idea was underlined in many contemporary accounts of particular groups of people largely dependent on fish. The evidence on the role of early fishing, however, is fairly limited. Some scholars suggest that, from the point of view of early humans, aquatic environments must have been both hostile and an inadequate source of food. Others suggest that coastal zones must have been quite attractive, providing a worthwhile challenge, and that people may have exploited them for a long time without leaving modern groundlings any evidence. While the archaeological evidence does not indicate that fishing predated the hunting of terrestrial animals, nonetheless among hunter-gatherers fishing may have played quite an important role. The abundance of resources in coastal zones may, indeed, have provided an opportunity for the development of complex civilizations. There is evidence for a transformation of hunter-gatherer social relations in coastal regimes although many of the important issues involved are far from settled (see *Hunting and Gathering Societies in Anthropology*). Rather than seeing aquatic resources themselves as determinants of complexity, however, one should regard coastal niches as just one possible avenue for intensification (Kelly 1992). With the advent of fieldwork and participant observation, evolutionary speculations tended to be replaced by descriptive accounts. The category of 'fishing' became much less visible than before—sometimes being subsumed under the label 'hunting and gathering.'

While the category of fishing may seem straightforward, it need not be taken for granted. The distinctiveness of fishing activities, it has been argued (Hewes 1948), has two aspects. First, objects behave in a particular manner while in an aquatic substance, due to special conditions of buoyancy, turbulence, solubility, and refraction of light. Second, fishers and their prey occupy different media. An emphasis on differ-

ences in habitat yields a definition of fishing as 'that category of human activity which is connected with the capture or gathering, of animals (or plants) which regularly dwell in the water' (Hewes 1948). The concept of fishing as a particular kind of hunting which happens to yield fish is one element of a common anthropological scheme for classifying types of subsistence: gathering, collecting, hunting (including trapping), husbandry (including fish farming), and plant cultivation. Much like medieval Europeans often distinguished between fishing, fowling, and hunting, modern anthropology tends to operate with three concepts of foraging—fishing, gathering, and hunting. For some scholars, fishing is a broad category that includes shell fishing and the pursuit of large aquatic animals. Others have argued to the contrary that the pursuit of large aquatic animals more properly is classified as hunting and that shell fishing should be classified as gathering.

While some linguistic evidence suggests that fishing and hunting are perceived as similar strategies, differing only with respect to the focus of subsistence efforts, indigenous terminologies do not distinguish necessarily between hunting and other subsistence activities, including trapping, collecting, and gathering. Also, the classification of aquatic organisms, including 'fish,' varies from one society to another. Although in many societies, fish are classified according to an elaborate taxonomy, depending on their shape and usefulness to humans, the category of 'fish' represents one of the largest and most diverse categories found in most environments; at times fish terms are extended to all creatures inhabiting aquatic environments, including turtles, frogs, crocodiles, crustacea, and fish-shaped mammals. In some languages, there is no term for fish; fish are included in a heterogeneous category along with snakes, worms, lizards, and other elongated creatures.

2. Natural and Social Models

Many approaches to fisheries, both of the present and the past, suggest a 'natural' model of fishing, emphasizing material context and ecological relations. Social scientists have often drawn attention to the fact that fishing takes place in a relatively uncertain environment (see Pálsson 1991). People who adapt to earning a living by exploiting marine resources, it is argued, often seem to manage their lives in similar ways. Thus, crew organization tends to be flexible and based more on voluntary ties than kinship obligations, to ensure cooperation and the right combination of skills, to reduce uncertainty, and spread the risks of production. Fishing societies, as a result, seem to have a range of characteristics in common due to the fact that their members have to adapt to corresponding environments and cope with identical problems. Such approaches are reminiscent of the method of cultural

ecology, according to which social life is rather mechanistically adapted to the material world.

In line with the assumptions of the natural model of fishing, there has been a strong tendency to look for parallels between trawling ('industrial hunting') and small-scale fishing on the basis of a common link to water. While focusing on the boundary between land and water is helpful for drawing contrasts between economic and social systems which are organized in similar ways—between, say, aquatic and terrestrial industrial systems—much recent research has emphasized that a unitary category of fishing diverts attention from important issues, in particular the social relations in which production is necessarily to be found. Although there are real empirical differences between fishing and other modes of subsistence, the use of such differences gives misplaced importance to ecology.

One way to understand similarities and differences among fishing systems, to emphasize the social context of production, is to distinguish between societies in terms of mode of circulation—the motivation of the producers and the destination of the products (see *Mode of Production*). Production may be primarily for exchange or primarily for use. Many fisheries conflicts derive from the different rationales of production of the household economy of local small-scale fishing and the market economy of industrial fleets (McGoodwin 1990). Another way to distinguish fishing systems is to underline different forms of managing access to fish. While knowledge of ecology and fishing techniques is important for the understanding of different forms of managing access, one should not ignore the social space in which they occur. The informal exclusion of outsiders by means of secrecy and formal ownership of local territories should not be regarded as functionally equivalent devices. Differences among fishing economies in terms of modes of circulation of products and access to resources insufficiently represent the variety of production systems there is. On the other hand, they help to illustrate the fundamental point that fisheries are embedded in social life. Any study of what happens on board fishing vessels must consider variations in fishing systems and the larger contexts of fisheries and fishing communities.

3. Gender Roles and Identities

In removing 'fishing' from the context of social relations, social scientists have often failed to recognize the importance of the relations of men and women and the significant economic role of women in fishing economies. The issue of gender tends to be suppressed or distorted. Nevertheless, in recent years an extensive literature on the role of gender in fishing and coastal communities has been developed (see, for instance, Nadel-Klein and Davis 1988). In fishing communities, it has sometimes been suggested, one can expect a

sharp division of labor along gender lines in that men do the fishing while women are responsible for the processing of the catch. A rigid sexual division of labor should not, however, be regarded as inevitable. In many cases women's participation in extraction is no less important than that of men. The ethnography of fishing is biased with respect to gender. This observer's bias is partly a reflection of the popular Western image of fishing as a particularly 'manly' activity. Thus, the pursuit of small aquatic animals has sometimes been regarded as mere 'gathering' or 'collecting' if performed by women, but as 'fishing' if performed by men.

Many indigenous accounts of fishing economies are no less gender-specific than observers' accounts, treating women's labor as not only different but also somehow secondary to that of men. While the devaluation of women's work is not restricted to market economies, the change from a subsistence economy to markets often has important implications for the social construction of gender. The 'domestic' labor of women tends to be presented as private labor, on the grounds that it takes place internal to the household and that its consumption products are 'perishable' use values, requiring immediate use. Conversely, only the non-domestic work of men is regarded as socialized since it alone involves cooperation between households, the production of exchange values and durable commodities. Domestic laborers are presented as natural beings outside the social life of the market economy. And women only become social or visible insiders to the extent that they engage in nondomestic production, outside the sphere of reproduction. Women, then, tend to be presented as unproductive (see Cole 1991). But just as extraction is a social activity whether or not it takes place in a group or by isolated individuals, labor is necessarily a social activity whether or not it takes place in a domestic or non-domestic context. The household is not a private place devoid of social content but a social institution constituted by a particular set of relations—the social relations of domestic production involved in the creation of use values, particularly cooking and child care (see *Gender and Feminist Studies in Anthropology*).

4. Ecological Knowledge

Since fish are relatively incapable of coping with human predators, their chances of using evasive tactics are small. The difference in medium, however, has contradictory implications and some observers have emphasized the disadvantages to fishers, pointing out that if humans are invisible to the fish so are fish to humans. Accordingly, it is argued, knowledge of marine organisms is likely to be relatively restricted. Nevertheless, as several ethnographic accounts have shown, in fishing communities people accumulate a

vast amount of knowledge of fish and aquatic habitat. The social distribution and nature of ecological knowledge and its importance for resource management, on the other hand, are matters of some debate.

In fisheries generally, catches vary from one boat to another. Accounts of and interest in such differences vary from one context to another, depending on the nature of the social system involved and the nature of the fishery. Some explanations emphasize differences in luck, others underline personal differences in fishing skill. Some anthropologists have attempted to statistically measure the personal contribution of the leaders of fishing operations, the 'skipper effect' (see Pálsson and Durrenberger 1990). It is difficult, however, to separate the factors involved and recent studies of fishing success have drawn attention to the continuity of the social world to which the skipper belongs, focusing on the importance of the crew for decision-making and, by extension, relative fishing success. In a very real sense, the practical skills of fishing skippers are the result of mutual enskillment, the collective property of a community of practice (Pálsson 1994). Not only has there been a growing theoretical interest in the practical learning of fishers and the general divide between indigenous and scientific knowledge, but fisheries science itself is increasingly seen as an embedded enterprise (Smith 1996). Some scholars argue that learning, whether in science or every-day life, primarily takes place with the internalization of a cognitive scheme (see *Cognitive Anthropology*). Others have emphasized that learning involves the whole person interacting with the social and natural environment (Lave 1990; see *Apprenticeship: Anthropological Aspects*).

In many fishing societies there is little attempt to draw upon the knowledge that practitioners have achieved in the course of their work, for the purpose of marine ecological research. Analyses of human-environmental relations are the privileged task of people formally trained in public institutions. There are important signs of change, however. A greater reliance on practical knowledge, it is sometimes argued, is a practical necessity, given the randomness and complexity of many marine ecosystems. Multispecies marine ecosystems, in particular, are highly unpredictable, with constant fluctuations in interactions among species and between species and their habitat (Acheson and Wilson 1996). This does not mean that governance is impossible; it suggests, however, increasing reliance on a finer spatial and temporal scale of observation and engagement—a scale that only the skilful practitioner is able to apply. There are good grounds, therefore, for exploring how fishers acquire knowledge about the ecosystem within which they operate, how their knowledge differs from that of professional biologists, and to what extent the former could be brought more systematically into the process of management for the purpose of ensuring sustainability.

5. Resource Management

Many fisheries are currently in a state of crisis. Increasingly social scientists are addressing the problems involved, both internationally and in particular ethnographic contexts (McCay and Acheson 1987). Two main positions can be identified, the market approach which gained momentum in the early 1970s and the approach of co-management which has received increased attention since the 1980s. Comparative studies of the implications and consequences of different management regimes represent an important and growing forum for social research.

The market approach assumes that the current crisis in fishing is due to the open-access nature of fisheries; in the absence of both government control and the institution of private property, it is assumed, people will inevitably drive the ecosystem beyond its sustainable limits. The paradigm further suggests that only the 'hidden' forces of the market can ensure efficiency and sustainable use of resources. Finally, the issues of ethics and equity typically are presented as irrelevant theoretical distractions. In several fisheries in different parts of the world, fishing stocks are being turned into private property. First, the resource is appropriated by regional or national authorities and later on the total allowable catch (TAC) for a season is divided into quota holdings and allocated to producers, often the owners of boats. At a still later stage, the temporary privileges of quota holders are turned into a marketable commodity, individual transferable quotas (ITQs).

In some fisheries the crew's contribution and its relevance for quota allocation has been a pressing political issue. Since fishing quotas are often allocated with reference to 'fishing history,' fishers, boat owners, and managers debate on what exactly makes fishing history, focusing on the relative importance of the vessel itself and its technology (favoring the owners of capital) and the crew (favoring labor power, in particular that of the skipper). A related contested issue involves the practical and theoretical merits of effort quotas (number of fishing trips) vis-à-vis catch quotas (size of catch).

The other main approach to fisheries management, co-management (Jentoft 1989), emphasizes democratic decision making and the importance of practical knowledge. A number of anthropological studies have shown that in many fishing societies people have developed indigenous means of regulating access to fishing grounds. In some cases, local groups of users successfully control the reproduction of renewable resources without external intervention, effectively co-managing local resources. In some parts of the world, especially Asia and the Pacific, co-management has a long tradition (Ruddle and Johannes 1985). Recently, such examples have attracted much attention among both researchers and managers, informed by renewed interest in 'indigenous' or 'traditional' ecological

knowledge on the international scene (Dyer and McGoodwin 1994) as well as growing doubts about the modernist concept of 'resource management,' with its assumptions of human mastery, control, and predictability.

Traditionally, anthropologists have focused on isolated communities. Modern anthropology, however, provides useful perspectives on cultural flows and the link between micro and macro levels. Recently, anthropological students of fishing societies have increasingly moved from the narrow context of face-to-face encounters and local communities to the nation state and beyond (LiPuma and Meltzoff 1994). Thus, several anthropological studies have drawn attention to the ways in which local and professional identities are constructed within larger contexts.

Bibliography

- Acheson J M, Wilson J A 1996 Order out of chaos: the case for parametric fisheries management. *American Anthropologist* **98**: 579–94
- Cole S 1991 *Women of the Praia: Work and Lives in a Portuguese Coastal Community*. Princeton University Press, Princeton, NJ
- Dyer C L, McGoodwin J R (eds.) 1994 *Folk Management in the World's Fisheries: Lessons for Modern Fisheries Management*. University Press of Colorado, Niwot, CO
- Hewes G 1948 The rubric 'fishing and fisheries.' *American Anthropologist* **50**: 238–46
- Jentoft S 1989 Fisheries co-management: delegating government responsibility to fishermen's organizations. *Marine Policy* **13**(2): 137–54
- Kelly R L 1992 Mobility/sedentism: concepts, archaeological measures, and effects. *Annual Reviews of Anthropology* **21**: 43–66
- Lave J 1990 The culture of acquisition and the practice of understanding. In: Stigler J W, Shweder R A, Herdt G (eds.) *Cultural Psychology: Essays on Comparative Human Development*. Cambridge University Press, Cambridge, UK
- LiPuma E, Meltzoff S K 1994 Economic mediation and the power of associations: toward a concept of encompassment. *American Anthropologist* **96**(1): 31–51
- McCay B M, Acheson J M (eds.) 1987 *The Question of the Commons: The Culture and Ecology of Communal Resources*. University of Arizona Press, Tucson, AZ
- McGoodwin J R 1990 *Crisis in the World's Fisheries: People, Problems, and Policies*. Stanford University Press, Stanford, CA
- Nadel-Klein J, Davis D L (eds.) 1988 *To Work and to Weep: Women in Fishing Economies*. Institute of Social and Economic Research, St. John's, Newfoundland
- Pálsson G 1991 *Coastal Economies, Cultural Accounts: Human Ecology and Icelandic Discourse*. Manchester University Press, Manchester, UK
- Pálsson G 1994 Enskilment at sea. *Man: The Journal of the Royal Anthropological Institute* **29**(4): 901–27
- Pálsson G, Durrenberger E P 1990 Systems of production and social discourse: the skipper effect revisited. *American Anthropologist* **92**: 130–41

- Ruddle K, Johannes R E 1985 *The Traditional Ecological Knowledge and Management of Coastal Systems in Asia and the Pacific*. UNESCO, Jakarta, Indonesia
- Smith M E 1996 Public policy, sciencing, and managing the future. In: Nader L (ed.) *Naked Science: Anthropological Inquiry Into Boundaries, Power, and Knowledge*. Routledge, London

G. Pálsson

Folk, Indigenous, and Customary Law

The term 'customary law' in its most general meaning refers to unwritten legal systems. Though its roots are much older, the theoretical foundations were laid in the eighteenth century by the German historical school of legal theory, which was based on the assumption that all law originated from customs. There, it obtained the meaning of law that does not emanate from the state. The term became fashionable and of eminent practical importance in the nineteenth century, when colonial powers had to determine their relationship to existing legal systems in their colonies. These systems were in the majority of cases unwritten and at best partly written down. With the political demands for self-determination of indigenous peoples, customary law, as one of the fundamental features of their culture and claims for sovereignty, has received renewed interest and political importance. This article will discuss the various terms used for unwritten law and the advantages and disadvantages of each of them. It will then present an overview of the political controversies around the recognition of customary law. Then follows a discussion of the historical developments in the research themes concerning customary law. Finally the debate about the question whether customary law is an invention by colonial powers will be presented.

1. Customary Law, Traditional Law, Indigenous Law, Folk Law, Local Law

The terms 'traditional law,' 'customary law,' 'indigenous law,' 'folk law,' and—for Indonesia—'*adat* law' are often used interchangeably. Each one carries its own connotations and has its specific advantages and disadvantages (van den Bergh 1986). The problem with terms like customary and traditional is that they refer to a presumed unchanged past, though in actual fact each kind of law is subject to continuous change. Sometimes and in some places changes occur rapidly, and at other times law changes slowly and gradually. But there is no legal system that does not change at all. To presume too much continuity poses analytical as

well as political problems. It leads too easily to a fixed and traditionalist representation of local law which has little to do with the social reality of such law. However, when people do not want to comply with proposals or orders from the government, often one of the few legitimate and accepted arguments to oppose government interference is to say government plans go against their traditional law. The real reasons for opposition may be quite different and have to do with distrust of corrupt government officials or differences as to the economic soundness of the proposed plans. This very argument strengthens the perception of government officials that traditional law is indeed an obstacle for development.

The struggle for self-determination of indigenous peoples also suffers from traditionalization of indigenous law. Because reference to long traditions strengthens the claim to self-determination, attention is called away from the flexibility and adjustability of such legal systems. The problem is complicated by the fact that usually there is more than one version of 'traditional' law; the law as it is used and maintained by the local communities themselves, and traditional law as it is interpreted by government institutions.

The Dutch equivalent for indigenous, *inheems*, has had strong colonial connotations and for that reason went out of use. However, in connection with the new discussions about the position of indigenous peoples, it has come into use again. The term 'folk law' (Dutch and German: *Volksrecht*) has been suggested, and in several European languages is a more neutral alternative. In English, it resembles the term 'folklore' too much and is therefore not a good alternative. For these reasons it has been proposed to use the neutral term 'local law' as a generic term for law that is being used and maintained at a local level, from whatever source it is derived.

2. Political Controversies Surrounding the Recognition of Customary Law

Until the nineteenth century colonial governments throughout the world had been mainly interested in controlling trade centers. However, with the change from mercantile to production colonies, colonial administrations had to deal increasingly with local populations at large. In their attempts to gain control they had to determine their relationship towards the existing political and legal systems. In general, they refrained from much interference with family law, but other sectors of law—such as criminal law and land law—pertaining to the political and economic position of colonists and their economic relationships towards the local populations could not simply be left to customary law. Different colonial powers developed very different policies. One extreme was France, with its strong centralistic tradition; it reduced recognition

of existing legal systems as much as it could. England represented the other extreme, with its indirect rule and its concomitant wide recognition of customary law. The Netherlands took a middle position. As in all other colonial countries heated debates took place in The Netherlands about the status and scope of influence of what was called '*adat-law*' of the Dutch East Indies. Issues of both public administration, and material, economic interests were involved in these debates, which touched upon control over large tracts of land. At an abstract level, the question concerned whether colonial law was the exclusive guarantee for economic development, or whether customary law could also support economic growth and development. Ultimately, it was also a debate about the legitimacy of the colonial legal order, the colonial administration, and its monopoly on the legitimate use of force. In a more concrete way, the struggle was about the extent and legitimacy of colonial control over the most important economic assets, i.e., land. See Hooker (1975) for an overview of the legal policies of various colonial powers.

After independence, new states had to reconsider the form and constitution of their legal order. Most countries opted for continuation of the policy of their former colonial motherland and simply took over the former colonial law, including its policy concerning recognition of customary law. Some, among them Malawi, tried to establish a more 'indigenous' legal order. But usually they reverted to the old system after a while. Indonesia, although it claimed to indigenize its legal system, by and large continued the development towards unification that had been initiated by the government of the Dutch East Indies. In most former colonies the issue of recognition is hardly debated any more; there are no serious attempts in countries of the Third World to found a national legal system upon customary law.

However, in countries with indigenous populations—such as Canada, the USA, Latin America, New Zealand, and Australia—the discussion has presented a high political profile since the 1980s. Human rights, in particular the right of self-determination and the protection of cultural rights of indigenous peoples, form the focus of the debate. Many countries negotiate forms of (internal) self-determination and a certain degree of autonomy for indigenous peoples. Debates about minorities in Eastern Europe and the former Soviet Union, and discussions about ethnic minorities in Western Europe, show parallel features with the debates about indigenous peoples (Finkler and K. von Benda-Beckmann 1999). Notions of cultural identity are increasingly used as an argument for autonomy and recognition of group rights. They are also used as an argument to explain and defend 'delinquent' behavior (i.e., the 'cultural defense').

The debate recurs in the context of natural resource management and environmental protection. Indigenous groups and their representatives are trying to

develop 'indigenous' forms of management of tropical forest, land, and water, and call for recognition of local communal rights. Participatory development programmes for natural resource management are also based on the—contested—idea that customary forms of regulation are more adequate than other types of regulation. Notions of participation, self-government, 'good governance,' and sustainability, developed in international law and administration, play an important role in these discussions.

3. Studying Customary Law

The empirical study of customary legal systems began at the end of the nineteenth century. During that century missionaries, travellers, and civil servants had provided a host of widely diverse descriptions of the customs that they encountered. The colonial governments had become particularly interested in land tenure systems, family and inheritance laws. Scholars in the then emerging social sciences developed an interest in customary law as a means to study the evolution of modern society. Thus it was a combination of practical concerns of the colonial government and scholarly interest that elicited a long and rich tradition of the study of customary law.

One problem was how to study unwritten legal systems. The ordinary approach of lawyers to look at legislative and court decisions was inappropriate, because there were virtually no documents. Over the years, different approaches to solve this problem have been developed with different aims. The German legal scholars Post and Kohler were the first to set up a systematic worldwide collection of unwritten legal systems to be used for comparative study. Unlike researchers of the twentieth century, who went out to do extensive field research to learn the ins and outs of particular legal systems, Post and Kohler sent out questionnaires with detailed questions on all aspects of a legal system, to missionaries, travellers, and civil servants. Another German scholar, Bachofen, and his contemporary, the British scholar Sir Henry Maine, each published a comparative study of legal systems in 1861—*Das Mutterrecht* and *Ancient Law* respectively—works firmly rooted in an evolutionary perspective.

From the beginning of the twentieth century a large number of empirical studies began to appear on customary legal systems. Until then the focus had been on a description of the rules of a particular society. But due to scholars like Ehrlich, Malinowski, and Van Vollenhoven (see Holleman 1981), the focus shifted towards a study of the working of law in society. It was their contention that the law could be distilled from ordinary social behaviour. Malinowski's *Crime and Custom in Savage Society* (1926) became one of the most famous examples in this tradition.

From the 1930s on, British, South African, and American scholars doing research in Africa developed

a rich research tradition in the study of customary law with a strong focus on disputes. Standing firmly in the evolutionary tradition, they sought to find the original, authentic customary law, unaffected by modernity and the law of the colonial state. Gluckmann (1965, 1967), Gulliver (1963), and Llewellyn and Hoebel (1967) carefully—sometimes even purposefully—disregarded the colonial state and its representatives. As a consequence, they did not consider it necessary to deal with complex normative structures. Pospišil (1958, 1967) and Bohannan (1965) did write about legal complexity, although they did not deal with the law of the colonial state. The disregard of the state came under vehement attack in the 1970s from critics such as Moore (1978), Spittler, and Von Trotha, who argued that, contrary to what these authors suggested, the state and its law were not at all absent.

Since the 1970s, several important developments have taken place in the study of customary law.

(a) While the earlier studies concentrated on disputes as a source for learning about law, later on the disputing process itself moved into the centre of attention. The 'Berkeley village law project' under Laura Nader led to the classic *Law in Ten Societies* (Nader and Todd 1978). Schott (1978) wrote *Das Recht gegen das Gesetz*. The main objective was no longer the study of customary rules and regulation as such, but the study of modes of dispute management outside courts or court-like institutions (Morse and Woodman 1987); Gulliver's work on negotiation and mediation (1979) is a famous example. These studies have been influential in the movement towards alternative dispute resolution (ADR) in the USA and Europe. ADR is becoming an important export article of the USA to developing countries. For a critical examination of these modern forms of customary law, see Nader (1996); see also K. von Benda-Beckmann (2000).

(b) Customary law is no longer exclusively studied in the context of disputes. Based on the theoretical insight as developed by Ehrlich, Malinowski, and Van Vollenhoven (see Holleman 1981), that law operates primarily in ordinary social life, the role of law in ordinary social relations and social behavior has become a major object of research (F von Benda-Beckmann 1979).

(c) Customary law is nowadays always seen as a part of complex legal systems. It is studied in its relation not only to the law of the state but also to religious law, with which it merged in some societies, while in others the two are considered to be different bodies of law. It embraces, besides the more classical forms of customary law, all kinds of self-regulation within companies and in business branches, ethnic communities, the Mafia, local communities, and organizations of various kinds. In many instances mixed forms of law develop, or unnamed new forms of law (De Sousa Santos 1977). International law, in-

cluding Human Rights of Indigenous Peoples, has provided a new stimulus in the study of customary law in its relation to state law and international law (Merry 1997).

4. The Construction of Customary Law

Many authors have warned that what is often thought to be authentic customary law, in the sense of untouched by Western influence, has in fact been fundamentally influenced by colonial rule. The terms 'invented' or 'fabricated' customary law are proposed to indicate the changed character of customary law as it is usually found today. Some authors maintain that it does not make sense to even speak of customary law as some form of changed or transformed indigenous law, because what has been constructed by the state institutions is an entirely new type of law, invented by the state (Moore 1986, Wiber 1993).

This is not a problem that belongs to the past. In former colonies most courts still apply customary law to many disputes in which the rural population is involved. Parts of customary law are thus incorporated into the law of the state. This is a case of constructed legal pluralism. However, application of customary law regards substantive rules only; procedures of state courts are always designed by the state. Thus 'traditional' substantive law is applied under procedures of state law, as a result of which the substantive rules have undergone important changes. In fact, two kinds of customary law coexist: local law as it is developed and applied by the local population itself; and 'customary law,' or 'traditional law,' as it is applied by government institutions. These two types of law are certainly not fully independent, but stand in complex relationships towards each other and towards the law of the state. Dutch authors writing about colonial law, such as Van Vollenhoven were well aware of what was happening (Holleman 1981). The term *adat* law was designed precisely to indicate the difference from *adat*—the Malay term for local customs in the widest sense of the word, the way of life. Today the term *adat* law is generally used for both the state defined and the local version of customary law.

Courts play a central role in the creation of customary law of the state. Other government institutions (notably land registration offices and extension services) as well as banks adopt the courts' interpretations, enlarging the realm of state-defined customary law. Sometimes these institutions even develop their own version which is slightly different from the courts' version. Thus the influence of international institutions like the World Bank and also internationally operating NGOs and donor agencies has not only renewed the interest in customary law, it also has made the study of customary law more complex.

5. Conclusion

For a long time, customary law was regarded as a vestige of the past that was bound to disappear, but it is clearly far more resilient than was expected. It is subject to change, and it always develops in relationship with, and in response to, other normative orders—of which most important are the state legal system and religious law—but increasingly so in response to international law. In addition, self-regulation and other forms of regulation not emanating from the state or state-like institutions are new twigs on the branch of customary law, giving the concept a far broader scope than it used to have.

See also: African Legal Systems; Common Law; Conventions and Norms: Philosophical Aspects; Law: Change and Evolution; Legal Systems, Classification of; Norms; Tradition, Anthropology of; Traditions in Sociology; Traditions: Social

Bibliography

- Benda-Beckmann F von 1979 *Property in Social Continuity: Continuity and Change in the Maintenance of Property Relationships Through Time in Minangkabau, West Sumatra*. Martinus Nijhoff, The Hague, The Netherlands
- Benda-Beckmann K von 2000 Transnational dimensions of legal pluralism. In: Fikentscher W (ed.) *Begegnung und Konflikt—eine kulturanthropologische Bestandaufnahme*. C H Beck, Munich, Germany
- Bergh G van den 1986 The concept of folk law in historical context. In: Benda-Beckmann K von, Srijbosch F (eds.) *Anthropology of Law in The Netherlands*. Foris, Verhandelungen van het Koninklijk Instituut voor Taal-, Land- en Volkenkunde 116, Dordrecht, The Netherlands, pp. 67–89
- Bohannon P 1965 The differing realms of the law. In: Nader L (ed.) The ethnography of law. *American Anthropologist* 67: 33–42
- Finkler H, Benda-Beckmann K von (eds.) 1999 *Folk Law and Legal Pluralism: Societies in Transformation*. Papers of the XIth International Congress. Department of Circumpolar Affairs, Ottawa, Canada
- Gluckman M 1965 *The Ideas of Barotse Jurisprudence*. Manchester University Press, Manchester, UK
- Gluckman M 1967 *The Judicial Process Among the Barotse of Northern Rhodesia*, 2nd edn. Manchester University Press, Manchester, UK
- Gulliver P H 1963 *Social Control in an African Society*. Boston University Press, Boston
- Gulliver P H 1979 *Disputes and Negotiations: A Cross-cultural Perspective*. Academic Press, New York
- Holleman J F (ed.) 1981 *Van Vollenhoven on Indonesian Adat Law*. Martinus Nijhoff, The Hague, The Netherlands
- Hooker M B 1975 *Legal Pluralism: An Introduction to Colonial and Neo-colonial Laws*. Clarendon Press, Oxford, UK
- Llewellyn L, Hoebel E A 1967 *The Cheyenne Way: Conflict and Case Law in Primitive Jurisprudence*, 2nd edn., 4th print. University of Oklahoma Press, Norman, OK
- Malinowski B 1926 *Crime and Custom in Savage Society*. Routledge, London

- Merry S 1997 Legal pluralism and transnational culture: The Ka Ho'okolokolonui Kanaka Maoli Tribunal, Hawai'i, 1993. In: Wilson R A (ed.) *Human Rights, Culture and Context: Anthropological Perspectives*. Pluto Press, London, pp. 28–48
- Moore S F 1978 Archaic law and modern times on the Zambezi: Some thoughts on Max Gluckman's interpretation of Barotse law. In: Gulliver P H (ed.) *Cross-examinations: Essays in Memory of Max Gluckman*. E. J. Brill, Leiden, The Netherlands, pp. 53–77
- Moore S F 1986 *Social Facts and Fabrications: 'Customary' Law on Kilimanjaro 1880–1980*. Cambridge University Press, Cambridge, UK
- Morse B W, Woodman G R (eds.) 1987 *Indigenous Law and the State*. Walter de Gruyter, Dordrecht, The Netherlands
- Nader L 1996 Civilization and its negotiations. In: Caplan P (ed.) *Understanding Disputes: The Politics of Argument*. Berg, Oxford, UK, pp. 39–63
- Nader L, Todd H F (eds.) 1978 *The Disputing Process: Law in Ten Societies*. Columbia University Press, New York
- Pospisil L 1958 *Kapauku Papuans and their Law*. Yale University Publications in Anthropology, New Haven, CT
- Pospisil L 1967 The attributes of law. In: Bohannon P (ed.) *Law and Warfare: Studies in the Anthropology of Conflict*. Natural History Press, Garden City, NY, pp. 25–41
- Schott R 1978 Das Recht gegen das Gesetz. In: Kuhlbach F, Krawitz W (eds.) *Recht und Gesellschaft. Festschrift für Helmut Schelsky zum 65. Geburtstag*. Duncker & Humblot Berlin, pp. 605–38
- Sousa Santos B de 1977 The law of the oppressed: The construction and reproduction of legality in Pasagarda. *Law and Society Review* 12(1): 5–126
- Wiber M 1993 *Politics, Property and Law in the Philippine Uplands*. Wildrid Laurier University Press, Waterloo, Ontario, Canada

K. von Benda-Beckmann

Folk Religion

The term 'folk religion' is used rarely today, as it is loosely defined, with the word folk bearing several overtones. It is associated primarily with the English word 'folk-lore,' first used in 1846, and referring to traditional popular beliefs, stories, customs, and superstitions. It relates to the nineteenth century belief that research into 'survivals,' or ancient customs and traditions, as practiced by, e.g., English peasants, and/or 'primitive' peoples, could provide reliable clues to the origins, evolution, or 'progress' of mankind. Folk religion, as part of folklore, is therefore linked to certain European Renaissance, Enlightenment, and post-Enlightenment concepts within psychology and the philosophy of history, concerning the nature of mankind and modes of cultural reasoning. These include human evolution and diversity, comparative religion, and the social functions of religion, including its cultural adaptations.

1. Folk Religion in England

In England and in many other Christian countries, folk religion referred to the survival of pre-Christian and medieval practices. The latter were rooted in the Roman Catholic traditions of pre-Reformation England, when peasants learned through performative ritual rather than intellectual adherence to theological beliefs. The Latin prayers of the medieval Church blurred into charms and superstitions when offered by an illiterate and uncomprehending laity to ward off disaster and disease. Fertility and purification rites, ranging from corn dollies to the churching of women, were customary in rural areas. They became part of folklore, as did continuing beliefs in fairies, portents, ghosts, and witches. Such beliefs often continued well into the twentieth century. In the sixteenth century, the Protestant Reformation brought about the religious use of the vernacular, advances in printing, and increasing literacy. Nevertheless, in the sixteenth and seventeenth centuries, the boundaries of magic, science, and technology overlapped. Dreams were often accepted as supernaturally inspired, on a par with divine judgments and the direct intervention of God in the physical world. From the eighteenth century onwards, such beliefs continued to bridge the divisions between folk and conventional religion.

2. Explorations and Learning in Europe

Throughout Europe, the geographical explorations and new learning of the Renaissance was followed, in the eighteenth and nineteenth centuries, by increasing interest in man's origins and in the natural world. Great interest was shown in the Americas, Asia, the Pacific islands, and in their inhabitants. Writers ranging from Michel de Montaigne to Jean-Jacques Rousseau drew numerous comparisons between these newly discovered peoples and the ancient founders of European civilization, the Greeks and the Romans. At first this took the form of social speculation, of antiquarianism, and of archaeological excavations, such as the eighteenth century excavations of Pompeii and Herculaneum by Johann Winckelmann, which attempted to prove the historical reality of places and events described by the ancient writers. In the nineteenth century, with the expansion of the French and British empires, the interest in antiquity began to include the Middle East. Efforts were made to verify sites mentioned in the Bible. This was partly in response to devastating scientific attacks upon the literal creationism of the Book of Genesis, notably those made by the concepts of geological time, and by evolutionism (as expressed in Charles Darwin's *The Origin of Species on the Basis of Natural Selection* 1859). As well as those interested in establishing the origins of mankind through Biblical truths, others sought the missionisation and Christian conversion of

technologically simple peoples recently encountered in Africa, Australia, and other newly colonized areas. Others saw possible parallels between such people's behavior, customs, and beliefs, and those of early humans.

3. Comparative Religion and Human Origins: Totemism and Evolution

Such beliefs included religion. Under this came polytheism, monotheism, and the concept of a high god; the after-life; theodicy; and divine agency. Opinions were divided as to whether there existed any peoples so 'primitive' as to recognize no god at all. Some commentators, such as John McLennan, and William Robertson-Smith, were preoccupied by sacrifice and cannibalism, sometimes relating these to the origins of Christianity. In *Lectures on the Religion of the Semites* (1889), Robertson-Smith argued that totemism was the earliest form of religion, and that the periodic eating of the totemic animal was both a form of sacrifice and a form of cannibalism, an idea later adopted by Sigmund Freud. Others, such as James Frazer (*The Golden Bough* 1900–22), believed that the answer lay in careful comparative study of all known religious and other rites. Frazer's attempt to elucidate the meaning of divine sacrifice assumed that the comparison of similar customs throughout the world might help to understand the behavior of early humans. This inferred the racial and psychic unity of all mankind. Frazer believed in evolution and the idea of progress, as did his contemporary, Edward Tylor. Tylor, the founder of what is now seen as modern British anthropology, was influenced by Auguste Comte, and became particularly interested in animism and the definition of culture. Frazer and Tylor theorized that the intellectual development of mankind could be described as a movement from magic to religion, and from religion to scientific thinking, an idea later taken up by Bronislaw Malinowski (1925).

4. Folk Religion: Social Functions, Evolution, and Diversity

Folk religion, like folklore, implies the fusion of ancient and modern religious practices, customs, and beliefs. As such it refers to what would today be called multiculturalism, cultural adaptation, and syncretism. These are studied with reference to their social functions, rather than to the theological and philosophical questions which beset man's search for meaning in the universe. As an idea, folk religion belongs both to the Enlightenment and counter-Enlightenment traditions within anthropology. Its most notable representatives include Tylor and Frazer, who tried to present religious beliefs in evolutionary

terms, and Johannes Herder and Max Müller, who insisted on the irreducible diversity of human cultures. Herder argued that cycles of growth and decay affected both man and nature, and that the art and literature of each nation, the 'Volksgeist,' had a unique cultural value that was incompatible with developmental hierarchies. Belief in the Volksgeist was to fuel nineteenth-century nationalism and claims that particular ethnic groups were entitled to their own homelands or nation-states. Müller, a folklorist and noted religious scholar, held that ethnic identity sprang from two human universals, religion and language.

5. *History and Psychology: Society's Impact on the Individual*

The belief that mankind consisted of one species, rather than several racial subspecies, meant that concepts of psychic unity coincided with questions about society's impact on the individual. This was already a well-established theme in the philosophy of history, with the idea that the course of history was ruled by something greater than the influence of individuals who participated in it. The Renaissance historian Giovanni Vico, who developed a cyclical theory of civilisation, attributed this to divine intervention. Georg Hegel saw competing ideas, 'thesis' and 'antithesis,' as determining mankind's behavior. Karl Marx, who had read both Vico and Hegel, claimed ideas were not autonomous but firmly rooted in material conditions which produced an ideological superstructure, a 'false consciousness,' which affected human action. In psychology, William James emphasized the social functions of religious practice. Freud, reworking Christian concepts of original sin, promulgated belief in the unconscious: Carl Jung, that of the collective unconscious. Emile Durkheim's *Les Formes élémentaires de la Vie Religieuse* (1912) (translation *The Elementary Forms of the Religious Life*) held that religious rites were mechanisms for affirming social solidarity, involving collective representations of the group. Durkheim's idea of collective representations influenced Lucien Lévy-Bruhl's concept of prelogical mentalities and discussion of the ways in which the mental capacities and/or, reasoning devices of so-called primitive people were thought to differ from those seen as more civilized and rationalistic. As Evans-Pritchard (1965) pointed out, those who were engaged in the sociological and anthropological study of religion generally regarded it from an Enlightenment, rationalist viewpoint, as a series of superstitions which could only be explained in terms of cognitive variation or social function.

Religion became linked to the study of emotions, to the nature of the self, and to psychological interpretations of social functions, but also to theories of economic exchange, social control, and political legiti-

mization: links which continue today. One effect of this was that religion tended to be seen less as a holistic phenomenon than as the sum of several parts, any of which could be used to support arguments relating to the functions and meanings of other social institutions.

6. *Cultural Groups and Modes of Reasoning*

The ways in which modes of reasoning appear to differ between cultural groups is a perennial theme in anthropology, even though the arguments have long since moved from questions of innate capacities, and from the relationship between magic, science, and religion, to the ways in which particular cultural premises favor different forms of interpretation. Malinowski and Arthur Radcliffe-Brown were the first two European anthropologists to base their arguments upon personal fieldwork rather than armchair speculation. Malinowski held that the religious practices of 'primitive' peoples displayed a rational response to their surroundings, rather than one infused by the mysticism of Lévy-Bruhl. Radcliffe-Brown, influenced by Durkheim, interested himself in the expressive and affective aspects of tribal ceremonies, rather than in their possible origins.

Evans-Pritchard, like his student Mary Douglas, emphasized the logic underlying belief in witchcraft and other quasireligious practices, showing that indigenous interpretations depended on analogies, rational deduction, and the influence of emotions upon a particular line of reasoning.

7. *Folk Religion: Tradition and Adaptation*

Even though the phrase is obsolete, folk religion usually denotes either the traditional beliefs and practices of a particular people, or, more often, remnants from an earlier age. In *Folkways* (1906), the political scientist William Sumner argued that folkways, or the belief in and enactment of folklore, were the result of illiteracy combined with an intellectual closure produced by overwhelming deference to tradition. He pointed out that all beliefs were constantly adapting themselves to the exigencies of everyday life, so that 'pure' customs did not exist. Today, folk religious beliefs and practices may refer to those which have survived attempts to destroy an indigenous culture by colonialism or other means. They are then often described in terms of ethnicity and cultural resistance to forms of oppression. Such oppression may range from the exclusion of certain ethnic groups from formal learning systems to the class exclusion of the illiterate or poorly educated, including those living in rural areas and traditionally defined as peasants. Cultural resistance may encourage the oppressed to cling to beliefs and practices discarded by other social

or ethnic groups, or to create particular meanings and interpretations from their experiences, which are not countered by the interpretations given in formal meaning systems. The result is cultural adaptation, as elements of both the dominant and traditional cultures fuse together. This may give rise to religious syncretism, which may be both formal and informal, encompassing both distinctly new forms of cultural expression (e.g., Brazilian spiritist movements such as Umbanda and Candomblé) and the reworking of the previously discarded beliefs and practices of formal religious systems. Such peasant remnants and reworkings may be ignored by rationalistic and educated believers with higher social status (for a description of this, see Christian 1972). The former may continue to exist without reworking, with few substantial changes, so that they effectively consist of religious survivals, the folk religion of many years ago.

See also: Anthropology; Anthropology, History of; Belief, Anthropology of; Evans-Pritchard, Sir Edward E (1902–73); Folklore; Freud, Sigmund (1856–1939); Healing; Magic, Anthropology of; Malinowski, Bronislaw (1884–1942); Millennialism; Myth in Religion; Popular Culture; Religion, Sociology of; Shamanism; Tradition, Anthropology of; Witchcraft

Bibliography

- Christian Jr. W A 1972 *Person and God in a Spanish Valley*. Seminar Press, New York
 Evans-Pritchard E E 1965 *Theories of Primitive Religion*. Clarendon Press, Oxford, UK
 Malinowski B 1974 *Magic, Science and Religion, and Other Essays* (1925). Souvenir Press, London
 Morris B 1987 *Anthropological Studies of Religion: An Introductory Text*. Cambridge University Press, Cambridge, UK
 Stocking Jr. G W 1987 *Victorian Anthropology*. Free Press, New York
 Thomas K 1971 *Religion and the Decline of Magic*. Weidenfeld & Nicolson, London

M. A. Taylor

Folklore

'Folklore' has four basic meanings. First, it denotes oral narration, rituals, crafts, and other forms of vernacular expressive culture. Second, folklore, or 'folkloristics,' names an academic discipline devoted to the study of such phenomena. Third, in everyday usage, folklore sometimes describes colorful 'folkloric' phenomena linked to the music, tourist, and fashion industries. Fourth, like myth, folklore can mean

falsehood. While in this text, focus is placed on the first two meanings, the other two are often present, implicitly or explicitly.

1. History

1.1 A Term is Invented

The word folklore was coined by British antiquarian William John Thoms who, in 1846, proposed that this 'good Saxon compound' be used to describe 'the manners, customs, observances, superstitions, ballads, proverbs' and other materials 'of the olden time.' Thoms was inspired by the works of the Brothers Grimm and 'folklore' is apparently a translation of the German 'Volkskunde' introduced in 1806.

Although the word was new, the interest in the phenomena themselves was not. European rulers had long ordered the collecting of traditions that would glorify them, and religious leaders had gathered dangerous beliefs in order to weed them out. However, during the national romantic currents of the late eighteenth and early nineteenth centuries, intellectuals added something new, when they insisted that the customs and arts of illiterate peasants were remnants from a remote human past and should be collected before they were bound to disappear as civilization advanced. Like Johann Gottfried von Herder, many intellectuals also saw a reflection of the true spirit of nations in the traditions of the anonymous folk living close to the national soil.

1.2 A Discipline is Founded

During the nineteenth century, many people were fascinated with folklore, among them scholars with interests in literature, philology, religion, psychology, history, archeology, anthropology, and other fields. In some countries enthusiasts formed societies: the British Folklore Society was founded in 1878, and the American equivalent 10 years later. In other countries, primarily in northern and central Europe, the idea of folklore was directly linked to nation formation. In Finland, which was in the process of freeing itself, first from Swedish and then from Russian domination, Elias Lönnrot pieced together the *Kalevala Epic* from songs collected in oral tradition. Published in 1835, the *Kalevala* became the first example of literature written in Finnish and a national symbol. In addition, the songs formed the core of a national folklore archive and, in 1898, the world's first professorship in folklore was established: the chair in 'Finnish and Comparative Folklore' at the University of Helsinki. Other departments were to follow in other European countries, often in conjunction with the founding of national museums and archives.

However, in many countries, folklore continued to be studied outside formal institutions. In some cases, such as England, separate university departments were never established, whereas in others, notably the United States, they were founded as late as after World War II (see Sect. 3).

1.3 A History of Approaches

On the whole, the ideas employed to explain and interpret folklore coincide with the major schools of thought in anthropology, beginning with evolutionism and diffusionism and moving on toward functionalism, structuralism, and later tendencies. Yet even more than in anthropology as a whole, the study of folklore was long focused on questions of origin; an array of explanations were proposed to account for the origin of such ancient and widely disseminated tales as *'The Dragon Slayer'* and other traditions. The Indianist school championed by Theodor Benfey sought to establish that all European folktales had originated in India and spread from there. The mythological school headed by Max Müller found that all folktales were survivals of ancient nature-myths, while ritualists such as Pierre Saintyves tried to demonstrate that all folktales had originated in rituals.

While these explanations now have mostly antiquarian interest, other early trends of thought remain important. Among these is, first, the diffusionist 'historical-geographical' method developed by Finnish scholars at the end of the nineteenth century. In this cumbersome procedure all recorded versions and variants of a tale type (or a ballad or a riddle) are studied so that the location of the earliest traits (and theoretically also the tale's place of origin) can be determined. The gathering of comparative evidence resulted in massive indexes and catalogues that are still often consulted (Aarne and Thompson 1961). A second intellectual movement that remains important to the understanding of folklore is psychoanalysis. Sigmund Freud, C. G. Jung, Otto Rank, and others came to regard folklore as essential expressions of the individual or collective human unconscious. A third essential strand of thought is the insight that forms of folklore are not only widely spread but also firmly localized and therefore must be studied in their immediate social contexts. Franz Boas' ethnographies of the stories of native peoples on the north west coast of the United States initiated an influential format for the study of folklore in local contexts.

Functionalism, not least Bronislaw Malinowski's analyses of story telling among the Trobrianders, is a fourth development that has had a lasting impact on the study of folklore. Particularly influential is the psycho-functional premise that folklore is 'socially sanctioned' fantasy; as such it permits action that is otherwise not approved (Dundes 1965, p. 277). Finally, structuralist thought has also endured among students of folklore. One structuralist approach deri-

ves from Russian formalism and emphasizes features of plot sequence; in an influential study Vladimir Propp demonstrated that all Russian fairytales share the same plot structure. A second approach derives from the work of Claude Lévi-Strauss and points to the ways in which logical relationships within myths and tales (such as binary oppositions) reflect universal processes of the human mind. Lévi-Strauss' thinking has been central to the analysis of a wide range of traditional forms, not least to Henry Glassie's monumental investigations of vernacular material culture.

2. Key Words

Despite the many decades of inquiry, it remains difficult to define folklore. In one recent attempt it is suggested that folklore embraces 'expressive forms, processes, and behaviors' that are learned, taught, and utilized 'during face-to-face interactions.' It is also suggested that people judge these phenomena to be traditional 'because they serve as evidence of continuities ... through time and space in human ... thought, belief, and feeling' (Georges and Jones 1995, p. 1).

However, many scholars avoid generalized statements and prefer to look at folklore as perspectives and processes that might be illuminated through such keywords as 'folk,' 'tradition,' 'genre,' 'performance,' 'text,' and 'art' (cf. Feintuch 1995).

2.1 Folk

Influenced by the typologies of early social thought and by such conceptualizations as Robert Redfield's 'ideal folk society,' students of folklore long took it for granted that the folk (i.e., the carriers and transmitters of folklore) were illiterate peasants or, to some extent, tribal peoples. However, during the last few decades, the word folk is increasingly regarded as a cumbersome inheritance from earlier thinking and most scholars agree that all human beings in some way or another are involved in folklore processes, as performers, participants, watchers or audiences—sometimes even as indifferent or disdainful audiences (see Sects. 2.4–2.6). But it has been difficult to abolish the word folk. For one thing, it remains important in studies of folklore in feudal systems or other class hierarchies. Second, it remains alive in such compounds as folk art, folk music, and folk dance; attempts to replace 'folk' with 'people,' 'vernacular,' or other words have not been successful (see Sect. 3). Furthermore, despite their conviction that all human beings partake in folklore processes, contemporary students also concentrate on marginal or subaltern peoples. But their reasons for doing so differ from those of nineteenth century predecessors; many students now argue that the role of folklorists today is to assist peoples who lack political power in their quest to gain a public voice.

2.2 Tradition

Used both to name a process of transmission and to name specific cultural elements (i.e., traditions), the concept of tradition is basic to the understanding of folklore and folkloristics. It has long been recognized that tradition not only involves historical continuity but also change. Because folklore forms are often transmitted orally or in other informal ways, they are extremely adaptable to fleeting moments of social interaction. One could even say that it is because of their adaptability that simple string-figures or guessing games have remained in circulation through millennia.

The materials deposited in archives and listed in comparative indexes demonstrate both the adaptability of some themes and their continuity through time and space. For example, comic book hero Superman travels through space in a womblike vessel only to be found and raised by a poor couple in the middle of North America. His story has parallels in the stories of Moses, Cyrus, and other heroes who are similarly found as children floating around in small containers. Furthermore, identity changes are not only characteristic of Superman but also of countless other heroes and heroines in world traditions. The transformation of an orphaned young girl into a rich and beautiful queen is reenacted over and over in oral tales, films, and real life. Folklore themes have a psychological force that makes people recreate them century after century in different communicative modes and media.

During recent decades, students of folklore increasingly utilize the verb 'traditionalize' to name processes in which the past is interpreted or reinvented in a given present. Many students of folklore now emphasize that they do not investigate given sets of traditions but all sorts of practices of traditionalization in social life. Furthermore, by focusing on active traditionalization processes, scholars avoid cementing an opposition between tradition and creativity. They see tradition as a source of creativity, not as a constraint to it (Briggs 1988).

2.3 Genre

Genre means 'kind' and a genre may be seen as a bounded distinguishable form characterized by a specific combination of stylistic traits and elements of content. The classification into genres was for a long time so central that scholars thought that enumerating genres was the easiest way to describe what folklore is (Dundes 1965).

When analyzing genres, students of folklore often find themselves caught in dilemmas familiar to all students of culture. On the one hand, all peoples make local, culture specific, or 'emic' distinctions between genres in their repertoires: coyote tales, superstitions, dirty jokes, little girls' games, gossip, songs while making baskets, and so on. On the other hand,

scholars may observe more overriding features in the same materials and may note how elements of content, style, or structure are shared across cultural boundaries. The extent to which local classifications are emphasized and the extent to which broader ones come into play have to vary with different analytic tasks.

One distinction that was long taken for granted but has proven problematic is that between 'fixed' genres, such as riddles and proverbs, and 'unfixed' genres, such as many narratives. While some genres, magic formulas for example, are fixed in the sense that they must be recited exactly as learned in order to have power, other genres that were long called fixed change too. For example, people often do not cite proverbs in full but merely hint at them through tone of voice. Thanks to performance-centered approaches (see Sect. 2.4), scholars have recently gained new insights into the ways in which oral genres are malleable social phenomena. In social interaction people mix or blur materials so that artful hybrids emerge. At the same time, genres are needed as 'social contracts' that aid memory and establish structure.

2.4 Performance

Since the 1960s, the concept of performance has transformed the study of folklore. Thanks to increasing refinements in recording techniques, scholars have been able to analyze folklore as part of communicative processes in ways that had not been possible before. Two performance-centered approaches will be recognized here; in both, careful ethnographic fieldwork is essential.

One approach is rooted in the ethnography of speaking and other sociolinguistic methods. A basic idea is that verbal art and other forms of folklore emerge in social interaction as aesthetically marked forms. They are recognized as something special thanks to cues made through tone of voice mimicry, gestures, and other devices. Arguing that performance is 'constitutive' of verbal art, Dell Hymes, Richard Bauman, and others have developed ways of studying folklore as artistic communication in everyday life, communication in which, as Hymes says, people shape 'deeply felt values into meaningful, apposite form.' If preachers, story-tellers, weavers, carvers, and carnival participants in the past were seen as transmitters of handed down traditions, they were now regarded as active artists with power to transform social life.

A second performance-centered approach is oral formulaic theory. This was first developed by Albert Lord and others who in studies of South Slavic oral epic poetry show how songs that are many thousands of lines long can live in tradition, although they have not been written down. They are remembered because singers learn to draw on a store of formulae and formulaic expressions and learn to adjust their singing

to different social events. Several scholars have continued along these paths and oral formulaic studies are now enriching the understanding of both ancient (in particular Homeric) epic singing and of contemporary epic performances in India, China, Africa, and Eastern Europe (Finnegan 1992, Honko 1998).

2.5 Text

It has long been recognized that the distinction between oral and written texts is far from firm and that all kinds of texts can be of interest to students of folklore. Some 'folklore texts' have always existed primarily in writing, for instance court records, religious documents, newspaper articles, and xeroxlore. Others have circulated widely as visual representations, for example fables painted inside medieval churches. Furthermore, literary authors have continuously made use of traditional themes and, in a sense, folklore texts can be found in the works of many celebrated writers. For example, contemporary African authors, such as Ayi Kwei Armah, acknowledge their profound debt to oral narrative art and Shakespeare often builds his plays on traditional materials. Actually, a great many art forms can profitably be studied as folklore texts. Classical ballet, for example, shares some of its stylized and repetitive character with the European folktale.

Recently, performance-centered approaches have opened up oral texts in unprecedented ways in that scholars have invented new methods to translate oral performances into writing. In procedures that have sometimes been named 'laying out texts ethnopoetically' students transcribe not only words but also pauses, cadences, gestures, mimicry, and other oral features (Briggs 1988). These procedures have led to many insights. One is that oral narration in many cultures tends to be closer to drama than to prose literature. Another is that repetitions and other features of oral narration (or visual folk art) are not to be dismissed as poor artistry as was often the case in the past. Rather, repetitions can be seen as ways to control complex materials. A repetitive structure simultaneously offers restrictions and creative freedom; it aids memory at the same time as it frees fantasy and improvisation.

During the last few years, the term 'textualization' is increasingly used to describe modes of translating oral performances onto the written page (Honko 1998). Long regarded as routine exercises, textualization processes are now seen as central analytic tasks.

2.6 Art

On some level, scholars always seem to have agreed that folklore has to do with art, creativity, and imagination. Even at times when the study of folklore

was reduced to enumerations of genres and tale types, there was a sense that folklore touches upon the aesthetic dimensions of life. This sense has become heightened during the last few decades, in part due to the interest in folklore as performance. This is particularly the case in the United States, where all kinds of vernacular expressions—from singing to cooking and quilting—are often called the 'the folk arts.'

The focus on art is linked to several perspectives. First, students make a distinction between less articulate folklore performers and artists who are celebrated in their communities and have the power to move others to action. Second, students of folklore have made important discoveries about social life by focusing, not on finished products, but on moments when material objects or community events are being planned or created. Third, the interest in art has led to new analyses of the languages of folklore. Whether verbal, visual, musical, kinesthetic, or tactile, these languages are bold in their use of intense and multi-valent metaphors and other tropes.

It is because of this expressive intensity that folklore, like other forms of art, can be a vehicle through which people can examine or critique the lives they lead. For example, to people who have fled from one part of the world to another, a few songs, stories, or dishes can become embodiments of a precious inheritance which must be preserved and reconstituted in new lands. Conversely folklore can incarnate some of the worst aspects of life: detested symbols of racism and humiliating jokes that outsiders direct at one's own group. Because of its power of presence, folklore can be invoked to hold communities together or to separate them from one another.

3. Folklore and the Politics of Culture

Among contemporary scholars there is both confidence and unease about the future of folklore and folklore study. There is confidence, for example, about the supply of material. The idea that folklore is 'dying out' is no more true today than it was 200 years ago. People shape rituals and tell stories, perhaps more than ever before, but they do so not only in face-to-face interactions but also on the electronic media. There is also confidence in the sense that a great deal of scholarship is being conducted. Indeed, some scholars are convinced that, regardless of discipline, those who study folklore today play a special role as advocates of the artistic expressions and traditions of all those groups and individuals who are caught in the cracks of modern life.

But, on the other hand, scholars are uncertain about the directions of folklore study. There are renewed critical reflections on the discipline's long and euro-centric history and on its often naive, and sometimes disastrous, ideological and political entanglements (Briggs and Shuman 1993). One indication of the

uncertainties are the current controversies surrounding the designation of the field. In Germany, the term 'Volkskunde' is now often avoided, because it is deeply implicated in fascist scholarship and in the creation of symbols for the Third Reich. Here and in some other countries, 'ethnology' is preferred to compounds containing 'folk' (see Sect. 2.1). In other countries, the word folklore is perceived as an inheritance from a colonial era and is being replaced by new coinages, such as the South Indian 'janapada.'

Many critical reflections concern the paradoxical situation that folkloristics regards itself an international field at the same time as it is almost everywhere a construction in the service of nation building and related local endeavors. This was true in the nineteenth century and it is true in the twenty-first. For example, in some newly formed or recently reconstituted nations, folklore and folklore archives play a critical part in efforts to construct visions of homogeneous and exclusive national heritages. In other situations, folklore and folklorists are linked to nation building in other ways. In the United States, for example, they have become important in affirming a vision of the country in which a multitude of diverse cultures can co-exist. This is particularly true of so-called public folklorists. Employed in museums, at the American Folklife Center in Washington DC, or as state folk arts coordinators, these professionals have special skills in collecting, preserving, and presenting the folklore of an immense diversity of groups. But also on a global level folklorists participate in cultural politics. Many work with UNESCO in efforts to safeguard folk traditions around the globe from the onslaughts of modernity.

Whether nationally or globally based, all these endeavors to collect, preserve, and present folklore raise many questions. In whose interest are these efforts being carried out? Whose traditions should be safeguarded and whose should be deleted? What happens to all those traditions that are not beautiful or presentable but offensive, frightening, or incomprehensible? In the view of many scholars, it is the many different kinds of heritage politics and ideological mazes that ought to become an important focus of attention for folklore scholarship in the years to come.

Bibliography

- Aarne A, Thompson S 1961 *The Types of the Folktale. A Classification and Bibliography*. FF Communications 184, Suomalainen Tiedakatemia, Helsinki
- Bausinger H 1961 *Volkskultur in der technischen Welt*. Kohlhammer, Stuttgart (1990 *Folk Culture in a World of Technology*. Indiana University Press, Bloomington, IN)
- Briggs C L 1988 *Competence in Performance: the Creativity of Tradition in Mexican Verbal Art*. University of Pennsylvania Press, Philadelphia, PA
- Briggs C L, Shuman A (eds.) 1993 Theorizing folklore: toward new perspectives on the politics of culture. *Western Folklore* 52(April): 109–400
- Cocchiara G 1952 *Storia del Folklore in Europa*. Editore Boringhieri, Torino (1981 *The History of Folklore in Europe*. Institute for the Study of Human Issues Press, Philadelphia, PA)
- Dundes A (ed.) 1965 *The Study of Folklore*. Prentice-Hall, Englewood Cliffs, NJ
- Feintuch B (ed.) 1995 Common ground. Keywords for the study of expressive culture. *Journal of American Folklore* 108(430): 391–550
- Finnegan, R 1992 *Oral Traditions and the Verbal Arts. A Guide to Research Practices*. Routledge, London
- Georges R A, Jones M O 1995 *Folkloristics: An Introduction*. Indiana University Press, Bloomington, IN
- Glassie H H 1982 *Passing the Time in Ballymenone. Culture and History of an Ulster Community*. University of Pennsylvania Press, Philadelphia, PA
- Honko L 1998 *Textualizing the Siri Epic*. FF Communications 264, Helsinki
- Toelken B 1996 *The Dynamics of Folklore*. Utah State University Press, Logan, UT

B. Klein

Food in Anthropology

The study of food in anthropology involves simultaneous consideration of both culture and biology. Work spans a wide array of emphases from largely biological to largely cultural, and comprehends a range of topics, methodologies, and theoretical approaches. Food and culture studies emphasize the social and symbolic dimensions of food to gather insights about gender, class, political economy, power, cultural ideology, and cultural identity (Counihan and Van Esterik 1997). Nutritional anthropology gives more weight to biological considerations and examines the interplay of food supply and hunger, nutritional status, breast-feeding, disease, household dynamics, and human evolution (Goodman et al. 1999). The two different approaches overlap and complement each other and reveal the power of anthropological holism to explain the human condition.

1. History and Scope of the Field

Many early ethnographers paid attention to food habits because food is a central focus of the economy in most cultures. In noncapitalist societies, the production, distribution, and consumption of food take up a great deal of people's time, energy, and attention. Marcel Mauss (1923) laid an important foundation of foodways research in anthropology with his pioneering study of exchange in *The Gift* (see also *Exchange in Anthropology*). Mauss directed our attention to the importance of giving, receiving, and repaying in initiating, maintaining, or breaking social and political relationships, and reminded us of how often food is a

key substance in exchanges. Gifts, Mauss pointed out, are 'total social phenomena,' a concept that applies beautifully to food. Mauss used ethnographic and historical literature to explore gifts in a variety of cultural forms, for example the Northwest Coast Indian potlatch and the Maori 'taonga.' He described the meaning of the ancient Hindu food offerings thus: 'The thing given brings a return in this life and the other ... Food given away means that food will return to the donor in this world; it also means food for him in the other world and in his series of reincarnations' (1967, pp. 54–5). Food gifts establish obligations of reciprocity in recipients whether they are gods or humans.

The centrality of food in gifts and economic transactions has ensured its central role in expressing religious beliefs, communicating with others, establishing and breaking relationships, defining social groups and families, and expressing cultural values. Yet few early anthropologists focused explicitly on food, the exception being Audrey Richards who conducted pioneering work on land, labor, diet, and hunger in Southern Africa (1932, 1939). She took a functionalist approach to food and examined the interconnectedness of food-centered production, distribution, consumption, and beliefs (see also *Functionalism in Anthropology*). She consistently examined not only food's important economic and nutritional dimensions, but also its role in constructing social relationships, showing for example how gifts of food declared and enacted the new obligations between families at marriage. Richards was the first anthropologist to do a full fieldwork study centered on food. Although her work was well regarded, it did not inspire followers for several decades.

The current burgeoning of studies in food anthropology began in the 1960s with the stunning work of Claude Lévi-Strauss (1969 [1964]) in France and Mary Douglas (1966) in England. Lévi-Strauss developed his structural theory by applying it to a symbolic analysis of food and other important symbols in hundreds of myths collected from Amazonian Indians. He introduced the idea of the 'gustatory code' (1969, p. 164) which is the most important of the five basic codes (stemming from the five senses) that humans use to transmit cultural messages. The gustatory code refers to the predominance of food symbols and metaphors in human expression which is due to 'the truly essential place of cooking in native thought: not only does cooking mark the transition from nature to culture, but through it and by means of it, the human state can be defined with all its attributes' (1969, p. 164). The study of myths and the consistent patterns that they display on diverse levels of meaning reveals that 'myths signify the mind that evolves them ... there is simultaneous production of myths themselves, by the mind that generates them and, by the myths, of an image of the world which is already inherent in the structure of the mind' (1969, p. 341).

Mary Douglas took a different approach to the symbolic study of foods by focusing on their ability to reveal not the mind's structure but social structure. She demonstrated how foodways revealed concepts of purity and pollution and constructed social boundaries in diverse cultural settings, most prominently in her study of Jewish dietary law. She argued that Jewish rules about food consumption continually reinforced concepts of holiness and renewed the community of the faithful. Together, the work of Douglas and Lévi-Strauss gave an influential demonstration of the role of food as 'good to think' as well as 'good to eat' and overcame the anthropological tendency to take food for granted.

Food studies in anthropology proliferated in the 1970s, 1980s, and 1990s. Researchers have used ethnographic, archaeological, and biocultural fieldwork methods to study foodways—the beliefs and behaviors surrounding the production, distribution, and consumption of food. Douglas and Lévi-Strauss' emphasis on the symbolic role of food did not go unchallenged. Marvin Harris (1985) responded to their work by arguing explicitly that food habits exist for utilitarian purposes. 'Preferred foods (good to eat) are foods that have a more favorable balance of practical benefits over costs than foods that are avoided (bad to eat)' (1985, p. 15). Harris gave priority to material over symbolic concerns in explaining Jewish dietary law. He argued that the pig was an economic competitor with humans in the arid lands of the Middle East where it made much better economic sense to raise ruminants who do not need high-protein foods to flourish as pigs do. Thus, he argued food aversions and preferences evolve primarily to support utilitarian practices based on nutritional and ecological concerns rather than to support religious belief systems.

Douglas and Harris's different explanations for Jewish dietary law show how the study of food provides diverse and complementary insights into human behavior. Anthropologists have examined foodways to gain insights into power in politics, family, gender, class, and race. They have studied human evolution, health, and disease; famine, hunger, and malnutrition; and the relationships between cultural ideology and behavior.

2. Food and Power

Because food is the most insistent and basic daily survival need, it is a central pawn in power struggles in many different settings. In Papua New Guinea, for example, men use food to acquire and demonstrate political leadership. In Kalauna, men 'fight with food' by shaming others through excessive gifts of comestibles (Young 1971). In Wamira, Papua New Guinea (Kahn 1986), men demonstrate their leadership and public renown by accumulating food and giving it away to followers and enemies alike at public feasts.

At the same time, they refrain from eating and thus demonstrate their self-control and prestige through 'the potency of [their] hunger-assuaging magic' and their ability 'to tame the most frightening of uncontrollable behaviors—namely ravenous, rapacious desire—both gustatory and sexual' (Kahn 1986, p. 86).

The attainment of power through giving food away typical of Melanesian and other subsistence cultures contrasts dramatically with the goal of capitalists, who attempt to withhold food and sell it for profit and power, as Mintz (1985) capably demonstrated in his anthropological study of the worldwide commodification of sugar. Sugar production involved Europeans in conquest of Third world countries, seizure of raw materials, and enslavement of Africans to provide land and labor to produce sugar cheaply enough to fuel a vast expansion of sugar consumption and exorbitant profits. The story of sugar provides a case study in how foodways research can uncover the interworkings of political, economic, and cultural systems of power.

Goody (1982) focused explicitly on the role of foodways in constituting class relations by examining the conditions for the emergence of high and low cuisine among some African, European, and Asian cultures. He examined how 'industrial food' can be a form of cultural domination of Third World cultures and explored the socioeconomic structures that enable them to resist. Meillassoux (1975) used a Marxist perspective and economic focus on food production and distribution to examine the persistence of the household economy under the corrosive conditions of colonization and capitalist labor exploitation. Weismantel (1988) showed how specific food behaviors identify class and race in Highland Ecuador. Indians regularly eat potatoes but only on special occasions do they consume the white rice that forms the staple food and symbol of class and race privilege of the whites. In the USA, distinctions between rich and poor, male and female, and white and 'ethnic' Americans are all enacted through foodways (Counihan 1999).

3. Food and Gender

Food is intimately connected with the definition and assertion of gender identity in many cultures (Counihan 1999). Food work ordinarily falls to women and whether this work is revered or denigrated is an important constituent in women's status. Among the Zumbagua of the Ecuadorian Highlands, women use their cooking to express approval or disapproval of their husbands, and they establish hierarchy in the family by their allocation of scarce and valued meat (Weismantel 1988). Among Wamirans of Papua New Guinea, women's roles in producing food and children are likened to each other and valued highly (Kahn 1986). To emulate women's important role in procreation and establish symbolic gender parity,

Wamiran men cultivate taro, which they call their 'children' and which represents their 'personal status and virility' (Kahn 1986, p. 91). Among the Hua of highland Papua New Guinea, men's elaborate system of food rules reveals how they construct gender and overcome their own fears of women's power of procreation (Meigs 1984). They define foods in two categories, *koroko* and *hakeri'a*, female and male. Female foods are wet, cold, fertile, soft, and fast growing; male foods are dry, hot, infertile, hard, and slow growing. Men publicly aver that women and the foods associated with them are polluting and dangerous, yet in rituals they secretly consume the female foods to gain the power, fertility, and vitality of women.

In all cultures, gender relations are clearly expressed through the attitudes of men and women towards eating, feeding, and the body. In contrast to the great majority of cultures, where women aspire to plumpness and pleasure in eating, in the USA, women reproduce their subordination by restricting their eating, reducing their bodies, and loathing their own appetites. As capitalism transforms women from producers into consumers and commodities, they lose social value and power, their bodies become objects, and they lose subjecthood and agency (Counihan 1999).

Anthropologists have shown how cultural practices and attitudes towards breast-feeding vs. formula feeding reveal much about women's power and status. Van Esterik (1989) used fieldwork data and documentary sources on cases in Thailand, Indonesia, Kenya, and Colombia to argue that the adoption of formula-feeding robs women of an important bodily ability, costs them money, and generates profits for multinational corporations at the expense of infant health, especially among the poor in the USA and the world. Empowerment of women depends on creating conditions where they can breast-feed if they choose with dignity and respect so as to ensure the best possible health of children and validation of women's corporeal abilities.

4. Nutritional Anthropology

The relationship between cultural practices, food consumption, health, and disease past and present has been a central focus of nutritional anthropology. Harris and Ross (1987) developed their edited volume *Food and Evolution* to demonstrate the important role in human development of the interplay of biological and behavioral 'repertoires' surrounding food. Goodman et al. (1999, p. 1) defined nutritional anthropology as the study of 'the interrelationships of biological and social forces in shaping human food use and the nutritional status of individuals and populations.' Their volume explored the evolution of human foodways with attention to both fossil and primate studies and the importance of plant and animal domestication

on diet and health. It looked at case studies on lactose tolerance, maize processing technology, and bitter cassava consumption to explore the development and impact of specific diets and food processing techniques and their ecological and health implications. They examined the biocultural causes and effects of hunger, malnutrition, and overeating.

5. Hunger and Malnutrition

Few anthropologists have written books on the causes and effects of hunger, the exception again being the pioneering work of Audrey Richards (1932). Recently there have been a growing number of good articles, some of which have been anthologized (see Counihan and Van Esterik 1997, Goodman et al. 1999, Harris and Ross 1987). Dettwyler (1994) is one of few anthropologists to produce a book-length study of hunger. She focused on feeding practices, health, and disease in Mali by combining ethnographic observations with nutritional methods. She used dietary nutrient assessment, height and weight measurements of children and adults, collection of health beliefs, observations of feeding and breast-feeding practices, and blood, urine, and stool sampling. She was able to assess health status and parasitic invasions and show their relationship to cultural practices and attitudes. For example, Malian beliefs about children's normal appearance being thin and pot-bellied obscured their ability to recognize or end malnutrition in children. Dettwyler's goal was to use her study to develop 'culturally appropriate nutrition education programs, aimed at improving people's understanding of the links between diet and health, and the importance of making sure young children get enough food, of sufficient quality, to meet their needs for growth and health' (1994, p.7). Nutritional anthropology demonstrates the analytic power of combining sensitive social observations with scientific data-gathering to understand the causes and potential cures of hunger, malnutrition, and the diseases of poverty.

6. Future Directions

Future studies of food in anthropology will continue to address links between cultural systems, social inequality, food, and hunger. Studies can do this by examining food in popular culture—particularly the use of food images in advertising, television, film and popular music—where the messages transmitted often penetrate public consciousness with little critical reflection. The study of McDonalds in East Asia by Watson (1997) and his collaborators links the anthropology of food to the theme of globalization. Anthropology's symbolic and biocultural approach is particularly useful in studying the growing epidemic of eating disorders, including anorexia nervosa, bulimia, and compulsive eating. These disorders strike women in

much greater numbers than men and seem to have psychocultural causes associated directly with women's subordinate status and socially constructed role as those who should feed others but not themselves. The cross-cultural study of the spread of, and resistance to, eating disorders in Third World countries where they were previously unknown will provide important insights into their causes, consequences, and cures.

As multinational corporations increasingly dominate food production and distribution, anthropologists will need to keep a close eye on the effects of food profit-making on access to and quality of food. Issues surrounding the bioengineering, genetic manipulation, loss of genetic diversity, and environmental pollution and protection in agricultural production will be paramount. Because of the centrality of food to human physical well being and cultural expression, anthropologists must continue to document and promote efforts to end hunger and to establish food as a human right.

See also: Consumption, History of; Fishing Societies; Food Production, Origins of; Functionalism in Anthropology; Horticultural Societies; Hunter-Gatherer Societies, Archaeology of; Hunting and Gathering Societies in Anthropology

Bibliography

- Counihan C M 1999 *The Anthropology of Food and Body: Gender, Meaning and Power*. Routledge, New York
- Counihan C, Van Esterik P (eds.) 1997 *Food and Culture: A Reader*. Routledge, New York
- Dettwyler K A 1994 *Dancing Skeletons: Life and Death in West Africa*. Waveland, Prospect Heights, IL
- Douglas M 1966 *Purity and Danger: An Analysis of the Concepts of Pollution and Taboo*. Praeger, New York
- Goodman A H, Dufour D L, Peltó G H (eds.) 1999 *Nutritional Anthropology: Biocultural Perspectives on Food and Nutrition*. Mayfield, Mountain View, CA
- Goody J 1982 *Cooking, Cuisine and Class: A Study in Comparative Sociology*. Cambridge University Press, Cambridge, UK
- Harris M 1985 *Good to Eat: Riddles of Food and Culture*. Simon and Schuster, New York
- Harris M, Ross E B (eds.) 1987 *Food and Evolution: Toward a Theory of Human Food Habits*. Temple University Press, Philadelphia, PA
- Kahn M 1986 *Always Hungry, Never Greedy: Food and the Expression of Gender in a Melanesian Society*. Cambridge University Press, Cambridge, UK
- Lévi-Strauss C 1964 *Le Cru e Le Cuit*. Librairie Plon, Paris [1969 *The Raw and the Cooked: Introduction to a Science of Mythology*. Harper and Row, New York]
- Mauss M 1923 Essai sur le don: Forme et raison de l'échange dans les sociétés archaïques. *L'Année Sociologique* 1: 30–186 [1967 *The Gift: Forms and Functions of Exchange in Archaic Societies*. Norton, New York]
- Meigs A S 1984 *Food, Sex, and Pollution: A New Guinea Religion*. Rutgers University Press, New Brunswick, NJ

- Meillassoux C 1975 *Femmes, greniers et capitaux*. F Maspero, Paris [1981 *Maidens, Meals and Money*. Cambridge University Press, Cambridge, UK]
- Mintz S W 1985 *Sweetness and Power: The Place of Sugar in Modern History*. Viking, New York
- Richards A I 1932 *Hunger and Work in a Savage Tribe*. Routledge, London
- Richards A I 1939 *Land, Labor and Diet in Northern Rhodesia*. Oxford University Press, London
- Van Esterik P 1989 *Beyond the Breast-Bottle Controversy*. Rutgers University Press, New Brunswick, NJ
- Watson J L (ed.) 1997 *Golden Arches East: McDonalds in East Asia*. Stanford University Press, Stanford, CA
- Weismantel M J 1988 *Food, Gender and Poverty in the Ecuadorian Andes*. University of Pennsylvania Press, Philadelphia, PA
- Young M W 1971 *Fighting with Food: Leadership, Values and Social Control in a Massim Society*. Cambridge University Press, Cambridge, UK

C. M. Counihan

Food Preference

Food and food choice constitute a major part of waking human activity, and, in the Third World, the major expense. Furthermore, food choice may be the single most important selective force in animal evolution, as suggested by the fact that many animal groups are named for their food habits (e.g., carnivores, insectivores). In human evolution, adaptations to enable the exploitation of large animals as foods played a significant role in the evolution of the human brain and human social organization. As argued persuasively by Jared Diamond (1999), early advances in food procurement, agriculture, and domestication set the stage for major advances in technology, the development of urban centers, and warfare. As a fundamental and frequent part of human life, food has served as a foundation for the cultural evolution of non-food systems; thus, food plays a major role in religion, ritual, and social exchange. It is also a source of metaphors, with words such as 'bitter,' 'meat,' and 'swallow' used in many non-food contexts. Thus, an understanding of human relationships to food promises to enlighten a major area of human concern, as well as other areas. The full range of food in human life is beautifully described in Leon Kass' book, *The Hungry Soul* (1994).

1. The Basic Relationship Between Humans and Their Food

For any person, the world can be divided into the self and everything else. Eating involves taking matter from outside the self and putting it inside the self. This

is a very intimate act. It is not surprising that people feel strongly about what they eat. The costs in terms of toxins, microorganisms, or imbalanced nutrients are high, but the benefits are at least as high: survival is at stake.

Humans (and rats and cockroaches) are food generalists. They eat a wide range of foods; virtually anything that can fit into the mouth is potential food. Generalists have few innate determinants of food choice, simply because it is not easy to predict the nutritive and toxic properties of a potential food on sensory grounds. There are only a few documented innate biases in human (and rat) food choice. First, there are innate predispositions to like or dislike certain tastes. There is an innate preference for sweet tastes, which, in nature, are predictive of calorie sources. The long history of sweetness in human culture, from fruit preferences, to cultivation of fruits, to extraction of sugar from fruits, to colonization of the Americas partly to get a source of sugar, to the development of sugar substitutes, is all driven by the innate liking for sugar.

There is also an innate tendency to dislike bitter tastes, and there are probably innate predispositions to reject very strong tastes including oral irritation (as from peppers) and to like fatty textures. Second, there is a suspicion about trying new foods (on account of potential toxicity), but also a conflicting interest in them (on account of their potential as a new nutrient source). This has been described as the generalist's dilemma. Third, there is a special learning mechanism that allows learning about the consequences of ingestion, even when these may occur hours after ingestion. Fourth, there are a few specific internal states that seem to signal the need for specific nutrients. The best example is thirst. Hunger, of course, signals a general need for energy, and there may be a specific internal signal associated with sodium deficiency.

Humans are mammals, and hence have a unique first food, milk. Humans and other mammalian generalists make a transition from this single food to a very wide range of foods. This transition is made under the guidance of parents and other caretakers, and usually consists of a scheduled and graduated introduction of new foods. For all mammal species except a subset of modern humans, milk is a food limited to infancy. The development of dairying afforded the opportunity for humans to exploit milk and its products as adult foods.

2. Determinants of Human Food Choice

For humans, food choice is accounted for by some combination of biological, psychological, and cultural factors. The role of biological factors is limited for two reasons. One, as discussed above, is the human generalist orientation to food. Second, the modern

developed-culture world presents to its members a food world that is vastly different from the ancestral world in which the biological predispositions evolved. A shortage of food has been replaced by a surplus of food. A modest range of natural, local foods has been replaced by an enormous range of foods, representing most of the world's cuisines at a neighborhood supermarket. A major risk of food poisoning from toxins and microorganisms has been sharply reduced; the longer lifetime that has resulted from this and other medical advances establishes a new and more subtle link between degenerative diseases and dietary patterns over decades. Limited knowledge about foods and their consequences is replaced by exposure through the media to massive amounts of evidence relating dietary pattern to wellness and disease. And finally, the linkage between energy expenditure and finding food (or finding anything else) has been almost eliminated. Thus, while the ancestral food environment focused on shortage and short-term consequences of eating, the modern environment stresses overabundance and long-term consequences.

Culture is the most powerful determinant of human food choice. More can be learned about an individual's food attitudes and preferences from his or her cultural identity than from any other single piece of information. Food choice is highly constrained by availability and cost, and both of these are primarily determined by culture. In addition, since availability and cost control exposure to food, they indirectly affect the developing of liking for foods, since that is substantially determined by exposure.

Cost and availability aside, human preferences and food attitudes can be framed by a psychological (as opposed to nutritional) taxonomy of foods, developed by Paul Rozin and April Fallon. A basic distinction can be drawn between preferences based on liking, that is on the intrinsic, sensory properties of foods, and instrumental preferences, based on the consequences of eating particular foods. There appear to be three types of reasons for rejecting or accepting a food: sensory-affective, that is, how pleasant a food tastes and smells; anticipated consequences, what the expected consequences are of eating a food; and ideational, what is known about a food (for example, where it comes from, what the nature of it is).

Food rejections can be understood in terms of the selection and interplay of these reasons. One category of rejections is called 'distaste.' These are entities rejected because of negative sensory-affective properties, such as lima beans, broccoli, beer, or chili pepper, for those who find these foods distasteful. A second category is 'danger.' These are things rejected primarily because they are believed to be harmful, because of acute or long-term consequences. The emotion of fear is often associated with their consumption. A third category of rejected things—the largest category—is called 'inappropriate.' These are things that the culture labels as inedible, such as

pencils, grass, paper, or cloth. They might taste good, and might be harmless, but they are rejected for ideational reasons. The fourth category is 'disgust.' Rejection of food as disgusting is, like the previous category, based on ideational, culturally transmitted information but, unlike inappropriates, there is a strong belief that disgusts taste bad and are harmful. Unlike the affectively neutral response to inappropriates, the response to disgusts is strongly negative and emotional. Disgust is the most powerful reaction people have to food. Disgusting entities are so powerful that if they touch an otherwise acceptable food, they render it undesirable, disgusting, and inedible (the principle of contamination or contagion; see *Disgust, Psychology of*).

On the positive side, there are four comparable categories. Good taste (acceptance principally because of sensory properties), beneficial (acceptance largely because of consequences), appropriate (acceptance because it is culturally designated as food, or food for a particular occasion), and transvalued (food enhanced because of its prior history). The transvalued category is much weaker and smaller than the disgust category, in most cultures. In Hindu India, food that has been 'shared' with the gods (via donation to the priests in the temple, and then returned, in part), called *prasad*, is an example of transvaluation.

3. The Acquisition of Food Preferences

For newborn infants, the only functioning categories are good taste (e.g., sweet) and distaste (e.g., bitter). Generally, infants will place in their mouth anything that might fit in, including feces and potentially toxic foods. They reject only on the basis of sensory properties. Gradually, they acquire distastes or good tastes, and learn about danger and beneficial foods. It is much later that the full manifestations of disgust appear. The full adult categorization is in place by roughly five to eight years of age.

Most is known about the distinction between distaste and danger. When ingestion of a food is followed by nausea, it tends to become disliked, that is, a distaste. However, if ingestion of a food is followed by most other negative symptoms (e.g., lower gut pain, skin rash, respiratory distress), the food typically becomes a danger. That is, people reject a food that has caused such symptoms, but it does not usually become a disliked taste. This distinction has also been demonstrated in the laboratory with rats. With respect to affect and emotion, it is notable that dislikes (distastes) and dangers have very different properties, although the outcome (rejection) is the same. The nausea-based acquired distaste (often called a conditioned taste aversion), unlike dangers, is not based on a legitimate sense of danger. Even if a person knows that the nausea/upper gastrointestinal illness was not produced by the food, the aversion remains.

Thus, people who get nauseous and often vomit after a meal usually develop an aversion to some food in the meal, even if they know that the illness was simply the onset of influenza.

The acquisition of good tastes is more complex and less understood. Mere exposure to a food, in itself, often seems sufficient to produce an acquired like. In addition, the pairing of a food with an already positive event (an already-liked food mixed with it, positive regard by a respected person, a pleasant environment), by a process called evaluative conditioning, can lead to acquired likes (or acquired dislikes, if the paired events are negative, as in conditioned taste aversions). Leann Birch has demonstrated that indications of liking by a significant other (peer, older child, teacher, parent) may cause acquisition of liking. The process at work here is not understood. It could be a form of evaluative conditioning. But it also may involve an important instance of communication of affect or emotion. The expressed pleasure by a significant other, on consuming a food, may directly induce a pleasant state in an observer, or it may induce a mimicked positive facial expression. Either of these responses may cause enhancement of liking. Birch has demonstrated, in the laboratory, that efforts by adults to promote liking by emphasizing the beneficial consequences (better health, a specific reward for eating) seem to block the acquisition of liking for the food. It seems that when a child observes respected others enjoying a food, this promotes liking; when she is rewarded for consuming it, this seems to block the acquisition of liking.

The acquisition of knowledge of appropriate and inappropriate foods seems to be largely cognitive, and affect free. The affect-laden acquisition of disgust contrasts with this. Some communication of affect (facial and other) is almost certainly involved, but there are also important cognitive aspects; it is the nature of the foods and their history that is central to this category. Disgusting foods have contamination properties. This property requires a realization that appearance does not equal reality: a contaminated food looks like any other food, but is distinguished by its history. Young children have difficulty making the appearance–reality distinction, and it is only by age four to seven that we see signs of the ability to appreciate contamination, and hence disgust.

Individuals in specific cultures develop likings (the good taste category) for some foods that are innately unpalatable (e.g., bitter, very strong tasting, irritating). Innately unpalatable foods are typically among the favorite foods: chili pepper, black pepper, ginger and other irritant spices, coffee, bitter chocolate, tobacco, alcohol, burnt foods, and highly salted foods. Such reversals of innate aversions are common in humans, and rare in animals. We do not know how these preference reversals occur. In part, it may be through the same processes that produce normal likes: mere exposure, evaluative conditioning, and social approval. But there may also be special mechanisms that

are involved in preference reversals. One is that innately negative foods may come to produce the secretion of endorphins, chemicals secreted in the brain in response to pain and irritation. These may modulate the displeasure, and like their pharmacological analog, morphine, can produce pleasure. Normally, one ceases to interact with a bad-tasting food. However, cultural forces continually reintroduce innately unpalatable foods to the child. It may be that with repeated exposures, the body endorphin response becomes stronger, and eventually cancels and then overwhelms the displeasure. Another account is that it is the very displeasure that is the source of the pleasure. This places liking for innately unpalatable foods in the same category as thrill-seeking activities such as roller coaster riding. That is, humans get pleasure out of situations in which innate aversions or fears are stimulated, but in which there is no real danger. Roller coasters are safe, and so is chili pepper. This may be case of pleasure derived from ‘mind-over-matter.’

4. *Family Influences and the Family Paradox*

There are strong arguments for high parent–child resemblance in food habits. There is the common genetic heritage, the fact that parents control access (and hence exposure) to foods, and that the principal affective signals about foods in early life come from the parents. Yet parent–child correlations in liking for foods are very low, in the range of 0 to 0.3, in comparison to parent–child correlations in values (such as attitudes to abortion), temperament, or abilities. These low parent–child correlations appear whether the children studied are four-year-olds or college students. This family paradox cannot presently be explained. However, it suggests that the second six years of life, when children are much more heavily influenced by peers and teachers, as opposed to parents, may be more critical for establishing adult preferences than the first six years.

5. *Preferences and Values: Moralization and Vegetarianism*

Selecting and consuming a food has social and moral implications, as well as physiological consequences. The source of a food (e.g., who prepared it, animal or vegetable origin) can powerfully influence acceptance. Particular food or food product boycotts in recent decades testify to this fact. Hindu Indian concerns about food fall largely within the moral domain. In the United States, there is a long history linking particular foods with the good life or immorality. Through the process of moralization, choice of a particular food may move from the domain of preferences to the domain of values. This has clearly happened with cigarette smoking in the United States, and it fre-

quently happens in the course of developing vegetarianism. For some vegetarians, eating of animals is an immoral act. When ingestion of a food becomes moralized, censure of consumers becomes licensed and governments and institutions may take limiting or prohibiting actions. Such foods often become offensive (disgusting), and the rejection becomes intrinsic as opposed to instrumental.

6. Food, Fear, and Pleasure: Cultural Perspectives

In the non-human animal world, and among traditional humans, food is basically a source of pleasure. Even in the developed world, cost and availability aside, the principal factor influencing food choice is taste, that is to say, pleasure. Convenience and health issues are also important determinants. However, with the major changes in the human environment in the First World, obesity and the consequences of long-term exposure to particular diets have become more salient. As a result, the earlier simple pleasure of eating has given way, for many people, to ambivalence, or even outright fear. This is most well developed in American women, who often fear food as much as they savor it. In some, there is now embarrassment and guilt about consuming high fat foods, great concern about thinness, and great concern about the healthfulness of every bite of food. This is not a necessary consequence of the modern environment since, in France, the pleasures of eating still dominate.

See also: Disgust, Psychology of; Eating Disorders, Determinants of; Genetic Aspects; Food in Anthropology; Food Production, Origins of; Hunger and Eating, Neural Basis of; Taste and Smell, Psychology of; Taste and Taste Culture; Taste, Neural Basis of

Bibliography

- Barker L M (ed.) *The Psychobiology of Human Food Selection*. AVI, Bridgeport, CT
- Birch L L, Fisher J O, Grimm-Thomas K 1996 The development of children's eating habits. In: MacFie H, Meiselman H (eds.) *Food Choice, Acceptance and Consumption*. Blackie, Glasgow, UK, pp. 161–206
- Booth D A 1994 *Psychology of Nutrition*. Taylor & Francis, London
- Diamond J 1999 *Guns, Germs, and Steel*. Norton, New York
- Fischler C 1990 *L'omnivore*. Editions Odile Jacob, Paris
- Kass L 1994 *The Hungry Soul*. Free Press, New York
- Logue A 1991 *The Psychology of Eating and Drinking*, 2nd edn. Freeman, New York
- MacFie H, Meiselman H (eds.) 1996 *Food Choice, Acceptance and Consumption*. Blackie, Glasgow, UK

- Rozin P 1982 Human food selection: The interaction of biology, culture and individual experience. In: Barker L M (ed.) *The Psychobiology of Human Food Selection*. AVI, Bridgeport, CT, pp. 225–54
- Rozin P 1990 The acquisition of stable food preferences. *Nutrition Reviews* **48**: 106–13
- Rozin P 1998 *Towards a Psychology of Food Choice*. Danone Chair Monograph. Institut Danone, Brussels, Belgium
- Rozin P 1999 Food is fundamental, fun, frightening, and far-reaching. *Social Research* **66**: 9–30
- Rozin P, Fallon A E 1987 A perspective on disgust. *Psychological Review* **94**: 23–41
- Shepherd R, Raats M M 1996 Attitudes and beliefs in food habits. In: MacFie H, Meiselman H (eds.) *Food Choice, Acceptance and Consumption*. Blackie, Glasgow, UK, pp. 346–64

P. Rozin

Food Production, Origins of

A bird's eye view of the human career from inception to now reveals four major landmarks: origins of hominids; origins of *Homo sapiens sapiens*; origins of food production; origins of urbanized, state-based civilizations. Like the other categories of origins research, the emergence of the first farmers and herders has long attracted attention from archaeologists, anthropologists, historians, and other scholars. Early in the twentieth century, a prominent archaeologist, Vere Gordon Childe, discussed the origins of agriculture in arid western Asia at the end of the Ice Age (Childe 1936, Chap. V 'The Neolithic Revolution,' and 1942, Chap. III 'Neolithic Barbarism'). It was not Childe, however, but another archaeologist who set himself the task of actually digging up evidence for the food producing revolution in western Asia.

1. The Origins of Food Production in Western Asia

Shortly after the end of World War II, Robert J. Braidwood (1960) organized the Iraq-Jarmo Project aimed at excavating the oldest agropastoral community known anywhere in the world at that time: Jarmo in Iraqi Kurdistan. Braidwood's expeditions to Jarmo between 1948 and 1955 are famous for their interdisciplinary nature, and for his suggestion that food production in western Asia began in environments like that of Jarmo, a situation he described as 'the hilly flanks of the Fertile Crescent.' He was referring to the foothill and piedmont zone, extending from the Levantine coast along the southern slopes of the Taurus and down the westerly face of the Zagros Mountains, where wild species ancestral to the earliest plant and animal domesticates were distributed.

Following Braidwood's example, interdisciplinary, international archaeological teams undertook intensive research on agricultural origins in the hilly flanks and other regions of western Asia. The work is ongoing today wherever political conditions permit; hence the evidential base is continually changing. At present, however, it appears that the first recognizably domestic food plants are small grains (rye, wheat, barley) and legumes (lentils, peas, vetches) from sites in the Levant, and in the hilly flanks zones of northwest Syria, southeastern Anatolia, and northern Iraq. Archaeobotanist Daniel Zohary (1996) thinks that most of the founder crops (wheat, barley, and legumes) were domesticated only once or twice, and then spread both east and west. The first domesticated food animals are sheep and goats from sites in the Zagros mountain portion of the hilly flanks, pigs and cattle appearing later. Although the archaeological and paleoenvironmental data are by no means unambiguous, most scholars invoke climatic factors prominently (especially fluctuations between warmer and colder intervals in the aftermath of the last Ice Age) to explain how and why food production arose in western Asia (for further details see Bar-Yosef and Meadow 1995, Miller 1992, Harris 1998, Hillman 2000).

2. The Origins of Food Production in other World Areas

Systematic research on agropastoral origins is now widespread around the globe. Interdisciplinary archaeological projects are recovering evidence for early food production in Africa, South and East Asia, New Guinea, and the Americas. At present, it appears that there may have been no more than seven independent trajectories to agriculture, or agropastoralism during the Early and Middle Holocene (i.e., between 10,000 and 5,000 years ago). A brief global survey follows.

2.1 Africa

The lower and middle Nile Valley portion of northeastern North Africa is linked to agropastoral developments in western Asia. Domestic species of wheat, barley, sheep, and goat were derived from the Levant by about 7,000 years ago, and were combined with the earlier gathering, hunting, fishing economy. It is likely, however, that cattle were domesticated in the eastern Sahara from local populations of wild *Bos primigenius* by 9,000 years ago. Many of the early herders of cattle, sheep, and goats were probably also systematically harvesting wild species such as sorghum, *Cenchrus* (krem krem), and the oil plant *Canarium*, but in ways (e.g., beating into trays or baskets) that did not result in morphologically identifiable domesticates.

On the present scanty evidence, there are no domesticated plants in Africa outside the Nile Valley prior to about 4,000/3,500 years ago. Domestic pearl

millet (*Pennisetum*) of that date is reported from archaeological sites across central Africa, probably introduced by Saharan herders. Other plants believed to have been first domesticated in Africa include African rice, sorghum, finger millet, tef, enset, and noog. Domestic sorghum and rice are present at the site of Jenne-Jeno in Mali about 2,000 years ago, whereas finger millet, tef, noog, and enset are East African plants first domesticated in the Ethiopian/Eritrean region (by ca. 1,200 years ago for finger millet; no data as yet for tef, enset, and noog).

There are several other plants, including a number of roots and tubers, that were used intensively, extensively, or both in various parts of sub-Saharan Africa, with little or nothing in the way of archaeological histories as yet. In South Africa, agriculture and pastoralism are thought to have arrived in the early Iron Age about 2,000 years ago with the spread of food producers into this part of the continent (for further details see Harlan 1992, Marshall 1998, Wetterstrom 1993, 1998).

2.2 South Asia

The origin and spread of agriculture and pastoralism in South Asia are complex, long-term processes beginning about 9,500 years ago. Most experts think domesticated wheat and barley as well as sheep and goats were introduced from western Asia, but humped cattle (*Bos indicus*) are thought to have been domesticated indigenously by 7,000 years ago, as perhaps were water buffalo.

Agriculture in the northerly portion of South Asia was altered significantly beginning about 4,500 years ago when a series of summer crops were introduced and grown in various combinations with the winter wheat and barley: finger millet, pearl millet, and sorghum, all from Africa; broomcorn millet, foxtail millet, and rice, all probably from East Asia. One other millet was apparently indigenously domesticated by about this same time: little millet (for further details see Meadow 1998).

2.3 East Asia

A great deal of archaeological work has been carried out at early village sites in China since World War II, but only very recently have explicit, systematic efforts been made to obtain archaeobotanical and zooarchaeological materials from them. Most of the present evidence comes from South China and centers upon rice. Of the 14 direct (AMS) radiocarbon dates on rice remains reported by Crawford and Shen (1998), the oldest is 9,000 years bp ('bp' = before present; this is a calibrated date) at the site of Hemudu (on the east coast of China south of the Yangtze delta) with 9,000 to 8,000bp determinations on rice grains at Pengtoushan (in Hunan Province), and the northerly site of

Jiahu (Henan Province). Crawford and Shen say that the people who lived in these three early sites were harvesting rice, but whether it had been planted, or was being reaped from wild stands is unknown. In any case, it appears that rice was probably first domesticated somewhere in the general region of the middle Yangtze River drainage, and domestic rice may have spread to Korea and Japan between 6,000 and 3,000 years ago.

The other indigenously domesticated Chinese grain crops are foxtail millet (believed to have been domesticated by ca. 7,500 (calibrated) years ago at one site in Hebei Province, and at three sites in Henan Province); and broomcorn millet, domesticated prior to 8,500 years ago in the Yellow River drainage. On the present evidence, rice use (and domestication) in the central Yangtze drainage of southern China seems to predate the appearance of domestic millet in the Yellow River drainage.

Even less is known about the origins of domestic food animals in China than about domestic plants, but dogs, pigs, and probably domestic chickens seem to be present in North China sites at about the same time as the millets. Although the data are not very detailed, the origin of food production in China appears to have entailed processes somewhat more like those of western Asia (i.e., a suite of founder species, which then spread widely) than like the patchier, mosaic pattern of Africa and South Asia (for further details see Cohen 1998, Crawford 1992, Crawford and Shen 1998).

2.4 Southeast Asia and the Islands of the Southwest Pacific

Although much of the older literature on the development of Southeast Asian agriculture focuses heavily on rice (now thought to have diffused there from China's central Yangtze River region), there was a complex and important array of early tuber, root, and tree crops in Southeast Asia, New Guinea, and the islands of the southwestern Pacific. Evidence of these is difficult to retrieve archaeologically, and much of the discussion about them revolves around ethnographic, ethnobotanic, and linguistic data as well as delineation of prehistoric agricultural terraces, ditches, and other landscape features. A strong case has been made for landscape modification in New Guinea, presumably in the interests of growing taro and native yam species, long preceding the ca. 3,500 bp expansion of people from Southeast Asia onto the Pacific islands, accompanied by domestic animals (probably dogs, chickens, and pigs) and an array of domesticated tuber and root species (see *Pacific Islands, Archaeology of*). There is, however, considerable disagreement about the exact nature of pre-3,500 bp highland New Guinea agriculture, and how significant it was prior to the arrival of Southeast Asian derived agricultural systems. (Hence, although some scholars regard New Guinea as one of the world

areas that witnessed an independent transition to agriculture in prehistory, that claim is still being debated. For further details see Bellwood 1996, Golson 1989, Hather 1996, Spriggs 1996).

2.5 The Americas: Mexico

Maize is the most famous crop of the Americas, and by far the most complex of the world-class grains domesticated in prehistory. Although its wild ancestor (teosinte) and its origin place (the Balsas River drainage in western Mexico) are generally agreed upon, the exact details and the chronology of initial domestication are still under debate. AMS dates on early maize from excavations directed by Richard MacNeish in the 1960s at two rockshelters in the Tehuacan Valley, Puebla, and from two other rockshelters in the Ocampo region of Tamaulipas have established that none of these specimens is more than about 5,000 years old, previous estimates having ranged from 7,000 to 5,000 bp or even older. Some researchers, persuaded by seemingly early pollen and phytolith data, continue to advocate very early estimates for domestic maize in spite of concerns by other scholars about the dating of these microremains. Similarly, there is lack of consensus on how, and how rapidly, domesticated maize spread north and south from its place of origin.

Archaeological work at the Mexican sites noted above, plus excavations by Kent Flannery in 1966 at another dry rockshelter—Guila Naquitz in Oaxaca—have resulted in information about several other cultigens: squashes, bottle gourd, beans, and chili peppers. Also present at some of these sites are *Amaranthus*, *Chenopodium*, *Agave*, and several root crops and tree fruits thought to have been encouraged and harvested if not actually domesticated in early prehistory. Recent AMS dates on squashes from MacNeish's excavations in Tehuacan and Tamaulipas (*Cucurbita pepo*, *C. argyrosperma*, *C. moshata*) and from Flannery's work at Guila Naquitz (*C. pepo* gourds) have returned 9,000 bp determinations for *C. pepo* there, and 4,000 to 5,000 bp for the other two cucurbits. No fieldwork on early Mexican agriculture comparable to MacNeish's and Flannery's has taken place since the 1960s (for further details see Fritz 1994, Piperno 1994, Smith 1997a, 1997b).

2.6 The Americas: Central and South America

Origins research on early agriculture in the American lowland neotropics is being actively pursued, with new information emerging at a rapid rate. The foci of attention in the lowlands are tubers and root crops (including potatoes, sweet potatoes, yams, and manioc), as well as tree fruits, pineapple, chili peppers, avocados, maize, squashes, beans, cotton, and bottle gourd. Because of preservation problems, macro-

botanical remains are less likely to be recovered from humid regions than from arid locales like those of the Mexican rockshelters. Hence, several researchers have turned to pollen, phytoliths, and starch grain analyses. This is important, innovative work, but does pose serious problems for secure dating, and—because these trajectories are so new—for consensus in identifications and interpretations.

The Andean highlands of South America are very different from the tropical lowlands, but similar in that the origins of food production there are archaeobotanically documented only preliminarily). The crops centrally involved are *Chenopodium quinoa*, squashes and gourds, beans, cotton, maize, and several tuber species including the potato, *Solanum tuberosum*. Early animal domesticates are llamas, alpacas, and guinea pigs. Chronologies and other details are not precise, but quinoa may have been domesticated by 5,000 to 4,000 years ago, gourd and squash perhaps somewhat earlier, common beans by 4,400 bp (calibrated). There are no direct AMS dates for prehistoric finds of possibly domestic potatoes, and the chronology for llama/alpaca and guinea pig domestication is also vague. Estimates for domestication of all three animal species range around the same time as for quinoa, and the potato may have been domesticated about the same period or even earlier.

The arid west coast of South America in Peru and Chile has its own food-producing story to tell, related to that of the highlands in some ways, but with significant differences. Preservation of even very delicate plant parts as well as textiles and human remains makes the archaeology attractive to professionals and to looters. Evidence for plant use begins early at some sites, perhaps as far back as the Early Holocene for gourds, and includes many of the same species referred to above for other areas of South America. The coastal valleys of Peru currently provide the oldest directly dated evidence for beans (one lima bean pod) at ca. 5,600 bp (calibrated); Mexican common and tepary beans are no older than 2,500 bp (calibrated) (for further details see Pearsall 1992, 1995, Piperno et al. 2000, Piperno and Pearsall 1998).

2.7 The Americas: Eastern North America, with Comparative Reference to the Southwestern United States

Since 1980, archaeological and paleoethnobotanical research has shown that an indigenous agricultural system was created in Eastern North America beginning about 4,500 years ago. Sunflower, sump weed, goosefoot, may grass, bottle gourd, and pepo gourd are the initial members of the complex at 3,000 bp, with knotweed, little barley, and tobacco being added several hundred years later. Maize first appears in Eastern North America 2,000 years ago, but is a very minor component of the archaeobotanical record until

ca. 1,200 bp when it became a major staple in communities of the Ohio and Mississippi River drainages. Common beans enter the picture even later (ca. 700 bp). The two gourds are present in a few places 7,000 years ago—well before the rest of the complex appears—but it is not clear whether they were domesticated that early, or were simply being harvested from wild plants. In any case, they were probably domesticated separately from parallel processes in Mesoamerica and South America, just as *Chenopodium* seems to have been domesticated separately in Eastern North America, Mexico, and South America. Fritz (1999) has argued persuasively that the early solo presence of cultivated or domesticated gourds of whatever kind does not mean the gourd cultivators were farmers. Full-scale agriculture may appear much later, as seems to be the case in several places within the Americas.

The archaeobotanical database in Eastern North America—arguably the best in the world—demonstrates use of wild plant and animal foods throughout the entire sequence from early Middle Holocene to the time of European contact. Domesticates appear in differing combinations and at different levels of intensity across space and time in an intricate mosaic.

Formerly, it was thought that the Mexican agricultural triad of maize, squash, and beans diffused north into the southwestern US, then much later moved across the Plains to Eastern North America. It is now abundantly clear, however, that the Southwest and the East were on separate trajectories, each area experiencing its own complex set of developments. Maize did arrive in the Southwest earlier than it appeared in the East (pre-3,000 bp vs. 2,000 bp) with squash and beans arriving perhaps about the same time, but these plants were adopted into very different kinds of pre-existing Archaic societies across the region: some small, mobile, and staying that way, others already somewhat sedentary and becoming more so. In all cases, considerable use was made of indigenous southwestern plants, some domesticated, others semi-domesticated or encouraged, and others harvested from wild stands. The only domesticated mammal was the dog, but turkeys were kept in some places. Hunting of deer, antelope, rabbits, and other small game was very important everywhere throughout the entire prehistoric sequence (for further details see Fritz 1990, Minnis 1992, Smith 1992, Yarnell 1994).

3. Discussion and Conclusions

There is a vast literature on agricultural origins and the development of prehistoric food-producing economies. The actual corpus of empirical, archaeological, archaeobotanical, zooarchaeological evidence about early food production is much, much smaller. Only in Western Asia and Eastern North America is there sufficient data to constrain conjecture in any signifi-

cant way. Nevertheless, it is clear that humans, plants, and animals everywhere were always interdependent in multiple ways, and that there were many specific pathways to food production. The areal sketches provided in the previous section indicate how diverse those pathways could be. The present evidence, viewed from a high level of abstraction, suggests that there were seven world areas where indigenous and distinctive food-producing economies were developed: Western Asia, Saharan and sub-Saharan Africa (perhaps with West African and East African foci), East Asia (perhaps with separate foci in North China and South China), Eastern North America, Mexico, Central and South America (with neotropical and highland foci), and—intriguing but enigmatic—highland New Guinea.

Discussion of early crops in all these places centers heavily upon those that dominate modern world economies (wheat, rice, maize), sometimes to the detriment of understanding the nature and functioning of ancient subsistence systems, whether food producing or not. Given present concerns about the grave dangers of monocropping on an areal, let alone a global, scale, this tendency is doubly unfortunate. Lost crops, lost breeds, and lost varieties of current crops and domestic animals are vital resources for modern agropastoralism, but can be retrieved only through the interdisciplinary work of archaeologists, paleoethnobotanists, zooarchaeologists, on the one hand, and molecular biologists and plant and animal breeders on the other.

Two more issues central to any consideration of ancient food-producing economies are chronology and explanation. Was the origin of food production a nearly simultaneous event around the world triggered by Late Pleistocene/Early Holocene climatic change? Were the causes and processes of all six or seven independent, indigenous transitions to food production essentially the same? How one answers these questions obviously depends almost entirely on how detailed and how secure areal chronologies and areal databases may be. There is, at best, only semi-security for two of the pristine-origins candidates. Nevertheless, one can argue a negative response to both questions. On the evidence of present chronologies, created and continuously refined via heavy dependence on AMS radiocarbon dating, it appears that only in West Asia and East Asia was there significant impact by end Pleistocene/Early Holocene climatic changes to impel intensification of the food-producing sort. In both places, food production economies seem to be phenomena of the earliest Holocene, taking shape in the first two millennia after the end of the last ice age, and spreading relatively rapidly, almost in package form. Elsewhere, when sufficient information is available, chronologies for the origins of food production are longer, with food-producing systems being more fluid and patchy or mosaic-like, varying considerably through time and across space. David

Harris (1996, pp. 557, 568) contrasts the temperate West Asian package (small grains, legumes, sheep, goat) with the vegeticultural systems (roots, tubers, rhizomes) nutritionally (the latter is heavy on carbohydrates, the former also includes fat and protein) and territorially (the former is expansive, the latter is not). Somewhat similarly but from a global perspective, Jared Diamond (1997) points out the great contrasts among geography, flora, and fauna that contextualize subsistence systems in different parts of the world.

Prime movers used to aid explanation of transitions to food production range from the older favorites, climatic or environmental change (see above) and population pressure (e.g., Cohen 1977) to newer models drawn from economics and evolutionary ecology, to competition among lineage or clan heads, or to the necessity of intercommunity trade for forging alliances guaranteeing mutual assistance in hard times (e.g., Bender 1978, Hayden 1996, Price and Brown 1985). Some of these models are quite ingenious, and plausible for specific origins cases, but they do not fit others at all, making them less than satisfactory as general explanations for *the* origin of food production in the human past. That abstraction is very useful for certain purposes, but self-destructs when placed in contact with all the evidential bases now available, meager though many of them are. Although we do not know nearly enough, we do know too much, and what we know is changing too rapidly, to sustain current general explanations for why food production happened in prehistory. We know, for example, that gatherer–hunter–fisher–foragers always manipulated and managed their physical and biological environments, that they often cultivated and/or harvested wild plants, tamed wild animals, and restructured the topography as well as the native plant and animal distributions in their home territories; and that they were quite capable of intensifying any or all of these management procedures if they perceived intensification to be necessary or desirable. We know that in some places and times the kind of intensification we call food production was implemented, but that mostly it was not; that gatherer–hunters could and did coexist for long periods (centuries, millennia) with food producers; and that food producers in some times and places became gatherer–hunters as well as vice versa.

In spite of this chronic mismatch between theory and data, inevitable in a field as broad and lively as agricultural origins research, it is essential that high-level, theoretical formulations be devised and tested against independent relevant data, explicitly sought if not already available, because that is the most efficient means to expand knowledge about what, in the abstract, is indeed the most important transition in human prehistory.

See also: Agricultural Change Theory; Cultural Evolution: Overview; Food in Anthropology; Food Security; Hunter–Gatherer Societies, Archaeology of;

Hunting and Gathering Societies in Anthropology; Pacific Islands, Archaeology of; Pastoralism in Anthropology

Bibliography

- Bar-Yosef O, Meadow R H 1995 The origins of agriculture in the Near East. In: Price T D, Gebauer A B (eds.) *Last Hunters, First Farmers: New Perspectives on the Prehistoric Transition to Agriculture*. SAR Press, Santa Fe, NM
- Bellwood P 1996 The origins and spread of agriculture in the Indo-Pacific region: Gradualism and diffusion or revolution and colonization? In: Harris D R (ed.) *The Origins and Spread of Agriculture and Pastoralism in Eurasia*. UCL Press, London
- Bender B 1978 From gatherer-hunter to farmer: A social perspective. *World Archaeology* 10: 204–22
- Braidwood R J 1960 The agricultural revolution. *Scientific American* 203: 130–48
- Childe V G 1936 *Man Makes Himself*. Watts, London
- Childe V G 1942 *What Happened in History*. Penguin, Harmondsworth, UK
- Cohen D 1998 The origins of domesticated cereals and the Pleistocene–Holocene transition in East Asia. In: Bar-Yosef O (ed.) *The Transition to Agriculture in the Old World. The Review of Archaeology* 19(2): 22–9
- Cohen M 1977 *The Food Crisis in Prehistory*. Yale University Press, New Haven, CT
- Crawford G W 1992 Prehistoric plant domestication in East Asia. In: Cowan C W, Watson P J (eds.) *The Origins of Agriculture, an International Perspective*. Smithsonian Institution Press, Washington, DC
- Crawford G W, Shen C 1998 The origins of rice agriculture: Recent progress in East Asia. *Antiquity* 72: 856–66
- Diamond J M 1997 *Guns, Germs and Steel: The Fates of Human Societies*. W W Norton, New York
- Fritz G J 1990 Multiple pathways to farming in precontact Eastern North America. *Journal of World Prehistory* 4: 387–435
- Fritz G J 1994 Are the first American farmers getting younger? *Current Anthropology* 35: 305–9, 639–43
- Fritz G J 1999 Gender and the early cultivation of gourds in Eastern North America. *American Antiquity* 417–29
- Golson J 1989 The origins and development of New Guinea agriculture. In: Harris D R, Hillman G C (eds.) *Foraging and Farming: The Evolution of Plant Exploitation*. Unwin Hyman, London
- Harlan J R 1992 Indigenous African agriculture. In: Cowan C W, Watson P J (eds.) *The Origins of Agriculture, an International Perspective*. Smithsonian Institution Press, Washington, DC
- Harris D R 1996 The origins and spread of agriculture and pastoralism in Eurasia: An overview. In: Harris D R (ed.) *The Origins and Spread of Agriculture and Pastoralism in Eurasia*. UCL Press, London
- Harris D R 1998 The origins of agriculture in Southwest Asia. In: Bar-Yosef O (ed.) *The Transition to Agriculture in the Old World. The Review of Archaeology* 19(2): 5–11
- Hather J G 1996 The origins of tropical vegetation: Zingiberaceae, Araceae and Dioscoreaceae in Southeast Asia. In: Harris D R (ed.) *The Origins and Spread of Agriculture and Pastoralism in Eurasia*. UCL Press, London
- Hayden B 1996 Pathways to power: Principles of creating socioeconomic inequalities. In: Price T D, Feinman G M (eds.) *Foundations of Social Inequality*. Plenum, New York
- Hillman G C 2000 Overview: The plant-based components of subsistence at Abu Hureyra 1 and 2. In: Moore A M, Hillman G C, Legge A J (eds.) *Abu Hureyra and the Advent of Agriculture*. Oxford University Press, Oxford, UK
- Marshall F 1998 Early food production in Africa. In: Bar-Yosef O (ed.) *The Transition to Agriculture in the Old World. The Review of Archaeology* 19(2): 47–58
- Meadow R H 1998 Pre- and proto-historic agricultural and pastoral transformations in northwestern South Asia. In: Bar-Yosef O (ed.) *The Transition to Agriculture in the Old World. The Review of Archaeology* 19(2): 12–21
- Miller N F 1992 The origins of plant cultivation in the Near East. In: Cowan C W, Watson P J (eds.) *The Origins of Agriculture, an International Perspective*. Smithsonian Institution Press, Washington, DC
- Minnis P E 1992 Earliest plant cultivation in the desert borderlands of North America. In: Cowan C W, Watson P J (eds.) *The Origins of Agriculture, an International Perspective*. Smithsonian Institution Press, Washington, DC
- Pearsall D M 1992 The origins of plant cultivation in South America. In: Cowan C W, Watson P J (eds.) *The Origins of Agriculture: An International Perspective*. Smithsonian Institution Press, Washington, DC
- Pearsall D M 1995 Domestication and agriculture in the New World tropics. In: Price T D, Gebauer A T (eds.) *Last Hunters, First Farmers: New Perspectives on the Prehistoric Transition to Agriculture*. SAR Press, Santa Fe, NM
- Piperno D R 1994 On the emergence of agriculture in the New World. *Current Anthropology* 5: 637–9
- Piperno D R, Pearsall D M 1998 *The Origins of Agriculture in the Lowland Neotropics*. Academic Press, San Diego, CA
- Piperno D R, Ranere A J, Holst I, Hansell P 2000 Starch grains reveal early root crop horticulture in the Panamanian tropical forest. *Nature* 407: 894–7
- Price T D, Brown J A 1985 *Prehistoric Hunter-Gatherers: The Emergence of Cultural Complexity*. Academic Press, Orlando, FL
- Smith B D 1992 *Rivers of Change: Essays on Early Agriculture in Eastern North America*. Smithsonian Institution Press, Washington, DC
- Smith B D 1997a The initial domestication of *Curcubita pepo* in the Americas 10,000 years ago. *Science* 276: 932–4
- Smith B D 1997b Reconsidering the Ocampo caves and the era of incipient cultivation in Mesoamerica. *Latin American Antiquity* 342–83
- Spriggs M 1996 Early agriculture and what went before in Island Melanesia: Continuity or intrusion? In: Harris D R (ed.) *The Origins and Spread of Agriculture and Pastoralism in Eurasia*. UCL Press, London
- Wetterstrom W 1993 Foraging and farming in Egypt: The transition from hunting and gathering to horticulture in the Nile Valley. In: Shaw T, Sinclair P, Andah B, Akpoko A (eds.) *The Archaeology of Africa: Food, Metals and Towns*. Routledge, London
- Wetterstrom 1998 The origins of agriculture in Africa, with particular reference to sorghum and pearl millet. *The Review of Archaeology* 2: 30–46
- Yarnell R A 1994 Investigations relevant to the native development of plant husbandry in eastern North America: A brief and reasonably true account. In: Green W, Arzigian C M

- (eds.) *Agricultural Origins and Development in the Mid-continent*. Office of the State Archaeologist, University of Iowa, Report 19, Iowa City, IA
- Zohary D 1996 The mode of domestication of the founder crops of southwestern Asian agriculture. In: Harris D R (ed.) *The Origins and Spread of Agriculture and Pastoralism in Eurasia*. UCL Press, London

P. J. Watson

Food Security

Food is undoubtedly one of the most basic human needs. Access to enough food to lead an active and healthy life is also one of the most fundamental human rights. Although the global food supply is generally satisfactory, the basic food needs of hundreds of millions of people are not satisfied, and the right to food is widely neglected. This article defines food security, assesses the current and future world food situation, and presents theoretical and conceptual approaches to analyze food security from the perspective of the social and behavioral sciences.

1. Definition of Food Security

Although nutritional scientists distinguish between 'food security' (availability of food on global, national, local, and household levels), on the one hand, and 'nutrition security' (satisfactory nutritional status of individuals), on the other (Oltersdorf and Weingärtner 1996, p. 23), social and behavioral scientists tend to consider 'food security' as a more comprehensive term that incorporates both concepts. This view is evident from most of the definitions available on food security, not only in social science, but also in official GO and NGO documents (see M. Smith et al. 1992). The following World Food Council (WFC) definition may serve as an example:

Food security implies two things. First, it implies that food is available, accessible, affordable—when and where needed—in sufficient quantity and quality. Second, it implies an assurance that this state of affairs can reasonably be expected to continue; or in other words, that it can be sustained. To put it simply, food security exists when adequate food is available to all people on a regular basis (World Food Council 1988, p. 2).

This definition will serve as the base for the following discussion on social and behavioral aspects of food security.

Along the line of this definition, food insecurity implies problems with the availability, accessibility, and affordability of food. This links the concept of 'food insecurity' to the notion of the 'food system'

which is widely used in the discussion of food (in)security (Cannon 1991). Food systems are defined as integrated systems of the four subsystems of production, exchange, delivery, and consumption (Cannon 1991, p. 298).

It is important to distinguish between two kinds of food insecurity, chronic and transitory. The difference between the two is emphasized by a World Bank definition (1986, p. 1):

Chronic food insecurity is a continuously inadequate diet caused by the inability to acquire food. It affects households that persistently lack the ability either to buy enough food or to produce their own. Transitory food insecurity is a temporary decline in a household's access to enough food. It results from instability in food prices, food production, or household incomes—and in its worst form it produces famine.

2. The World Food Situation

A comprehensive assessment of the world food situation by experts from the International Food Policy Research Institute (Pinstrup-Andersen et al. 1997) addresses recent developments, emerging issues, and long-term prospects, and also distinguishes between production, distribution, and consumption issues. On the one hand, in the past couple of years, developments in global food supply, demand, and trade have raised concerns about the world's future food supply. Prices of grain rose rapidly, and at the same time global cereal stocks fell sharply. Many countries of Eastern Europe and the former Soviet Union failed to make significant advances in economic transition and agricultural development. Moreover, flows of food aid and official development finance to developing countries declined. In sub-Saharan Africa, the population growth rate has exceeded the rate of growth in food production considerably. Weather fluctuations and climate change could lead to sizeable food production shortfalls and deterioration in food security in many parts of the world. Declining natural base, first of all growing water scarcity and declining soil fertility, put severe threats on future agricultural development. Stagnating income levels of the poor, high rates of underemployment, and severe cuts in the social security budgets put further stress on the food security of the poor and vulnerable sections of society. South Asia alone is home to around 50 percent of the world's poor and food insecure (ACC/SCN 1997). Moreover, there are escalating concerns about genetic engineering and food quality.

On the other hand, other developments offer glimmers of hope. The United Nations, for example, has once again revised its population projections downward, thus reducing expected pressures on future food supplies. Progress has also been made on international trade liberalization, although developing countries do not equally benefit as developed countries. The World

Food Summit of November 1996 has raised awareness for world food security problems and is expected to stimulate commitments for action.

Projections of food production, distribution, and consumption to the year 2020 offer some signs of progress, but prospects of a food-secure world, a world in which each and every person is assured of access at all times to the food required to lead a healthy and productive life, remain bleak if the global community continues with business as usual (Pinstrup-Andersen et al. 1997, p. 8). Projections for the future world food situation under several scenarios show that, under the most likely scenario, 150 million children under the age of six years will be malnourished in 2020, just 20 percent fewer than in 1993. Projections by the Food and Agricultural Organization (FAO) on the number of food insecure people paint a similarly mixed picture. FAO projects that 680 million people could be food insecure in 2010, down from 840 million in 1990–2. Worldwide, per capita availability of food is projected to increase around 7 percent between 1993 and 2020. However, the future world food situation will be significantly influenced by issues such as rapid urbanization, declining resource base, policy-induced draw down of cereal stocks, and greater variability in agricultural production caused by climate change. Widespread conflict and political instability are increasingly adding to food insecurity in many countries of the world.

To conclude with the IFPRI study (Pinstrup-Andersen et al. 1997, p. 7):

All of these issues suggest potentially larger fluctuations in food production and prices, and higher associated risks of food insecurity for the world's most vulnerable countries and people. The challenge for policy makers, researchers, and others is how best to minimize these risks to achieve food security for all people.

3. Food Security Research and Theoretical Approaches

The scientific debate on the causes of food insecurity was opened by Thomas Malthus (1798) more than 200 years ago. Malthus's basic argument was that the world's natural resources could not assure expansions in food supply that would match population growth. This argument was turned around by Boserup (1965) who postulated that population growth is one of the most powerful driving forces for agricultural development and increased food production.

Experience has proved that Malthus certainly underestimated both technical innovations in agricultural productions and checks in population growth. Although the Malthusian arguments are still prevalent, first of all in the context of the current situation in sub-Saharan Africa, the debate has now shifted away from perspectives of food availability to concerns with food

entitlements. This paradigmatic change can be traced back to Amartya Sen's seminal work on *Poverty and Famines* (1981) which won him the Nobel Prize in Economics in 1998. The entitlement concept has primarily addressed the conditions under which chronic food insecurity collapses into mass starvation and in this sense, it endeavors to provide a theory of food crisis and famine causation. A person must necessarily starve if his/her entitlement set does not include a commodity bundle with enough food; a person is reduced to starvation if his/her endowment or exchange entitlement makes it no longer possible to acquire a commodity bundle with sufficient food (Sen 1990, p. 37).

Sen's entitlement theory can be regarded as a milestone in social science risk research, in general, and in the field of food vulnerability analysis, in particular. However, when addressing issues of vulnerability to food insecurity and famine, recent debates have clearly shown that there can be no single theory of food (in)security, because the problems are simply too complex. Consequently, Watts and Bohle (1993) and Downing et al. (1994) have tried to integrate three theoretical vantage points, namely human ecology, entitlement theory, and political economy, into a comprehensive concept of vulnerability analysis which combines issues of demography, ecology, poverty, empowerment, and long-term structural crisis proneness into one causal model.

A new strand of research is to systematically perceive food systems from two different perspectives, namely an 'external' and an 'internal' side. As Chambers had already pointed out in 1989, vulnerability is:

the exposure to contingencies and stress, and difficulty coping with them. Vulnerability has thus two sides: an external side of risks, shocks and stress to which an individual or household is subject; and an internal side which is defenselessness, meaning a lack of means to cope without damaging loss (Chambers 1989, p. 1).

Contrary to the 'external' side of vulnerability, the 'internal' side of coping is still awaiting a firm theoretical base. Recent empirical research has revealed the enormous complexity in the scope, mechanisms, and efficiency of people's coping strategies for food security (for example, Spittler 1989, Bohle and Adhikari 1998). Thus, concepts are being developed currently that address the structural and behavioral context of coping strategies in more general terms, for example the 'access' model (Blaikie et al. 1994) or the notion of 'assets' as propagated by UNDP (HDR 1997/1999).

4. Future Directions

Although there is a wealth of empirical social science and behavioral studies on food security, their theoretical foundation is still rather weak, especially in the

fields of risk analysis and coping behavior. Moreover, the link between scientific approaches and practical needs is still underdeveloped. The theoretical frameworks are generally not operational for practitioners, although vulnerability analysis as developed by social sciences has become a standard procedure in food security projects and programs. Social and behavioral sciences should systematically provide insights into processes that allow goal-oriented structural changes for enhanced food security (e.g., promoting access and providing relevant assets). Finally, food security concerns have to be integrated more systematically into broader issues of 'human security' (HDR 1997/1999). In the field of transitory food insecurity (food crisis and famines), social and behavioral science has started to develop concepts for early warning and mitigation, such as risk mapping, to contribute to more precise social and geographical targeting procedures for food security. There is still broad scope for social and behavioral science in these fields.

See also: Agricultural Change Theory; Agricultural Sciences and Technology; Agriculture, Economics of; Ecological Economics; Ecology, Cultural; Feminist Political Ecology; Food Production, Origins of; Human–Environment Relationship: Carrying Capacity; Population and Technological Change in Agriculture

Bibliography

- ACC/SCN United Nations Administrative Committee on Coordination. Sub-Committee on Nutrition 1997 Third Report on the World Nutrition Situation. WHO, Geneva
- Blaikie P, Cannon T, Davis I, Wisner B 1994 *At Risk. Natural Hazards, People's Vulnerability, and Disasters*. Routledge, London and New York
- Bohle H G, Adhikari J 1998 Rural livelihoods at risk. How Nepalese mountain farmers cope with food insecurity. In: *Mountain Research and Development*, Vol. 18.4, pp. 321–32

- Boserup E 1965 *The Conditions of Agricultural Growth: The Economics of Agrarian Change Under Population Pressure*. Allen & Unwin, London
- Cannon T 1991 Hunger and famine: Using a food systems model to analyse vulnerability. In: Bohle H G, Cannon T, Hugo G, Ibrahim F N (eds.) *Famine and Food Security in Africa and Asia*. Naturwissenschaftliche Gesellschaft, Bayreuth
- Chambers R 1989 Vulnerability, coping and policy. *IDS Bulletin* 20(2): 1–7
- Downing T E, Bohle H G, Watts M J 1994 Climate change and social vulnerability: Toward a sociology and geography of food insecurity. In: *Global Environmental Change*, Vol. 4, pp. 37–48
- Kasperson J X, Kasperson R E, Turner I B L (eds.) 1995 *Regions at Risk: Comparisons of Threatened Environments*. United Nations University, Tokyo
- Malthus T 1798 *An Essay on the Principle of Population*. London
- Oltersdorf U, Weingärtner L 1996 *Handbuch der Welternährung. Die zwei Gesichter der globalen Nahrungssituation*. Deutsche Welthungerhilfe, Bonn, Germany
- Pinstrup-Andersen P, Pandya-Loren R, Rosegrant M W 1997 *The World Food Situation: Recent developments, Emerging Issues, and Long-term Prospects*. Food Policy Report, IFPRI, Washington, DC
- Sen A K 1981 *Poverty and Famines: An Essay on Entitlement and Deprivation*. Clarendon Press, Oxford, UK
- Sen A K 1990 Food, economics, and entitlements. In: Drèze J, Sen A K (eds.) *The Political Economy of Hunger*. Clarendon Press, London
- Smith M, Pointing J, Maxwell S 1992 *Household Food Security: Concepts and Definitions. An Annotated Bibliography*. The Institute of Development Studies, Brighton, UK
- Spittler G 1989 *Handeln in einer Hungerkrise: Tuaregnomaden und die Große Dürre von 1984*. Opladen, Germany
- UNDP 1997/1999 *Human Development Report*. United Nations Development Program, New York
- Watts M J, Bohle H G 1993 Hunger, famine and the space of vulnerability. *Geojournal* 30(2): 117–25
- World Bank 1986 *Poverty and Hunger*. World Bank, Washington, DC
- World Food Council 1988 *Towards Sustainable Food Security: Critical Issues*. Report by the Secretariat, 14th Ministerial Session, Nicosia, Cyprus, 23–26 May
- World Food Council 1999 *Current World Food Situation*, 116th Session, Food and Agricultural Organization, Rome, 14–19 June

H.-G. Bohle

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

For

Foreign Investment: Direct

Direct Foreign Investment or Foreign Direct Investment (FDI) is defined as the ownership (partial or full) and control of assets in one country by foreign residents. FDI in the USA is defined as the foreign ownership or control of, directly or indirectly, at least 10 percent of the voting securities of an incorporated US business enterprise or an equivalent interest in an unincorporated US business enterprise. Two major forms of FDI are the acquisition of an existing firm or real property or the establishment of new operations. The OECD countries account for most FDI flows in the world. Furthermore, large-scale mergers and acquisitions, particularly between British firms, and US firms have been an important development in the 1990s. The USA is the top ranking country in terms of both FDI inflows and outflows followed by the UK. Other dominant sources of FDI are Canada, Japan and the West European countries. With the exception of Japan, both Canada and West European countries are also the top ranking recipients of FDI. Notable trends in the 1990s include the flow of FDI in East European economies and China in response to national level policy changes towards foreign investments. Another development has been the growth of FDI in East and Southeast Asia in the 1990s by non-Asian firms in response to deregulation, as well as opportunities for low-priced acquisition following the Asian financial crisis. Prior to the Asian crisis, intra-Asian investments dominated FDI flows in East and Southeast Asia.

FDI studies can be classified into two broad categories: (a) analyses of factors affecting the flow of FDI from one country to another country; and (b) the location characteristics of FDI within a host economy. Other important topics include the analysis of the impact of FDI (e.g., employment and technology transfer) in source and host countries. However, due to data limitations, impact analyses are limited to corporate surveys.

Most explanations of FDI have been based on the trans-nationalization process of large USA based multinationals in manufacturing industries (Caves 1971, Hymer 1976, Vernon 1966). In the 1970s and 1980s, the upsurge of FDI from Western Europe, Japan, and the newly industrializing nations, as well as FDI in service industries called for a critique of the existing explanations and further theorization of the determinants of FDI (Dunning 1989, Lall 1983). In

the 1980s, the USA became the top recipient of FDI and several studies focused on the spatial and temporal dynamics of FDI in the USA with a special emphasis on FDI from Japan (Bagchi-Sen and Wheeler 1989, McConnell 1980, O'Huallachain 1985). The worldwide growth of FDI also led to the analysis of multinational activity in other destinations (e.g., the UK, Australia, Eastern Europe, and East and Southeast Asia) (Dicken 1998, Edgington 1990).

1. *Determinants of FDI*

Initially, the principle of comparative advantage was presumed to govern foreign investment flows. However, the neoclassical explanation did not account for the continuation of the bilateral flow of investment among capital-rich nations such as the USA and the countries of Western Europe. The neoclassical explanation of trade also failed to recognize the importance of technology, know-how, and managerial or marketing expertise, as significant determinants of FDI

1.1 *The Industrial Organization Approach*

Hymer's analysis of the trans-nationalization process of firms based on the theory of market imperfections suggested that foreign direct investors must possess specific advantages over competing host country firms (Hymer 1976). These competitive advantages can be derived from the factor market (e.g., discriminatory access to capital, technology, and managerial expertise), product market (through product differentiation and pricing policy), internal and external economies of scale, and government regulations (e.g., limitations on output or entry). These advantages will allow the direct investor to overcome barriers (e.g., the lack of information about local markets, legal issues, infrastructure and logistics, and culture among others) encountered in the host country. Caves (1971) suggested that in order to compete with domestic firms, foreign firms must internalize certain firm-specific advantages, such as patented technology, managerial know-how, marketing strategy, product differentiation techniques and economies of scale. Internalization refers to an extension of firm's activities rather than the externalization of such activities through contracts with independent host country firms. Alternatives to internalization, such as licensing, man-

agement contracts, franchises, and subcontracts are not feasible methods of appropriating firm-specific advantages.

1.2 The Product Cycle Theory

One of the major explanations of the internationalization process of firms as a sequential development is the product cycle theory (Vernon 1966). The product cycle theory explains why market-oriented firms replaced their export markets with direct investment in overseas facilities. The relevance of this theory becomes questionable when a firm's investment approach involves risk diversification through the integration of worldwide operations rather than the dependence on one market alone. Furthermore, some firms do not follow the sequential process and may skip the export phase altogether and undertake a direct investment. In the case of Japanese FDI, some products (e.g., color television) were developed exclusively for the US market. Europeans on the other hand reacted to US investor's pricing strategy in European markets by undertaking FDI in the USA (exchange of threat). The analysis of multinationals from developing countries shows that innovative power and wealth are not necessary preconditions of FDI—these firms often undertake FDI following defensive, imitative, dependent, and opportunist strategies.

1.3 The Eclectic Approach

In recent years, Dunning's eclectic paradigm has been used to explain FDI in both manufacturing and services from all possible sources (Bagchi-Sen and Sen 1997). Ownership, location, and internalization advantages are the three determinants of trans-nationalization (Dunning 1977). Once a firm possesses ownership-specific advantages (knowledge, management-marketing-financial skills, control of resources and markets, and the ability to diversify risks) and successfully internalizes its proprietary know-how, remaining variables that further influence the choice of a destination are location or country-specific characteristics (e.g., market demand, trade barriers). The analysis of firm structure and behavior of US-based, German and Japanese firms by Pauly and Reich (1997) highlights source-country based differences in corporate governance, financing, research and development, direct investment activities and intra-firm trade. The findings emphasize that a global model of internationalization is not the norm.

2. Spatial and Temporal Variations in FDI

Studies on location preferences of foreign investors offer a variety of explanations based on samples drawn from different regions, time periods, and groups of

firms (Hayter 1997). Selected findings from these location analyses over the past three decades are discussed. The most prevalent pattern is the location of foreign firms in the core regions of host countries (Blackbourn 1974, McConnell 1980). These core regions offer economies of agglomeration and lower search cost as well as uncertainty for newcomers. In the USA, for example, the major mode of entry is acquisitions—therefore, core regions are preferred because these regions offer an array of possibilities for mergers and acquisitions. In both developed and developing countries, the large metropolitan centers attract FDI, especially FDI in service sectors (Bagchi-Sen and Sen 1997, Edgington 1990). Latecomers in manufacturing have shown the tendency to locate in emerging industrial spaces (e.g., the preference of foreign manufacturers to locate in the southern and western part of the USA as opposed to the north-eastern and New England states) or sometimes foreign investors create their own industrial corridors in a host country (e.g., the location of Japanese automakers and suppliers in southern Ontario, Michigan, Indiana, Ohio, Illinois, Kentucky and Tennessee in North America). A preference for peripheral areas for new investments is often influenced by the incentive packages offered by state and local governments. Investments by state/local governments to attract Japanese autoproducers ranged between US\$33 million to US\$149 million in the USA. Similar investments have been incurred by local governments in other countries such as Setubal (Portugal), West Midlands and north-east England (UK), and Lorraine (France), among others, (Dicken 1998) to attract automakers from the USA, Korea, and Germany. A further observation is that foreign investors often prefer particular regions based on sector characteristics of FDI, physical proximity of the investor to the host, and personal choice/bias of the decision-maker. For example, Japanese automakers avoided places with existing US auto plants and States where another Japanese autoproducer had located their manufacturing operations.

3. Conclusion

The growth of FDI in the 1990s surpassed the growth of exports in goods and services, world output and gross domestic capital formation. The importance of FDI in host economies cannot be ignored because in several countries FDI's share of gross domestic product is well over 10 percent. Furthermore, the impact of FDI is far reaching. A nation's capital and financial markets, employment pattern, industrial structure and competitiveness, technology, as well as trade and development patterns, are affected by FDI. The nature of the impact depends on the characteristics of FDI (e.g., mode of entry, function, and attributes), the timing of FDI and the bargaining relationship between the foreign entity and the nation state (Cantwell 1989 Dicken 1998).

See also: Development and the State; Development, Economics of; Economics: Overview; Financial Institutions in Economic Development; Industrialization; International Business; International Organization; International Trade: Economic Integration; Multinational Corporations

Bibliography

- Bagchi-Sen S, Wheeler J O 1989 A spatial and temporal model of foreign direct investment in the United States. *Economic Geography* **65**: 113–29
- Bagchi-Sen S, Sen J 1997 The current state of knowledge in international business in producer services. *Environment and Planning A* **29**: 1153–74
- Blackbourn A 1974 The spatial behavior of American firms in Western Europe. In: Hamilton, F E I (ed.) *Spatial Perspectives on Industrial Organization and Decision-Making*. John Wiley, London, pp. 245–64
- Cantwell J 1989 *Technological Innovation and the Multinational Corporations*. Blackwell, London
- Caves R E 1971 International corporations: the industrial economics of foreign investment. *Economica* (New Series) **38**: 1–27
- Dicken P 1998 *Global Shift. Transforming the World Economy* 3rd edn. Guilford, London
- Dunning J H 1977 Trade, location of economic activity and the MNE: a search for an eclectic approach. In: Ohlin B, Hesselborn P O, Wijkman P M (eds.) *The International Allocation of Economic Activity*. Macmillan, London, Chap. 12
- Dunning J H 1989 *Transnational Corporations and the Growth of Services: Some Conceptual and Theoretical Issues*. United Nations, New York
- Edgington D W 1990 *Japanese Business Down Under: Patterns of Japanese Investment in Australia*. Routledge, London
- Hayter R 1997 *The Dynamics of Industrial Location. The Factory, the Firm and the Production System*. Wiley, New York
- Hymer S 1976 *The International Operation of National Firms: A Study of Direct Foreign Investment*. MIT Press Cambridge, MA (Ph.D. thesis 1960, MIT)
- Lall S 1983 *The New Multinationals: The Spread of Third World Enterprises*. Wiley, Chichester, UK
- McConnell J E 1980 Foreign direct investment in the United States. *Annals of the Association of American Geographers* **70**: 259–70
- O'Huallachain B 1985 Spatial patterns of foreign investment in the United States. *Professional Geographer* **37**: 155–63
- Pauly L W, Reich S 1997 National structures and multinational corporate behavior: enduring differences in the age of globalization. *International Organization* **51**: 1–30
- Vernon R J 1966 International investment and international trade in product cycle. *Quarterly Journal of Economics* **80**: 190–207

S. Bagchi-Sen

Foreign Language Teaching and Learning

Foreign language learning and teaching refer to the teaching or learning of a non-native language outside

of the environment where it is commonly spoken. A distinction is often made between 'foreign' and 'second' language learning. A second language implies that the learner resides in an environment where the acquired language is spoken. In the area of research, however, the term second language acquisition is a general term that embraces foreign language learning. Research into the acquisition of a non-native language includes the disciplines of psychology, linguistics, education, neurobiology, sociology, and anthropology. Inquiries of learning and teaching innovations have provided new insights into successful language learning strategies and environments designed to increase language achievement and proficiency.

1. Definition

A language is considered foreign if it is learned largely in the classroom and is not spoken in the society where the teaching occurs. Study of another language allows the individual to communicate effectively and creatively and to participate in real-life situations through the language of the authentic culture itself. Learning another language provides access into a perspective other than one's own, increases the ability to see connections across content areas, and promotes an interdisciplinary perspective while gaining intercultural understandings. Language is the vehicle required for effective human-to-human interactions and yields a better understanding of one's own language and culture. Studying a language provides the learner with the opportunity to gain linguistic and social knowledge and to know when, how, and why to say what to whom (National Standards in Foreign Language Education Project 1996).

Language scholars distinguish between the terms acquisition and learning: 'acquisition' refers to the process of learning first and second languages naturally, without formal instruction, whereas 'learning' is reserved for the formal study of second or foreign languages in classroom settings. One usually distinguishes between the relatively effortless process of second language acquisition by children and the more formal and difficult second language acquisition by adults. Foreign language education refers to the teaching of a modern language that is neither an official language nor the mother tongue of a significant part of the population.

2. Theories of Language Learning

Foreign language learning and teaching have undergone a significant paradigm shift as a result of the research and experiences that have expanded the scientific and theoretical knowledge base on how students learn and acquire a foreign language. Traditionally, learning a foreign language was thought to be a 'mimetic' activity, a process that involved students repeating, or imitating new information. Grounded in

behaviorist theories of learning and structural linguistics, the quality and quantity of language and feedback were regarded as the major determinants of language learning success. A popular method of teaching in the 1950s, called the audio-lingual approach (ALM), promoted an imitation and practice approach to language development. The major figure in the ALM classroom was the instructor who was cast into the role of drill sergeant, expert, and authority figure. Students were relegated to practicing and imitating patterns to a point of automatic response in the belief that the learner would then merely have to slot in lexical items appropriate to the conversational situation. It was believed that the first language interfered with the acquisition of the second language and that a transfer would take place from the first to the second language, resulting in errors. In 1959, Noam Chomsky's review (Chomsky 1959) of B. F. Skinner's (1957) *Verbal Behavior* dramatically changed the way of looking at language by arguing that language was a rule-governed activity, not a set of habits. Chomsky argued that stimulus-response psychology could not adequately account for creativity involved in generating novel utterances using internalized rules. The creative aspect of language behavior implies that the human mind is involved in deep processing of meaning rather than in memorized responses to environmental stimuli. Chomsky's view of language and cognitive psychology, dubbed generative transformational grammar, regarded language acquisition as an internal thinking-learning process. Chomsky claimed that children are biologically programmed for language and have an innate ability to discover for themselves the underlying rules of a language system. Chomsky's ideas led to the demise of structural linguistics, behaviorist psychology, and the ALM approach to language learning.

An alternative theoretical position emerged centered on the role of the linguistic environment in combination with the child's innate capacities in acquiring language. This position (interactionist) viewed language development as the result of a complex interplay between innate language capacities of the learner and the learner's environment. Unlike the innatist position (e.g., Chomsky 1959), the interactionists claimed that language had to be modified to the ability of the learner. According to Long (1985), language input was made comprehensible by simplifying the input, by using linguistic and extralinguistic cues, and by modifying the interactional structure of the conversation. Long maintained that speakers adjust their language as they interact or negotiate meaning with others. Through negotiation of meaning, interactions are changed and redirected, leading to enhanced comprehensibility. Long proposed that learners, in order to acquire language, cannot simply listen to input, rather they must be active co-constructive participants who interact and negotiate the type of input they receive.

Each of these theories of language acquisition addresses a different aspect of a learner's ability to acquire a language. Behaviorist explanations explain systematic aspects, whereas innatist explanations explain the acquisition of complex grammar. Interactionist explanations assist in understanding how learners relate form and meaning in language, how they interact in conversation, and how they use language appropriately.

3. *Emphases in Second Language Research and Teaching*

Research has revealed that knowledge of language structures demonstrated on discrete-point tests does not ensure communicative ability when the measure of language knowledge is one of more spontaneous language use. Further studies have shown that there is little correlation between the rules learners are taught and their developing knowledge of the second language. Language scholars have demonstrated that certain aspects of second language learning cannot be altered through instruction and that intermediate, non-native-like second language competencies, known as stages of interlanguage, characterize the progression of second language acquisition. Selinker (1974) viewed interlanguage as an intermediate system located on a continuum stretching from the native language to the target language. Corder (1978) stated that, in the interlanguage process, the learner constantly and progressively adjusts the native language system to approximate the target language system more closely (restructuring continuum). Corder noted that not all learners showed evidence of transfer from native language to target language and suggested that there was a uniformity about the way second language learners progress and that they follow approximately the same sequence of development regardless of their native language (developmental continuum). According to this view of second language acquisition, the controlling factor is the innate ability for learning language that all human beings possess. Pica (1983) determined that all language learners progressed through a fixed series of stages, known as developmental sequences, in learning particular linguistic subsystems, such as word order, negation, or relative clauses. In English negation, for example, when communicative samples were examined, it was revealed that both foreign language and second language learners progressed through the same four-stage sequence, defined in terms of placement of negation. Ellis (1986) reviewed several studies that involved Japanese, Spanish, German, and Norwegian children, adolescents, and adult learners. He concluded that all English-as-a-second language learners pass through the following prescribed set of stages: (a) no + phrase, e.g., 'No drink'; (b) negator moves inside the phrase, e.g., 'I no can swim'; (c) negator is attached

to modals, e.g., 'I can't play this one'; and (d) auxiliary system is developed and learner acquires correct use of not and contractions, e.g., 'He doesn't know anything.' This suggests that learners make particular kinds of errors at particular stages in the acquisition of a structure. Each stage marks some kind of restructuring in the mind of the learner regarding that particular structure. Structure evolves over time. These studies had significant implications for how foreign languages should be taught.

4. Learner-centered Instruction

Two communicative approaches, the input model and the input interaction model represent two models of foreign language theory and teaching that investigate the language acquisition process from the perspective of the learner. Krashen (1982) is the principal advocate of the input model of foreign language teaching. His theory is grounded in (a) Chomsky's generative linguistics, (b) research on the effectiveness of different second/foreign teaching methods, and (c) research on affective factors (motivation, anxiety, personality, etc.). Krashen posited that second language acquisition occurs when the learner comprehends the language input in a low-anxiety, high-motivation situation, and proposed that the teacher's role is to create such a learning environment. Krashen further claimed that conscious grammar teaching/learning is effective only in a monitoring capacity to check for grammatical accuracy, not in the acquisition of the second language itself.

Because classrooms remained a major setting for language learning, the pursuit to determine those elements that enhanced classroom language achievement became particularly important. Why do two learners who seemingly have the same instructional opportunity achieve varying levels of language proficiency? Investigations focused on individual skills or abilities and environmental factors that may impact foreign language achievement and proficiency.

Individual cognitive (e.g., intelligence, aptitude, or ability) and affective (e.g., attitude and personality variables) factors were analyzed. Skehan (1986) noted a fairly strong relationship between cognitive variables such as aptitude, intelligence, and language achievement for learners in foreign language classrooms. The predictive power of such traits, however, decreased as the criteria for language proficiency became more communicative and the learning setting became more natural (versus formal and instructional). The most avid pursuit in research occurred in investigations of the role of motivation in learning language and the learner's attitude toward the target language and culture. Using Gardner and Lambert's (1972) differentiation between integrative and instrumental motivation, researchers reported no significant advantage for an integrative (intrinsic) motive and others

reported an advantage when the learner was driven by instrumental (extrinsic) motives. Integrative motivation was defined as one in which the target language was being learned by an individual in order to be accepted by the native speaker community. Instrumental motivation was one in which the language was being learned for external benefits, such as securing a better job.

Results of studies investigating environmental factors reported on the effect on achievement scores. Carroll (1975) conducted a survey of French instruction in eight countries and noted effects on achievement by gender, school type, and teacher gender, and mixed effects according to parental interest. Social factors outside the school were determined to have a significant impact on the development of language proficiency. Both cognitive and affective factors were investigated to explain the variance in foreign language achievement. Motivation, attitudes, anxiety, self-esteem, tolerance of ambiguity, risk-taking, cooperation, and competition proved to be key variables that explained individual differences in foreign language learning (Ellis 1994). Successful language learning was determined to be largely dependent on who was learning the language, under what circumstances, and for what purposes. Foreign language acquisition was revealed to be a complex, multidimensional process influenced by both learner and environment variables. The questions generated by these theories and research studies began to focus on significant new responsibilities on the part of the teacher in the design and support of individual and personalized learning tasks.

5. Learning and Measurement

Language teaching throughout the 1990s has experienced numerous curricular innovations in response to the importance of providing students with opportunities to acquire and practice the foreign language in contextualized and meaningful language communicative tasks at all stages of the second or foreign language acquisition process. Communicative language teaching (CLT), the term most associated with current discussion of method, emerged as a significant approach that found universal resonance and support in theory and application in many contexts and across disciplines (linguists, methodologists, curriculum developers). Central to the rise of CLT was the realization that linguistic competence does not on its own achieve communicative competence (Canale and Swain 1980) and that language used in meaningful, authentic contexts is more readily acquired.

Pair work, group work, cooperative/collaborative learning settings, authentic materials, culturally integrated lesson content and interactive tasks focused on the cognitive and affective domains were integrated into foreign language classrooms. Constructivist

teaching practices, influenced by Vygotsky's emphasis on social interaction in learning and development, helped learners to internalize and reshape new information. The theoretical underpinnings of Vygotsky's (1978) view of language learning that maintained contextualized input in cooperative, meaningful interactions with others enhanced language acquisition took hold in classrooms around the globe. The interaction between an expert (teacher) and novice (learner) in a problem-solving task (scaffolding) in which the expert's role was to provide the novice with scaffolded help became the model for communicative tasks in the foreign language classrooms. The expert's/teacher's role was to gain the learner's interest in the task, simplify the task, keep the learner motivated, point out important features, reduce anxiety and frustration during problem solving, and model appropriate form.

In accordance with the new responsibilities, the role of the classroom teacher shifted to that of an architect, creating meaningful, interactive, cooperative learning tasks designed to engage the learner actively in negotiating language meaning in authentic contexts.

The focus on student language proficiency as measured through performance-based tasks made itself felt both in language learning research and in teaching. Questions emerged regarding how language proficiency could be enhanced and how best to measure the level of language proficiency.

As the proficiency movement gained momentum in the 1980s, most especially in the USA, consensus was sought about describing and measuring language abilities. The development of the Provisional Proficiency Guidelines by the American Council on the Teaching of Foreign Languages (ACTFL) defined what language users are able to do with language in speaking, listening, reading, and writing at various levels of performance. These Guidelines marked a major shift in language pedagogy from methodology to measurement and a focus on learner outcomes. In 1996, content standards were published (National Standards in Foreign Language Education Project 1996) that delineated what learners should know and be able to do. More recently, the ACTFL Performance Guidelines for K-12 Learners (ACTFL 1999) described language performance within three modes of communication (interpersonal, interpretive, presentational) to assist teachers in understanding how well students demonstrate language ability at various points along the language-learning continuum. A similar effort by the International Association for the Evaluation of Educational Achievement's Language Education Study is seeking a comparison and evaluation of the outcomes of different educational systems across Europe. The Standards Movement, seeking to promote the establishment of guidelines for the teaching of foreign languages for all learners, indicates the growing concern with learner outcomes and accountability. In a standards-driven environment the shift to

student performance will require that teachers have a repertoire of approaches that target specific goal areas or standards.

6. Future Directions

Future directions in research should investigate the potential use and effect of computer technology on foreign language learning. As classroom tasks become more focused on real-world issues, texts, or events, and problem-solving-based tasks, computers introduce a new dimension to the teaching and learning process. Technology allows students to manipulate learning materials/language at their own pace and according to individual needs. Students examine reports, authentic documents and web pages to find information that can be synthesized and discussed later and can collaborate electronically with youth from around the world. In such a learning environment the role of the teacher changes from one of authority figure/expert who delivers knowledge to one who facilitates guides and supports student learning. The teacher assumes greater responsibilities in designing and supporting individual and personalized learning tasks. This has tremendous implications for teacher educators and teacher trainers to act as agents of change.

One of the most effective research methodologies that emerged in the last few years has been action research. Inquiring into one's own instructional practices through classroom-based investigations, teachers actively contribute to the research endeavor and change practices based on findings. Such research promises to improve teaching practices that are of interest to both researchers and teachers.

Methodologically classroom-oriented research has been largely conducted within the framework of correlational approaches, case studies, survey research, ethnographic research, experiments, and discourse analysis (Johnson 1992). While the choice of research method is largely determined by the nature of the research question to be investigated, or by the hypothesis to be tested, thoughtful combinations of qualitative and quantitative research on foreign/second language learning conditions will provide valuable insights into language-acquisition processes. Greater use of qualitative methods investigating students in their classrooms with special attention to cultural, situational, and longitudinal contexts is needed and recommended. As foreign language research draws on related disciplines (psychology, psycholinguistics, neurobiology, neurolinguistics, sociology, linguistics) to better explain conditions that lead to greater language proficiency and differential success among foreign language learners, a deeper understanding of how languages are acquired and consequently how they should be taught will be gained. As learning and teaching innovations are

tested and researched, new insights will be gained that will influence teaching practices globally.

See also: Communication and Social Psychology; Language Acquisition; Second Language Acquisition; Verbal Communication, Psychology of

Bibliography

- ACTFL 1999 *ACTFL Performance Guidelines for K-12 Learners*. American Council on the Teaching of Foreign Languages, Yonkers, NY
- Canale M, Swain M 1980 Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1: 1-47
- Carroll J B 1975 *The Teaching of French as a Foreign Language in Eight Countries*. Wiley, New York
- Chomsky N 1959 Review of *Verbal Behavior* by B. F. Skinner. *Language* 35: 26-58
- Corder S P 1978 Language-learning language. In: Richards J C (ed.) *Understanding Second and Foreign Language Learning: Issues and Approaches*. Newbury House, Rowley, MA
- Ellis R 1986 *Understanding Second Language Acquisition*. Oxford University Press, Oxford, UK
- Ellis R 1994 *The Study of Second Language Acquisition*. Oxford University Press, Oxford, UK
- Gardner R C, Lambert W E 1972 *Attitude and Motivation in Second Language Learning*. Newbury House, Rowley, MA
- Jonson D M 1992 *Approaches to Research in Second Language Learning*. Longman, White Plains, NY
- Krashen S 1982 *Principles and Practice in Second Language Acquisition*. Pergamon, Oxford, UK
- Long M H 1985 Input and second language acquisition theory. In: Gass S, Madden C (eds.) *Input in Second Language Acquisition*. Newbury House, Rowley, MA
- National Standards in Foreign Language Education Project 1996 *Standards for Foreign Language Learning: Preparing for the 21st Century*. Allen Press, Lawrence, KS
- Pica T 1983 Adult acquisition of English as a second language under different conditions of exposure. *Language Learning* 33: 465-97
- Selinker L 1974 Interlanguage. In: Schumann J H, Stenson N (eds.) *New Frontiers in Second-language Learning*. Newbury House, Rowley, MA
- Skehan P 1986 The role of foreign language aptitude in a model of school learning. *Language Testing* 3: 188-221
- Skinner B F 1957 *Verbal Behavior*. Methuen, London
- Vygotsky L S 1978 *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA

A. K. Moeller and E. Koubek

Foreign Policy Analysis

What causes states to adopt certain foreign policies? Why do some nations seek to extend their influence abroad while others define their interests in narrower,

more inwardly focused terms? Statesmen, historians, and political scientists have long pondered these questions. Though there is no shortage of answers, most observers have followed one of two distinct paths: the first emphasizes states' external or international situation; the second stresses their internal or domestic makeup. This article examines each of these traditions or approaches in turn.

1. Realpolitik

'Realism,' which emphasizes a state's relative position in the international distribution of power, is usually regarded as the dominant approach to the study of foreign policy. Certainly, it is the oldest. Ever since Thucydides (1972) wrote about the Greek city-states' struggle for power and security, many have embraced realism's basic postulates as the best guide to explaining state behavior: anarchy is the basic, irreducible fact of international life (Waltz 1959). In the absence of a central authority to settle disputes over territory and wealth, states must provide for their own security and welfare. Exactly how they go about this, and the conditions that make success most likely, are the central issues animating realism.

1.1 The Classical School

Although scholars classify realism into different variants ('offensive' and 'defensive' realism and 'classical' and 'structural' realism), in the foreign policy field, the term is most often associated with classical realists, scholars like E. H. Carr (1939), Nicholas Spykman (1942), and Hans J. Morgenthau (1978). Writing during the years between World Wars I and II, these self-proclaimed realists sought to catalogue the 'do's' and 'don'ts' of international power politics. Their work continues to be a touchstone of realist thinking about foreign policy.

Classical realists start from the assumption that states are influence-maximizers: they seek, first and foremost, to control and shape their external environment and, all things being equal, will seek more rather than less influence. How ambitious and single-minded their leaders are in this regard depends on many things, but for the realist none is more essential than a nation's power, or more precisely, its relative power, where power refers to its share of material resources such as military strength, world trade, and population.

For realists, power is indispensable. Knowing how much a nation has, they argue, tells much about the foreign policies it will likely pursue. In general, as states' international power or capabilities increase, they will seek to extend their political interests abroad; as their power ebbs, their actions and aspirations will be scaled back accordingly. In short, power

shapes intentions: the stronger the state's international position, the more expansive its foreign policy interests will be.

1.2 Realism's Strengths and Weaknesses

One of realism's great virtues is that it reminds us of the importance of power in understanding state behavior. Big changes in a nation's international position—such as Britain's rise to great powerdom in the late eighteenth century, Germany's ascent under Otto von Bismark after 1870, and American dominance after World War II—can lead to big changes in its foreign policies. The theory that a country's power determines its intentions is crisp and parsimonious, and much of realism's appeal stems from this fact.

A major weakness of realism, however, is that it is underdetermining. A nation's international position may well create incentives for action. But rarely is a state's 'relative power' vis-à-vis its allies, competitors, and neighbors clear and unambiguous, and rarely is there a consensus among its leaders over the appropriate response to foreign policy opportunity or challenge. For every rising state that has acted as realism predicts, there is one that has not. The same is true about states in decline (Kennedy 1987). Declining powers often continue to expand abroad, failing to make the adjustments realism predicts. Sometimes this results in disaster as realism would predict, but sometimes they get away with it.

Realists try to address this shortcoming in one of two ways. The first involves substituting 'elite perception' about international strength for objective reality. By dispensing with objective measures of power in favor of policy makers' perceptions of power, it is possible, realists contend, to explain gaps or lags in a country's response to increases (or decreases) in its power as assessment failures (Friedberg 1988, Jervis 1976). Strategic miscalculations or misjudgments arise because of difficulties inherent in judging military strength, commercial prowess, and so on accurately.

A second solution to the problem of indeterminacy involves the introduction of domestic decision-making structures into the model. Even if leaders assess their nation's power correctly, they may not be able to respond appropriately because of domestic institutional constraints (Krasner 1978, Zakaria 1998). The weaker (i.e., the more fragmented and decentralized) a state's political institutions, the harder for its leaders to mobilize internal resources and act with dispatch. Here, state strength functions as an 'intervening variable,' mediating the transmission of capabilities (realism's independent variable) into behavior (its dependent variable).

Such efforts to bring theory and reality into line are useful. But introducing variables such as elite perception and state strength also carries a methodological price, most notably a loss of parsimony. One

needs to know a great deal about the country in question, as well as its decision makers, to make the theory work. Even more serious is the inevitable loss of distinctiveness that accompanies such broadening of the theory. The more realists turn to domestic variables, the harder it becomes to distinguish *Realpolitik* from its competitor: *Innenpolitik*.

2. *Innenpolitik*

If relative power in the international arena is realism's core idea, *Innenpolitik*'s is domestic political structure. Like realism, this approach claims that the context in which foreign policy decisions are made shapes policy. Where they differ is how they understand 'context' or, more accurately, in their judgment about which setting, the international or the domestic, should receive pride of place in the analysis. For realists, a state's external situation is determinant; for *Innenpolitikers* a country's internal makeup is decisive.

2.1 Many Theories, One Approach

Like realism, *Innenpolitik*'s roots are European. The approach traces its origins back to early-twentieth-century accounts of British imperialism and German *Weltpolitik*, though at least one variant of the general approach, the so-called 'democratic peace theory,' can be traced to Immanuel Kant's discourses. Vladimir I. Lenin's (1916) theory of colonialism, and economist Joseph A. Schumpeter's (1955) theory of imperialism are classic examples of the *Innenpolitik* genre. In the North American context, the earliest example is historian Charles Beard's (1934) revisionist account of America's entry into World War I.

As these examples suggest, *Innenpolitik* is less a single theory than a family of theories, each stressing a different, specific domestic variable in the explanation of foreign policy outcomes. Some argue for the importance of a nation's culture, others its socioeconomic structure, and still others its political institutions. What ties these theories together is the common assumption that foreign policy is best understood from the 'inside out,' as a product of domestic conditions and circumstances.

In its strongest forms, *Innenpolitik* makes predictions about foreign policy based solely on information about a state's domestic makeup. The notion that democracies behave differently from non-democracies is a popular, if overdrawn, example. A more subtle version of the argument is democratic peace theory (Doyle 1983). Liberal institutions and norms, it is claimed, make democracies less inclined to fight each other, even if they show no such reluctance when dealing with authoritarian regimes.

The key distinction in democratic peace theory is regime type—democratic or authoritarian. How war-

prone states are depends upon their domestic political institutions. The implication is clear: change a state's domestic structure, and its foreign policy behavior will change accordingly. International conditions, while not completely ignored, are secondary. They neither constrain nor dictate foreign policy; they provide opportunities. Whether states seek to exploit international opportunities or not depends on how constrained or motivated their leaders are by 'checks and balances,' interests of domestic constituencies, and pressures to gain office in a competitive political environment.

A focus on states' domestic makeup also informs many theories of imperialism, the most well-known of which, Marxism, claims that capitalist states, due to problems of overproduction and underconsumption, are more prone to expansionism than non-capitalist ones. Whether or not they pursue an aggressive imperialist foreign policy, and how likely they are to succeed if they do, may depend on the international balance of power, but their expansionist impulses do not. They are internally generated. Again, domestic structure is determinative.

2.2 Choice and Decision-making

Theories of democratic peace and imperialism are structural theories, and in this regard are like realism. Strictly speaking, information about the beliefs, perceptions, and prejudices of decision makers is unnecessary. For *Innenpolitik*, the implicit claim is that states with similar domestic structures will pursue similar foreign policies. Conversely, dissimilar states will differ in their foreign policies.

The problem is that there are many examples of states with similar domestic structures that have acted differently in foreign policy. The reverse is true, too: dissimilar states sometimes act similarly. It turns out that just as a state's external position rarely precludes a range of responses, its internal makeup leaves room for meaningful choice. To understand how those choices are made, domestic-level theorists, like realists, have sought to incorporate decision-making into their models.

The decision-making theory that has had the biggest impact on the *Innenpolitik* approach to studying foreign policy is 'bureaucratic politics.' A form of interest-group theory, its central claim is that foreign-policy decisions are the result of bargaining among government officials and agencies (Allison 1971). Foreign-policy choices are thus seen as a product of the distribution of power and initiative among governmental actors, which varies by policy issue.

That government officials' choices matter in the making of foreign policy is clear. Yet how useful bureaucratic politics is in explaining those choices is less obvious. One problem is that the decision-making process inside government is not as autonomous and insulated from larger political processes as bureau-

cratic politics theorists would have it: 'palace politics' often reflects larger social interests and powers. Moreover, the theory seems to work best when it matters least: on issues that are not seen as crucial to the state's security, and that are therefore handled mostly inside the foreign-policy-making bureaucracy.

3. The State of the Art

As structuralist theories, realism and *Innenpolitik* both seem to flatten the sense of choice and politics that characterizes statecraft. The future of foreign policy studies lies in finding systematic ways to transcend this problem: that is, to introduce choice and politics into models that recognize that foreign-policy making is constrained by domestic institutions and international balances of power.

Some possible strategies have been proposed. One is to introduce politics by analyzing the uneven impact of international forces on the different constituencies, sectors, or interest groups within one country (Gourevitch 1986, Trubowitz 1998). Groups compete within the framework of domestic political institutions to promote what is, in the final analysis, an inherently partisan vision of 'the national interest.'

Another is Robert Putnam's (1988) solution, which is to envision statesmen as rational maximizers who are engaged in 'two-level games.' Foreign-policy makers must strategize in domestic political arenas, as 'democratic peace' theorists suggest, and at the same time seek stability or advantage in the competitive international settings that are the centerpiece of realist analysis.

Finally, there are constructive theories that dissolve the distinction between 'international' and 'domestic' (Katzenstein 1996). Constructivist writers see 'structural constraints' themselves—understandings of international power balances, state interests, and domestic priorities—as the product of ideologies, cultural bias, and discourses about state and nation.

The main empirical challenge for foreign policy analysts is to figure out ways to conceptualize 'international context' and 'national interest' at a time when international relations is fluid and dominated increasingly by forces not controlled directly by governments (such as international capital flows), and when traditional definitions of security must be broadened to include environmental, economic, and other non-state sources of challenge.

See also: Balance of Power: Political; Conflict/Consensus; Decision-making Systems: Personal and Collective; Diplomacy; Dispute Resolution in Economics; Imperialism: Political Aspects; International Law and Treaties; Military and Politics; National Security Studies and War Potential of Nations; Peace; Power: Political; Risk: Theories of Decision and Choice; War, Sociology of

Bibliography

- Allison G T 1971 *Essence of Decision: Explaining the Cuban Missile Crisis*. Little Brown, Boston
- Beard C A 1934 *The Idea of National Interest: An Analytical Study in American Foreign Policy*. Macmillan, New York
- Carr E H 1939 *The Twenty Years' Crisis, 1919–1939: An Introduction to the Study of International Relations*. Macmillan, London
- Doyle M W 1983 Kant, liberal legacies, and foreign affairs. *Philosophy and Public Affairs* 12: 205–35
- Friedberg A L 1988 *The Weary Titan: Britain and the Experience of Relative Decline, 1895–1905*. Princeton University Press, Princeton, NJ
- Gourevitch P 1986 *Politics in Hard Time: Comparative Responses to International Economic Crisis*. Cornell University Press, Ithaca, NY
- Jervis R 1976 *Perception and Misperception in International Politics*. Princeton University Press, Princeton, NJ
- Katzenstein P (ed.) 1996 *The Culture of National Security: Norms and Identity in World Politics*. Columbia University Press, New York
- Kennedy P 1987 *The Rise and Fall of the Great Powers: Economic Change and Military Conflict from 1500 to 2000*. Random House, New York
- Krasner S 1978 *Defending the National Interest: Raw Materials Investments and US Foreign Policy*. Princeton University Press, Princeton, NJ
- Lenin V I 1916 (1939) *Imperialism, the Highest Stage of Capitalism*. International Publishers, New York
- Morgenthau H J 1978 *Politics Among Nations: The Struggle for Power and Peace*, 5th edn. Knopf, New York
- Putnam R 1988 Diplomacy and domestic politics: The logic of two-level games. *International Organization* 42(3): 427–60
- Schumpeter J A 1955 *Imperialism and Social Class*. Trans. H Norden. Meridian, New York
- Spykman N J 1942 *America's Strategy in World Politics: The United States and Balance of Power*. Harcourt, Brace, New York
- Thucydides 1972 *History of the Peloponnesian War*. Trans. R. Warner. Penguin, New York
- Trubowitz P 1998 *Defining the National Interest: Conflict and Change in American Foreign Policy*. University of Chicago Press, Chicago
- Waltz k 1959 *Man, the State, and War*. Columbia University Press, New York
- Zakaria F 1998 *From Wealth to Power: The Unusual Origins of America's World Role*. Princeton University Press, Princeton, NJ

P. Trubowitz

Forensic Psychiatry

A number of definitions of forensic psychiatry must be considered. The American Academy of Psychiatry and the Law (AAPL) (1985) endorses the definition of forensic psychiatry adopted by the American Board of Forensic Psychiatry Incorporated which is:

Forensic Psychiatry is a subspecialty of psychiatry in which scientific and clinical expertise is applied to legal issues in legal contexts embracing civil, criminal, correctional or legislative matters: forensic psychiatry should be practiced in accordance with the guidelines and ethical principles enunciated by the profession of psychiatry.

In the preface to the British textbook of forensic psychiatry, its field of operation is defined as

the overlap, interface, and interaction of psychiatry and the law in all its aspects; criminal behaviour, civil litigation, family law, the diagnosis, care and treatment of psychiatric patients where the disorder is associated with abnormalities of behaviour, legislation and numerous other problems such as the management of violence in the study of sexual deviation.

It adds that

forensic psychiatry continues to be concerned primarily with mentally disordered offenders, but now encompasses a wide range of offences and gives much more prominence to diagnosis, management and treatment in prison, hospital or the community using the skills of psychologists, nurses, social workers, and probation officers in addition to psychiatrists.

The Canadian approach to forensic psychiatry mirrors the British definition as is seen in the recent application for subspecialty status by forensic psychiatry under the auspices of the Canadian Academy of Psychiatry and the Law. The goal of forensic psychiatry is defined as

promoting the mental health of Canadians involved in legal matters. We aim to provide a functional interface between psychiatry and the law and to improve Canadian society as a whole through early intervention programs, better diagnosis and treatment and ultimately a reduction in offending.

In twentieth-century Europe, forensic psychiatry has generally centered on advances in the fields of neurology, psychiatry, and sciences related to these disciplines. In general terms, although there is no single definition of forensic psychiatry in Europe, it would tend to follow the British and Canadian approaches. In summary therefore, as we move into the twenty-first century there appears to be a convergence of the American definition of forensic psychiatry with the British, Canadian, and European definitions which relate to the evaluation and treatment of the mentally abnormal offender while at the same time being involved in the psychiatric assessment (for legal purposes) of fitness to stand trial, criminal responsibility in forensic psychiatry dealing with criminal matters, and similar assessments involving the civil aspects of the law. As we approach the twenty-first century, forensic psychiatry in all parts of the world appears to be moving and establishing its roots as a medical specialty and a subspecialty of psychiatry.

Even though the medical roots are emphasized, it is quite clear that the forensic psychiatrists must be familiar with basic issues in law.

1. The Concept

The forensic psychiatrist needs to be an expert clinician as well as being well versed in law and psychiatry as it applies in the jurisdiction where he or she lives. Law is understood to encompass criminal, civil, and mental health law. In medicine as well as the social sciences there is an increasing number of attempts to complete research studies involving large samples in order better to define the characteristics of disease and aberrant as well as normal behavior. This has led to progress in many areas including forensic psychiatry. This allows the individual to be classified according to whether they have certain clinical and other characteristics in common with a certain reference group. An estimate of the individual's personality characteristics can be measured against a reference sample. The difficulty in forensic psychiatry is that the legal system looks at unique characteristics of a given individual before the courts, rather than any theoretical concept of how he or she may fit a given reference group. This can potentially create a conflict where the scientific aspect of forensic psychiatry could be seen as more prejudicial than probative in the context of the legal system.

The development of forensic psychiatry in Britain, Europe, and the United States was based on historical events in the eighteenth and nineteenth centuries. In Britain the writings of Sir Mathew Hale and the trials of Arnold, Ferris, Hadfield, Oxford, and McNaughton were major landmarks in the recognition of forensic psychiatry. The development of the Bethlem Hospital and later Broadmoor as special facilities for the criminally insane were also fundamental to the growth and development of forensic psychiatry. This was extended to other facilities such as the special psychiatric facilities at Rampton and Parkhurst in the early part of the twentieth century.

The development of forensic psychiatry in Europe followed a similar path. Forensic psychiatry emerged from events in the eighteenth and nineteenth centuries. The French Revolution brought to light documents about prisons, workhouses, and civil hospitals being used to detain members of the public. This was directed specifically to the way and rationale for the detention of persons who had been found insane. Concerns over the method of detention led to pressure for scientists of the day to define insanity more precisely. The influence of religion declined and this allowed the scientific roots of modern clinical psychiatry to develop. Philippe Pinel had an enormous influence in his *Medico-Philosophical Treatise on Mental Alienation or Mania*, which was rapidly translated into English and German. This led to a system of classifying mental disorders that was soon adopted in Europe and

England. This helped in the development of forensic psychiatry as it brought more consistency to the testimony in insanity hearings before the courts. Esquirol, Morel, Lombroso, Krafft-Ebing, and others contributed to these important advances. As in Britain, special institutions for the criminally insane were developed and further research occurred in these institutions.

The United States saw a similar development with Benjamin Rush delivering the first lecture on forensic psychiatry under the guise of medical jurisprudence. The lecture covered topics ranging from civil incompetence caused by psychiatric disorder, and financial incompetence, to mental disorder and criminal responsibility. Rush was followed by Beck who in 1823 published *Elements of Medical Jurisprudence*. He referred to British and French psychiatry as well as Benjamin Rush in his book. Isaac Ray followed Rush and Beck and published the *A Treatise on the Medical Jurisprudence of Insanity* in 1838. This was used to defend Daniel McNaughton. Ray's approach in which he carefully studied the law and the judicial opinions of the day—in particular, anything in the law that obstructed the psychiatrically ill (the insane) being able to receive treatment. He became the superintendent of the Maine Insane Asylum and later the Butler Hospital. Perhaps most importantly he was one of the original founders of the American Psychiatric Association, and was its President from 1855 to 1859. In 1869 Ray went on to write a paper *Confinement of the Insane* that has remained fundamental to forensic psychiatry throughout the twentieth century.

In the first place, the law should put no hindrance in the way of the prompt use of those instrumentalities which are regarded as the most effectual in promoting the comfort and restoration of the patient. Secondly, it should spare all unnecessary exposure of private troubles and all unnecessary conflict with popular prejudices. Thirdly, it should protect individuals from wrongful imprisonment. It would be objection enough to any legal provision, that it failed to secure these objects, in the completest manner. (Ray 1869)

Highly significantly, Quen writing in *The Principles and Practices of Forensic Psychiatry* (1990) reports 'These elements remain necessary, valid, and unachieved in America today.'

As outlined above forensic psychiatry was fundamental to the development of psychiatry in the eighteenth and nineteenth centuries and was adopted by the founders of British, European, and American psychiatry as part of their role to promote psychiatry as a medical specialty.

2. Criminal Law and Forensic Psychiatry

This involves fitness-to-stand-trial evaluations and assessments for criminal responsibility. It involves the forensic psychiatric assessment of mentally abnormal

offenders. The forensic psychiatrist is, by virtue of his or her professional training, an expert in psychiatry and specifically in mentally abnormal offenders. This expertise is applied to legal matters such as fitness to stand trial and criminal responsibility. The challenge to the forensic psychiatrist is that the traditional doctor-patient relationship does not exist under these circumstances. The traditional aims of psychiatry of treating or helping the patient are also not obvious and may even be contradicted. The definition of forensic psychiatry by the well-known American forensic psychiatrist Dr Seymour Pollack (1974) addresses these contradictions. Pollack defined forensic psychiatry thus:

forensic psychiatry is limited to the application of psychiatry to evaluations for legal purposes. Psychiatric evaluation of the patient is directed primarily to the legal issues in which he is involved, and consultation is concerned primarily with the ends of the legal system, rather than the therapeutic objectives of the medical system.

Pollack is correct in terms of evaluations for fitness to stand trial and criminal responsibility. In the former, the initial role of psychiatric evaluation is to establish the diagnosis, from the standpoint of either a medical illness or a psychiatric disorder affecting the person's mental functioning. Once this has been established the next task for the forensic psychiatrist is to assess whether the degree of disorder present affects the person's capacity to proceed to trial. The legal basis may vary slightly between jurisdictions but all are similar and would be analogous to Section 2 of the Criminal Code of Canada, which defines 'unfit to stand trial' as follows:

unfit to stand trial means unable on account of mental disorder to conduct a defence at any stage of the proceedings before a verdict is rendered or to instruct counsel to do so, and, in particular, unable on account of mental disorder to
(a) understand the nature and object of the proceedings
(b) understand the possible consequences of the proceedings, or
(c) communicate with counsel (Section 2 Criminal Code of Canada 2000).

The process is clearly conducted for a legal purpose and, arguably, not necessarily in the person's best interest, as it may mean finding them fit for trial which may in turn result in a guilty verdict and possibly lengthy incarceration in a correctional setting. Alternatively, if they were found unfit and remained unfit, while they would still be incarcerated, they would probably be detained in a hospital setting. In Canada and most other jurisdictions, a finding of unfitness means that a judicial hearing would follow as to whether psychiatric treatment would render the individual fit to stand trial. In Canada this could result in a treatment order issued by a judge, ordering psychiatric treatment (most commonly antipsychotic medication) to render the person fit to stand trial. This

can be done independently of the person's capacity to consent or refuse psychiatric treatment.

An assessment of criminal responsibility involves judging whether the person's mental status at the time the crime took place rendered him or her incapable of appreciating the nature of their actions as well as knowing whether their actions were wrong. It is beyond the scope of this article to go into criminal responsibility in detail as the legal basis varies considerably in different jurisdictions. This particularly true in the United States where the criminal law is under state jurisdiction and therefore each state varies in terms of what it uses as the legal basis for a finding of not criminally responsible on the basis of mental disorder. The most commonly used standard in Western countries is the M'Naughton Test or some modification of the test. In recent years a number of American states have either abolished or severely limited the application of 'insanity' (mental disorder) to reduce the *mens rea* component of criminal responsibility. In order to be criminally responsible for a crime the state must prove the accused committed the unlawful act (*actus reus*) and that they also had the intent to commit the crime (*mens rea*). Mental disorder if present at the time the crime was committed can negate *mens rea*, leading to a finding of 'not guilty by reason of insanity,' not criminally responsible on the basis of mental disorder, or guilty but mentally ill. Such a verdict usually results in confinement in a forensic psychiatric facility until the acqutee is regarded as having recovered and no longer being dangerous to the community at large. In other cases, a mental disorder may result in a diminishing of criminal responsibility, allowing the accused to be found guilty of a lesser offence carrying a less severe penalty. This is seen in situations where a homicide has been committed and the individual suffered from a mental disorder at the time of the homicide. The degree of mental disorder is regarded as not sufficient to negate *mens rea* but to reduce it. This could result in a finding of manslaughter as opposed to murder where the intent to kill was reduced although the accused was still criminally responsible for the death of the victim. The original M'Naughton Test was strictly a cognitive test. In its original form it was:

To establish a defence on the grounds of insanity, it must be clearly proved that at the time of committing the act, the party accused was labouring under such a defect of reason, from disease of the mind, as not to know the nature and quality of the act he was doing, or if he did know it, that he did not know he was doing what was wrong. (Quoted in Hunter and MacAlpine 1963.)

Various modifications of the original M'Naughton Test have occurred in the United States and elsewhere. The most important of these has been the adoption, in the United States, of the American Law Institute Test (ALI Test), which includes both cognitive and volitional aspects. To complicate things even further the

actual legal interpretation of the meanings of the various tests of criminal responsibility is subject to case law. This means the interpretation may change, depending on a specific interpretation by a particular judge in a specific court case that is upheld by various appeal courts. Forensic psychiatrists who act as expert witnesses in cases of criminal responsibility based on mental disorder are therefore becoming involved in a very complicated process where they have to use their psychiatric skills initially to establish a psychiatric or medical diagnosis. Drawing on the observations of others as well as the medical history of the accused, they must establish, by a process of reconstruction, whether the disorder was present at the time of the offence. If it was present, they must assess its degree or severity and how that would affect the operating mind of the accused both cognitively and possibly volitionally. Then they must form an opinion with reference to the legal test for criminal responsibility in the particular jurisdiction where the trial is being held. He or she will need a working knowledge of the case law interpretations as they apply in the specific jurisdiction and in the specific case. Finally, as expert witness he or she must understand the facts of the case as they emerge at the time of the trial and be able to incorporate them into the opinion being rendered.

In addition to criminal responsibility and fitness to proceed with trial, forensic psychiatrists are also involved in the sentencing of mentally abnormal offenders. Here the role of an expert is to address the third principle of sentencing, which is rehabilitation (the first two being retribution and deterrence). In certain jurisdictions this may be the most prominent role for the forensic psychiatrist as it is used to mitigate the sentence while at the same time allowing the judge to direct treatment in a correctional facility for a psychiatric disorder. It therefore provides a link between forensic psychiatry and correctional psychiatry. This linkage varies in different jurisdictions and countries according to the degree of development of forensic psychiatry and the availability of treatment resources in the correctional system. A highly contentious issue related to sentencing is the participation of forensic psychiatrists in the death penalty process in the United States, including pretrial assessment, the sentencing phase, the determination of competence to be executed, and most controversially the assessment of restoration of competence to be executed as a result of psychiatric treatment.

3. *Civil Law and Forensic Psychiatry*

This branch of forensic psychiatry involves psychiatric malpractice; psychiatric disability determinations; personal injury evaluations; testamentary capacity and guardianship; and family law matters, including parental competence and child custody evaluations. Included here is the specialty of the child and ado-

lescent forensic psychiatrist. These individuals must combine the skills of two psychiatric subspecialties, namely, child and adolescent psychiatry, and forensic psychiatry. In some university divisions of forensic psychiatry these skills are available through a family court clinic. This clinic would receive consultations from the family court system to evaluate parental competence in child abuse and neglect cases and also assist in determining the best interests of the child in divorce custody cases. Family court clinics have also traditionally provided evaluation of young offenders. The evaluation of young offenders, principally adolescents, requires the clinical skills and experience necessary for evaluating children and adolescents. The importance of these dual skills cannot be underestimated and at times has called for the recognition of a sub-subspecialty of child and adolescent forensic psychiatry in some countries.

Testamentary capacity may be a relatively small component of forensic psychiatry, but there is no doubt as to its complexity and the special skills necessary for these retrospective evaluations and being able to communicate the evaluation to the legal authorities.

4. *The Legal Regulation of Psychiatry and the Forensic Psychiatrist*

The legal regulation of psychiatry involves informed consent and competence; the right to refuse psychiatric treatment and the right to psychiatric treatment; psychiatric hospitalization and involuntary civil commitment; outpatient civil commitment; confidentiality; the duty to warn or protect against professional misconduct, including the impaired physician, competence to practice, and professional sexual misconduct. These are areas of psychiatry and the law where the average psychiatrist also needs to develop skills and competence and are, therefore, not an exclusive area of practice for the forensic psychiatrist. The forensic psychiatrist, however, contributes special skills and training in the legal system, and the skill to be a competent expert witness. This means their role is to assist the non-forensic psychiatrist by consultation and advice and, in the more complex cases, forensic psychiatric expertise may be desirable.

5. *Special Clinical Skills of the Forensic Psychiatrist*

A wide range of special clinical skills is unique to the forensic psychiatrist. Special skills in evaluating and treating violent psychiatric patients are a fundamental requirement for all forensic psychiatrists. This is also expanded to include the assessment of the danger posed by an offender and risk assessment, as well as the risk management of violent mentally abnormal

offenders. Associated with these clinical skills are skills in assessing antisocial personality disorder as well as impulse control disorders. Because of the association between organic brain dysfunction and impulsivity and violence, forensic psychiatrists need to develop skills in assessing the neuropsychiatric aspects of psychiatric conditions that can result in violence or can affect competence. They also need to be able to evaluate and treat the most severe and violent major mental disorders. Forensic psychiatrists must develop special skills in evaluating malingering and the use of hypnosis. The forensic psychiatric aspects of certain psychiatric disorders (e.g., multiple personality disorder) are also within the skill set of the average forensic psychiatrist. The assessment and treatment of sexual deviation is also a required skill.

6. Correctional Psychiatry

The strict definition of correctional psychiatry is 'psychiatry practiced within a correctional system.' Broader definitions would include any agency of social control that is involved in the treatment and rehabilitation of adult or juvenile offenders. These populations have significant psychiatric morbidity and clearly would benefit from psychiatric intervention and this had led to a growth of correctional psychiatry in the last few decades. There is a close relationship between correctional psychiatrists and forensic psychiatrists with correctional psychiatric programs often being outreach programs of university departments of forensic psychiatry (as, for example, in Canada).

7. Organization and Future of Forensic Psychiatry

This is a subject of debate in North America at the beginning of the twenty-first century. There is a consensus that the future growth and development of forensic psychiatry must lie within the field of medicine as a subspecialty of psychiatry. This is already well established in general terms in Canada and Europe. If we accept this consensus, some comment as to the ideal forensic program is warranted. This author strongly endorses the subspecialty status of forensic psychiatry as a critical component of its future development. This implies that the ideal forensic program would be an academic division within a university department of psychiatry. In turn, this would mean that education, clinical training, and research would be fundamental components of the program. General psychiatric residents should be required to do a rotation as part of their residency requirement. As the largest interface between medicine and the law is between psychiatry and the law, it should be part of the basic training of all psychiatric residents. It is almost impossible for the average psychiatrist to go through his career without some

interaction with law and psychiatry tribunals, the criminal justice system, or the civil legal system.

An ideal program would provide fellowship training in forensic psychiatry in an accredited program for a minimum of one year. The forensic clinical program should have a medium or maximum inpatient assessment unit to provide pretrial evaluations of fitness and criminal responsibility. In addition, inpatient and outpatient facilities are required for the treatment and rehabilitation of individuals who have been found unfit to stand trial or not criminally responsible on the basis of mental disorder. The length of stay should allow for both acute and chronic care so a cautious approach to treatment and rehabilitation would prevail. There also need to be outreach programs to correctional facilities and an active participation in pretrial diversion programs. In an ideal situation, specialized clinics for the assessment and treatment of sexual deviation, and aggression and impulse control disorders should also be present. A family court clinic with child and adolescent forensic psychiatrists on the staff should also be included. A multidisciplinary team approach to assessment and treatment is essential. Research into all aspects of forensic psychiatry is critical for its future growth and the development of its profile as a medical subspecialty.

See also: Child and Adolescent Psychiatry, Principles of; Civil Law; Criminal Defenses; Criminal Justice, Ethics of; Aspects of Human Dignity; Expert Witness and the Legal System: Psychological Aspects; Law: Defense of Insanity

Bibliography

- American Academy of Psychiatry and the Law (AAPL) 1985 Definition of Forensic Psychiatry.
- Barras V, Bernheim J 1990 The history of law and psychiatry in Europe. In: Blueglass R, Bowden P (eds.) *Principles and Practice of Forensic Psychiatry*. Churchill Livingstone, London, pp. 103–9
- Blueglass R, Bowden P (eds.) 1990 *Principles and Practice of Forensic Psychiatry*. Churchill Livingstone, London
- Bradford J M W, Glancy G 2000 Commentary on 'Ensuring that forensic psychiatry thrives as a medical specialty in the 21st century.' *The Journal of the American Academy of Psychiatry and the Law* 28: 20–2
- Forshaw D, Rollin H 1990 The history of forensic psychiatry in England. In: Blueglas R, Bowden P (eds.) *Principles and Practice of Forensic Psychiatry*. Churchill Livingstone, London, pp. 61–101
- Glancy G, Bradford J, Gagne P 2000 Psychiatry and the law in Canada: A proposal for subspecialization: Part I. *Bulletin of the Canadian Psychiatric Association*, pp. 20–6
- Hunter R, MacAlpine I 1963 *Three Hundred Years of Psychiatry 1535–1860. A History Presented in Selected English Texts*. Oxford University Press, London
- Quen J M 1990 The history of law and psychiatry in America. In: Blueglass R, Bowden P (eds.) *Principles and Practice of Forensic Psychiatry*. Churchill Livingstone, London, pp. 111–16

- Pollack S 1974 *Psychiatry in Criminal Law*. University of Southern California, Los Angeles
- Ray I 1869 Confinement of the insane. *American Law Review* 3: 193–217
- Rosner R (ed.) 1994 *Principles and Practice of Forensic Psychiatry*. Chapman and Hall, New York
- Weinstock R, Leong D B, Silva A J 1994 Defining forensic psychiatry: Roles and responsibilities. In: Rosner R (ed.) *Principles and Practice of Forensic Psychiatry*. Chapman and Hall, New York
- Watt D, Fuerst M 1998 *The 1998 Annotated Tremear's Criminal Code*. Carswell, Toronto, ON

J. Bradford and G. Glancy

Forensic Psychology

Forensic psychology is the branch of psychology concerned with the production and application of psychological knowledge and principles within the legal process. (The term 'forensic,' meaning 'legal' or 'judicial,' is derived from the Latin word *forensis*, which literally means 'relating to the forum or market.' In ancient Rome, a forum was an open square, which served as a market, as well as a place where legal matters were disposed of.)

1. Historical Background

Since both psychology and law are concerned with human behavior, they have been intertwined since the dawn of recorded history (Gudjonsson and Haward 1998). Indications of interest shown in the law by philosophers, from whose ranks psychology would later develop, are found in the earliest Asian, Egyptian, Greek, and Latin writings. Later Greek philosophers continued to express themselves on practical questions of law and topics that today constitute the subject matter of psychology. For example, Plato (ca. 427–347 BC) in his *Republic* considered the question whether mentally disordered people were responsible for their behavior (Allan 1995). However, it took more than 2000 years for psychological knowledge to mature to a level where it was officially allowed in the courtroom for the first time. This happened in 1896 when Albert Von Schrenk-Notzing, student of Wilhelm Wundt who founded the first psychology laboratory in Germany in 1879, testified in a murder trial concerning the effect of pretrial publicity on memory and suggestibility (Blau 1998). Since then, forensic psychology has made great strides. As we enter the twenty-first century, there is hardly an area of the law where psychological expert testimony is not implemented. Ewing (1985) believes that the modern justice system could not function without significant participation by behavioral scientists such as psychologists.

2. Roles of the Forensic Psychologist

As both psychology and law essentially cover all aspects of human behavior, it is understandable that the forensic psychologist serves as an expert on a variety of matters including those considered in the following sections.

2.1 Competency to Stand Trial

Also known as fitness to plead, competency to stand trial refers to the defendant's mental capacity to make a proper defense, i.e., to understand the legal proceedings and the charges against him or her.

Contrary to popular opinion, the mere existence of a mental disorder *per se* does not render an individual incompetent to stand trial. Many mental disorders are of such a nature and intensity that the defendant's competency is not significantly influenced. The crucial question is whether the mental disorder affects the defendant's current functional abilities within the context of relevant legal criteria for determining competency (Gudjonsson and Haward 1998).

As laws and 'legal cultures' differ from country to country, it is understandable that statistics concerning competence to stand trial will vary considerably. For example, proportionally speaking, about 100 times more defendants in the USA are found incompetent than in England and Wales. Even within one country, the percentage of defendants found to be incompetent often varies between 1 percent and 80 percent across jurisdictions.

It does seem that individuals adjudicated incompetent to stand trial often have certain characteristics in common, such as a history of mental health treatment (especially for schizophrenia and depression), lower intellectual functioning, and a present charge of violent crime.

The evaluation of competency is usually very reliable and the agreement rate between psychologists often exceeds 80 percent. Next to the traditional assessment techniques, particularly the clinical interview, several competency assessment instruments are also used.

2.2 Criminal Responsibility

Criminal responsibility (liability/accountability) should not be confused with competency to stand trial. Competency to stand trial is specifically concerned with the defendant's mental capacity during the legal proceedings. Criminal responsibility, also known as the insanity defense, refers to the time of the alleged offense and more specifically asks the question whether the defendant had the mental capacity to distinguish between right and wrong at the time of the alleged offense and could control his or her conduct accordingly. Should the defendant be found not criminally

responsible, he or she rarely goes free. The person is usually referred to a mental health facility where he or she will remain as long as the commitment criteria are met.

The insanity defense is one of the most controversial issues in criminal law. One reason for this are the many misconceptions surrounding this plea. For example, although the public believes (probably as a result of the publicity given to such cases) that the insanity plea is very often entered, such cases are rare: often fewer than 1 in every 1,000 criminal cases. The success rate of the defense varies from country to country and from jurisdiction to jurisdiction, with 25–50 percent as a broad average.

Most defendants whose insanity pleas are successful are diagnosed with major psychotic disorders and have extensive mental health histories, often accompanied by prior civil commitments or findings of incompetence (Golding et al. 1999).

Insanity evaluations are usually regarded as more complex than competency evaluations. It is, after all, exceedingly difficult to determine what another person was thinking, feeling, and doing at the time of an offense that usually happened weeks, months, or even years earlier (Bartol and Bartol 1994). However, it seems that specialized training and experience can lead to an agreement rate between mental health professionals of more than 80 percent. Several measuring instruments have been developed to assist mental health professionals in their assessment of the insanity plea.

2.3 Other Defenses

Next to competency to stand trial and criminal responsibility, the expertise of psychologists is employed in several other defenses in both criminal and civil cases.

(a) *Intoxication* by alcohol or other drugs, because it is self-induced and temporary, is seldom given complete exculpatory effect (Melton et al. 1997). The major exception is when the substance abuse has been prolonged to a point where it has produced a serious mental disorder. However, in most countries voluntary intoxication can be used as a diminished capacity defense claim. In such cases mental health professionals have to produce evidence that the intoxication did indeed affect the defendant's mental capacity significantly.

(b) *Amnesia* (the total or partial loss of memory) at the time of the offense is one of the most common defenses heard in the criminal court, especially concerning violent crimes (see *Amnesia*). It is also one of the most difficult defenses to assess. It is not only difficult to determine the weight that should be given to the claimed amnesia, but also to distinguish between genuine and simulated amnesia. An even more complex task is to determine whether the defendant was

really amnesic when committing the crime, or whether he or she only developed the amnesia afterwards. Although psychologists frequently give expert testimony on amnesia, their ability to assist the court with scientific evidence is questioned. Courts are therefore often skeptical about accepting their testimony as a viable defense element.

(c) *Dissociative identity disorder*, formerly called *multiple personality disorder*, is primarily characterized by the presence of two or more distinct personalities that in turn take control of the person's behavior (American Psychiatric Association 1994). However, the disorder is highly controversial and sometimes called the 'UFO of mental health.' Many mental health professionals believe that the disorder does not exist and amounts to no more than role playing. Although screening instruments and structural interviews have been developed and hypnosis is used to diagnose dissociative identity disorder, experts admit that it is very difficult to distinguish between real cases and malingering (Alloy et al. 1999). It is therefore understandable that legal professionals tend to be very cautious about making use of this defense. However, in rare cases in several countries, dissociative identity disorder has been accepted as a defense, ranging from an insanity plea to extenuating circumstances.

(d) *Post-traumatic stress disorder* (PTSD) is characterized by intensely distressing symptoms (e.g., irritability and sleeplessness) caused by exposure to a traumatic event such as rape, death, combat, accident, and even harassment and difficult work conditions (see *Post-traumatic Stress Disorder*). PTSD can therefore be a factor in both criminal and civil cases. However, establishing the existence of PTSD poses several difficulties for the forensic psychologist. For example, the validity of the diagnoses is often questioned. Although psychological measuring instruments and techniques to assess PTSD are available, they are usually based on self-report. The result is that malingering cannot be ruled out easily.

2.4 Prediction of Violent Behavior

As one of the primary duties of the court is to protect society, assessing the offender's potential for future violence often forms an integral part of the judicial process. This is especially true if the offender is charged with a violent crime. Determining the propensity for future violence is therefore an important part of the forensic psychologist's functions. However, the role of the mental health professional in violence prediction is controversial (see *Clinical Psychology: Validity of Judgment*). The main reason for this is that past research showed that mental health professionals have a bad track record when it comes to predicting the probability of a person committing future violence, and even more so when it comes to predicting the type and severity of such potential violence. As such,

professionals can pay a heavy social, professional, and legal price for failing to predict violent behavior, they tend to 'play it safe' and therefore often overpredict. The result is that research has shown that more than 50 percent of persons predicted likely to commit future violence, do not do so.

Ongoing research in this field is fortunately beginning to bear fruit. For example, several variables have been identified which can be regarded as risk factors for violence, e.g., a history of violence, substance abuse, symptoms of a psychosis or antisocial personality disorder (psychopathy), weapon availability, early onset of violent behavior, and absence of social support. Measuring instruments to assist psychologists have also been developed.

2.5 Eyewitness Testimony

'It is...inevitable that what a witness says in court will be a subjective and incomplete version of 'the truth' interspersed with small portions of fiction.' Although most psychologists will agree with this statement by Ainsworth (1998, p. 3), the testimony of an eyewitness still remains a very important part of evidence presented in both criminal and civil cases. However, as a result of a rapidly growing data bank in this widely researched area, especially in the 1980s and 1990s, the courts are currently examining some of their assumptions about the reliability and thus the value of eyewitness testimony. It is therefore understandable that courts increasingly rely on psychologists to provide expert testimony in this regard (see *Expert Witness and the Legal System: Psychological Aspects*).

The following are some of the more important research findings that have been used in court by psychologists: (a) eyewitnesses tend to be less accurate in stressful situations; (b) people experience difficulty making cross-racial identifications; (c) eyewitnesses focus on weapons rather than faces; (d) memory decay starts immediately after a perception; (e) memory gaps are often and easily replaced by unconscious fabrication or by external suggestions; (f) there is not necessarily a correlation between the degree of certainty evinced by the eyewitness and accuracy (Melton et al. 1997); and (g) although younger children are often more suggestible and remember less detail than older children and adults, they do not necessarily have poorer memories (Saywitz and Goodman 1996).

2.6 Personal Injury

The world is progressively becoming a litigious society in which monetary compensation is sought for almost any maloccurrence (Tsushima and Anderson 1996). As many such claims involve 'psychological pain and suffering,' psychologists are increasingly becoming

involved in so-called mental injury cases. Although determining the presence and cause of mental injury constitutes the bulk of psychologists' work in this field, they are also involved in injury claims of a more physical nature, e.g., to assess brain damage or determine the psychological effect of losing a limb.

To assess whether the plaintiff does indeed suffer from a mental injury is not always easy. First, it has to be proved that there is a causal relationship between the specific incident and the plaintiffs symptoms, i.e., that the symptoms did not already exist before the incident or were not caused by other factors after the accident. Second, as possible financial remuneration is involved, plaintiffs often tend to exaggerate or even fake their symptoms. No foolproof method to determine such relationships and distinctions is available. However, by using the most objective, valid, and reliable tests appropriate to the complaints, conducting thorough interviews, and carefully recording case notes and test reports, psychologists can make a valuable contribution towards clarifying the extent to which psychological factors may be involved in personal injury cases (Butcher and Miller 1999).

2.7 Sentencing

If a defendant is found guilty of the charge against him or her, the court has to impose an appropriate punishment. In most countries the aim of such punishment consists of a combination or all of the following: deterrence, retribution, prevention, and rehabilitation. Since these goals have a strong psychological element and psychologists are supposed to be well equipped to provide the court with a psychological profile of the defendant, they often testify in this regard. For example, the psychologist can provide the court with information concerning the defendant's potential for future violence, the sincerity of the repentance shown, the possibility of rehabilitation, a variety of mitigating factors, and the effect of different sentences on the defendant as a unique individual.

Although there seems to be a high correlation between psychologists' recommendations and the final sentencing decision of the courts, it remains an open question as to what extent the courts do indeed take the testimony of the psychologist into account or merely use it to 'scientifically' justify decisions they have already made (Melton et al 1997).

2.8 Profiling

Criminal profiling refers to the process in which the nature of a crime is used to make inferences about the personality and other characteristics of the likely offender.

The sensational and dramatic elements of profiling portrayed in various well-known movies, television

series, and books (usually those depicting a serial killer) seem to be in sharp contrast with reality. Profiling is not widely accepted in the psychological and legal community, and some courts have even ruled profiling testimony inadmissible. There are two main reasons for this (Gudjonsson and Haward 1998). First, a criminal profile only gives a broad indication of the type of person who may have committed the crime. It does not indicate a specific individual who happens to fit the profile. The profiler is therefore unable to say whether it is more probable than not that a specific offender did, in fact, commit the crime. Second, there is no scientific evidence to support the reliability and validity of criminal profiling in solving crimes. Indeed, it seems that when profiling does assist the police in solving a case or in opening up new lines of inquiry, it is the exception rather than the rule.

On the other hand, it cannot be denied that criminal profiling has proven helpful in some, albeit exceptional, cases. However, much more research needs to be done before criminal profiling will earn its place as a valuable forensic tool.

A variation of profiling that has evoked some interest in the news media is psychological autopsy, which comprises the compilation of a psychological profile of especially well-known deceased individuals. It is also used in suicide cases, for example, to determine whether the deceased could indeed have committed suicide. However, as in the case of criminal profiling, its reliability and validity are also questioned.

2.9 Child Custody

With the worldwide increase in divorce rates and the resulting child custody battles between divorcing parents, mental health professionals have come to play an increasingly important role in such disputes.

Two principles play a major role in courts' decisions in child custody cases: the tender-years principle and the best-interest-of-the-child principle. The tender-years principle reflects the traditional belief that the mother is inherently better equipped to care for children of a 'tender age' (which particularly refers to children around or under the age of seven). However, several factors have led to the questioning and even abolishment of this principle in some courts. For example, research has shown that fathers and mothers are equally capable of fulfilling the psychological and other needs of children in their 'tender years.' The principle is also seen as a violation of equal gender rights. Most courts base their decisions in child custody cases on the best-interest-of-the-child principle, which involves criteria such as (a) the love, affection, and other emotional ties between parent and child, as well as between the child and his or her siblings or any other significant person; (b) the ability of the parent to provide for the child's psychological, educational, cultural, and environmental develop-

ment; (c) the ability of the parent to provide the basic physical needs of the child, e.g., food, clothing, and housing; (d) the mental and physical health and moral fitness of the parent; and (e) the child's preference as to his or her custodian. However, this principle has also come under fire from critics who, for example, claim that many of the criteria are vague, difficult to assess, and open to subjective interpretation.

Assessment in child custody disputes usually includes interviewing the parents and children, as well as any other third-party source that may provide valuable information, observation of the interaction between especially parents and child, and the use of traditional psychological and specialized tests. However, the interview is usually recommended as the best method. The reasons are the following. The observation technique is regarded as artificial and thus unreliable (the individuals know that they are being observed). The traditional psychological tests are often irrelevant (e.g., there is no correlation between good parenthood and IQ). The specialized tests are also not very reliable and valid. It is generally accepted that a psychologist, acting according to professional guidelines, can make a valuable and often decisive contribution in child custody cases.

3. Forensic Assessment

Forensic assessment refers to the procedures, methods, and techniques that psychologists use to gather information in order to assist the court in coming to a final decision. The nature of the assessment will depend on the theoretical orientation, training, experience, and preference of the psychologist, as well as the nature of the specific case. Usually a combination of procedures, methods, and techniques are used.

The forensic interview can be considered the central element of the assessment process. The interview can be structured (the questions being determined and systematized in advance), unstructured (the interview being allowed to develop spontaneously), or a combination of the two. The forensic interview is much more than just an 'ordinary' interview. It is a scientific method for gathering information and requires specialized training and experience. As mentioned, it is usually necessary to interview not only the specific individual, but also family members, friends, and other third-party sources. Such collateral information may be helpful not only in acquiring a broader picture, but also in determining the reliability of the information provided by the different sources.

In addition to this basic assessment method, several others are employed, and examples are given below.

3.1 Psychological Testing

Virtually all forensic psychologists incorporate psychological tests in their assessment armament. Psycho-

logical test results can be powerfully persuasive instruments because of society's long-standing high regard and respect for these measuring instruments (Tsushima and Anderson 1996). Many courts also seem to expect that such tests will be used.

However, it has also become clear that psychological tests have not always conformed to the high qualities attributed to them. Not only are the reliability and validity of many of the tests used by psychologists generally disputed, but also they have not been developed specifically for forensic issues. The result is that their usefulness in forensic settings is questioned. Attempts to develop specialized forensic assessment instruments have also not yet been very successful, mainly owing to psychometric flaws. Another problem especially encountered in underdeveloped and developing countries is that tests are used on groups for which they have not been developed. An example is the worldwide use of various American tests developed for the US population on groups for which they were not standardized. Cultural differences cast doubt on the reliability of such results.

On the other hand, can there be little doubt that, in the hands of a well-trained and experienced examiner who is well aware of the pitfalls, psychological tests can be very valuable tools and can provide valuable information.

3.2 Forensic Hypnosis

There are few topics in psychology more shrouded in mystery and misconception than hypnosis. Mainly responsible for this is the sensational representation of this technique by the movie industry, news media, and stage hypnotists. Even in the scientific world controversy surrounds the hypnosis phenomenon, with some regarding it as a valuable tool in many areas, while others do not believe in the existence of such a state, and mainly contribute it to suggestion and role-playing (see *Hypnosis, Psychiatry of*).

The primary use of forensic hypnosis centers on enhancing the memory of a witness, victim, or even defendant, especially with a view to helping him or her remember forgotten or vague details of a crime. Most research shows that, compared with people who have not been hypnotized, people under hypnosis tend to recall more detail about an event. They also usually feel confident that their recollections are correct. However, their stories contain a great deal of inaccurate information, even more so than in the case of nonhypnotized subjects. It is generally accepted that forensic hypnosis is seldom more efficient in improving memory than any other standard psychoforensic procedure. Courts therefore hesitate to accept testimony acquired through hypnosis. In some countries such testimony is even inadmissible. Others do allow it, but usually only if certain safeguards and guidelines

are applied. It is therefore important that forensic psychologists meet these preconditions if they wish to employ this controversial forensic tool.

3.3 Polygraph

The polygraph, also called 'lie detector,' is an apparatus that measures physiological variables such as heart rate, blood pressure, and respiration rate. Polygraphers believe that when people are asked threatening questions to which they respond by lying, physiological changes will occur which will show up on the polygraph record. This makes the polygraph a potentially valuable forensic tool.

Unfortunately, the validity of the polygraph leaves much to be desired. Polygraph results seem to be wrong in at least 30 percent of all cases. This concerns innocent people who are wrongly accused, as well as guilty persons who plead innocence. For this reason, most courts do not admit evidence based on polygraph results, or admit it only if it adheres to certain stipulations (e.g., that all parties agree to the test before it is administered).

Ironically, it seems that one of the polygraph's biggest strengths lies in people's belief that it is reliable. It seems that about 30 percent of suspects confess after they have failed a polygraph test (Memon et al. 1998). However, it does not necessarily mean that these suspects had been lying. It could also be that innocent suspects tend to confess because, having failed the test, they see little possibility of convincing the court of their innocence thereafter. By confessing they hope to get a lighter sentence.

4. The Future

Only a few decades ago the relationship between psychology and law was compared to an unhappy marriage. In the twenty-first century, it is rather regarded as a 'normal' marriage with its ups and downs, but definitely not heading for a divorce. There can be no doubt that both professions find the interaction and collaboration enriching.

However, in order to make full use of this interdependent relationship, much more needs to be done. The key lies in research, aimed not only at developing new procedures, methods, and techniques, but also at determining weaknesses in the general professional approach of forensic psychologists. Forensic psychologists should especially try to move away from the 'hired guns' stigma attached to them. This can only be achieved by telling 'the truth and nothing but the truth.' This principle can be best served by presenting objective and critical both-sides-of-the-story evidence without subjectively favoring one party.

The saying, 'It is not a question whether forensic psychology is good, but what it is good at,' should also serve as a principle guideline for acknowledging that there are areas where little or no contribution can be made and sticking to those areas where it is clear that psychologists are held in high esteem by courts. With such an approach and quality research, there can be no doubt that forensic psychology will flourish in the future.

See also: Confessions: Psychological and Forensic Aspects; Criminal Defenses; Expert Testimony; Expert Witness and the Legal System: Psychological Aspects; Forensic Psychiatry; Liability: Legal

Bibliography

- Ainsworth P B 1998 *Psychology, Law and Eyewitness Testimony*. Wiley, New York
- Allan A 1995 The role of the psychologist in South African courts. Ph.D. dissertation, University of the Orange Free State, Bloemfontein
- Alloy L B, Jacobson N S, Acocella J 1999 *Abnormal Psychology. Current Perspectives*, 8th edn. McGraw-Hill, New York
- American Psychiatric Association 1994 *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV*, 4th edn. American Psychiatric Association, Washington, DC
- Bartol C R, Bartol A M 1994 *Psychology and Law. Research and Application*, 2nd edn. Brooks/Cole, Pacific Grove, CA
- Blau T H 1998 *The Psychologist as Expert Witness*, 2nd edn. Wiley, New York
- Butcher J N, Miller K B 1999 Personality assessment in personal injury litigation. In: Hess A K, Weiner I B (eds.) *The Handbook of Forensic Psychology*, 2nd edn. New York, Wiley, pp. 104–26
- Ewing C P 1985 *Psychology Psychiatry and the Law: a Clinical and Forensic Handbook*. Professional Resource Exchange, Sarasota, FL
- Golding S L, Skeen J L, Roesch R, Zapf P A 1999 The assessment of criminal responsibility: current controversies. In: Hess A K, Weiner I B (eds.) *The Handbook of Forensic Psychology*, 2nd edn. Wiley, New York, pp. 379–408
- Gudjonsson G H, Haward L C R 1998 *Forensic Psychology. A Guide to Practice*. Routledge, London
- Melton G B, Petrila J, Poythress N G, Slobogin C 1997 *Psychological Evaluations for the Courts*, 2nd edn. Guilford Press, New York
- Memon A, Vrij A, Bull R 1998 *Psychology and Law. Truthfulness, Accuracy and Credibility*. McGraw-Hill, New York
- Saywitz K J, Goodman G S 1996 Interviewing children in and out of court. In: Briere J, Berliner L, Bulkley J A, Jenny C (eds.) *The APSAC Handbook of Child Maltreatment*. Sage, Thousand Oaks, CA
- Tsushima W T, Anderson R M Jr 1996 *Mastering Expert Testimony*. Erlbaum, Mahwah, NJ

D. Louw

Foucault, Michel (1926–84)

Michel Foucault is one of the most influential, and also most controversial, social thinkers of the century. His work represents a sustained effort to reconstruct the way in which the specifically modern Western form of subjectivity has become what it is through institutionalized practices of truth-telling and the search for knowledge.

1. Foucault's Intellectual Trajectory

Foucault was born on October 15, 1926 in Poitiers. His upbringing was organized around knowledge since his early childhood (he started to attend school at the age of three), while major political events, especially wars (the Ethiopian war, the Spanish Civil War, and World War II) also had a major impact on his formation. In September 1945 he moved to Paris where he listened to the lectures of Jean Hyppolite during a preparatory high-school year, and was then admitted to the prestigious *École Normale Supérieure*. These were troubled years for Foucault, intellectually, personally, and politically alike. He had a short spell with the Communist Party and had difficulties in managing his homosexuality.

After graduating in philosophy in 1951, he received further qualifications in 1952 and 1953 in psychopathology and experimental psychology. He then started teaching, working on his thesis and also attending the St Anne psychiatric hospital. However, under the impact of an intensive reading of Nietzsche, his interests gained focus, and he decided to leave the standard academic career, even the country.

In between 1955 and 1960 Foucault worked as a cultural diplomat, first in Uppsala (at the invitation of Georges Dumézil), then in Warsaw and Hamburg. He thus experienced personally the different economic and political regimes of the period, noticing their latent similarities, and also worked on his doctoral thesis. This was defended in 1961, under the supervision of Georges Canguilhem, and published as *Folie et déraison (Madness and Unreason)* with the support of Philippe Ariès.

After completing the book, Foucault continued his work on the reconstruction of the joint emergence of modern forms of scientific (especially medical) knowledge and its institutional context, and published in 1963 *The Birth of the Clinic*. At the same time, he also got increasingly involved in literary criticism, publishing various essays and a book on Raymond Roussel. As a result, he came to 'bracket' institutional practices temporarily in his work and turned towards language and discourse—a shift he would later much regret. The result was *The Order of Things*. Published in 1966, at the height of structuralism, the book immediately became a huge success, identifying Foucault as a main protagonist of the field.

Becoming increasingly uneasy about this label, Foucault soon embarked on a major self-reflexive exercise, attempting to clarify his project and its underlying methods for others and for himself. This eventually led to the *Archaeology of Knowledge* and a series of major methodological papers. At the same time, he moved from Paris to Tunis, and this is where he lived during the events of May 1968. However, instead of a systematic and final clarification, the reflexive exercises rather brought about a major reorganization of Foucault's entire work. Appointed as Chair of the History of Systems of Thought at the prestigious *Collège de France*, a few months after his inaugural lecture, Foucault suddenly shifted his intellectual interest from knowledge to power, and at the same time he became engaged in political struggles as the founder of GIP (group for information on prisons).

His new interests led to the publication of *Discipline and Punish*, easily his most widely read and influential work. However, around the time of its publication, Foucault was again already reflecting on his work. This yielded a book-length study published as the first of six planned volumes on the *History of Sexuality* and a series of significant and quite influential interviews and shorter pieces (see Foucault 1980). But perhaps most importantly, it led to another major reorientation of his project. When trying to specify the main characteristics of his approach to the study of power, he eventually came to realize that the question of power was not central to his work. This change—often interpreted as Foucault's 'crisis'—is visible in successive displacements of the time horizon of the work, first to the fifteenth–sixteenth centuries, then to early Christianity, and finally to Greco-Roman Antiquity; and also in a reorganization at the level of central concepts: from power-knowledge to government by truth; from the genealogy of power to the hermeneutic of the subject; from technologies of power to techniques of self. It was strongly influenced by Foucault's encounter with the works of major scholars of antiquity such as Pierre Hadot and Peter Brown. At the same time he also increasingly recognized affinities with the works of social scientists like Norbert Elias and especially Max Weber. These encounters, together with a systematic reflection on the trajectory of his past work, helped Foucault to identify the underlying problem of his work as being concerned neither with power, nor with discourse, rather with the link between truth and subjectivity (Foucault 1997, p. 281).

The new orientation is visible in the published Volumes Two and Three of the *History of Sexuality* project that discuss the way practices related to pleasure were problematized in ancient Greece and Rome, respectively. Volume Four, covering early Christianity, only exists in the form of an uncorrected manuscript, as Foucault was hospitalized, terminally ill, just days after he started to revise it for publication.

The best guides for the last, arguably most significant period of Foucault's work are the courses he

gave at the Collège de France in between 1978 and 1984, especially the last two courses on *parrhesia* or frankness of speech and truth-telling. Foucault died in Paris on June 25, 1984.

2. The Work and its Context

Throughout his writings, covering over three decades, Foucault's work focused around the reasons and manner in which the individual subject, especially its 'true being,' has become the center of attention in the pursuit of knowledge and the search for social order, considered as an evident starting point and foundation. Thus proceeding, Foucault posed question marks to what has been taken for granted by the main available approaches in philosophy and the social sciences, without dismissing them as mere ideology or illusion. This explains much of both the dynamics of Foucault's intellectual trajectory (discussed in this section) and its reception (covered in the next).

The trajectory has three major peculiarities. First of all, far from following a straightforward path of development, it was quite bumpy and curvy, progressing through crises (see Szokolczai 1998). This was probably unavoidable as Foucault's work did not simply represent an addition to the existing stock of knowledge, neither a general attack on it, rather a problematization of the taken for granted. Furthermore, and for much the same reasons, the trajectory cannot be separated from the substance of the ideas. Given its detours, it would seem particularly desirable to get over such details and focus on the main results. However, this would undermine the effectivity of Foucault's undertaking, as doing so one would fail to problematize one's own hidden related preconceptions. Finally, Foucault was quite aware of the pitfalls of his project, and throughout his career kept reflecting on its directions and stakes. These reflections had a significant effect on the trajectory and contain a concise introduction to Foucault by Foucault, a starting point for any interpretative effort.

For these reasons the exact mapping of Foucault's early steps is of crucial relevance, as this not simply defines the starting point of the investigation that would later be assumed as given, rather indicates the way in which Foucault plunged into the questioning and mapping of the aspect of the taken for granted he singled out for attention as problematic. The first anchorage point to be questioned was academic philosophy that assumed, from Descartes through Kant to Husserl and Merleau-Ponty that the 'subject of knowledge' or the 'transcendental ego' was the starting point of any philosophical undertaking. Foucault recognized that behind the epistemological assumptions of this approach there lay an entire anthropology, and was particularly interested in Kant's related work, which he translated and introduced as his minor thesis.

Given his problems with academic philosophy, he turned to psychology, or an empirical approach to the Kantian question ‘what is man.’ At the same time, he experimented with the two main alternative theoretical frameworks to academic philosophy, Marxism and psychoanalysis, finding them unsatisfactory. It is in this combination that the profound significance and originality of Foucault’s work lies. His background incorporates both theoretical and empirical work: the philosophy of the subject in its various kinds and positivism on the one hand, and their main critiques, the works of Marx and Freud on the other, while being orthogonal to this entire epistemological field.

In his attempts to go beyond the available approaches, Foucault found inspiration in the works of Nietzsche. Instead of starting from the experiences of the subject, gained through a combination of sense perception and the application of ‘universally’ valid categories of mind, Foucault rather posed the question of how this subject has been actually historically formed *through* experience.

In his last writings, Foucault would argue that there are three main axes of experience: knowledge, power, and subjectivity. This conceptualization, however, was only the fruit of late reflection on his own work. In his first major book, *Folie et déraison*, all three axes were dealt with, and their separation was not at all clear-cut. Furthermore, instead of directly reconstructing the subject of knowledge, he rather started at the opposite end, as if following a reflexive strategy, by reconstructing the opposite or the mirror, the subject of nonknowledge or the ‘nonsubject,’ the mad, the carrier of ‘unreason.’ The aim was to show that ‘mental illness’ was not simply an objective aspect of the human condition that was there, waiting to be discovered finally by the methods of true science, but as a complex form of behavior was itself shaped by these very methods and their institutional context.

In this book and the follow-up work, *The Birth of the Clinic*, Foucault tied closely together the reconstruction of the views concerning the aim and methods of collecting knowledge about individual human beings and its context of closed or total institutions, or the knowledge and the power axis. This changed with his third main book, *The Order of Things*, where the institutional dimension or actual scientific and social practice was left out of the picture. On the one hand, this enabled Foucault to sharpen his tools here and in the following, reflexive works, defining his archaeological method as ‘an attempt to introduce the diversity of systems and the play of discontinuities into the history of discourses’ (Foucault 1991). On the other hand, however, this also meant a certain one-sidedness, the suspension of the study of the link between displacements in thought and broader historical processes.

As a result, in the 1970s, Foucault reversed his approach and turned to the power axis, developing conceptual tools like power-knowledge (which never

meant a simple identification between knowledge and power, only the assertion of the impossibility of a complete separation) or the microphysics of power. This meant a shift of disciplinary allegiances from philosophy and the life sciences to sociology and the social sciences, and a reorientation of focus from the ‘subject of knowledge’ of phenomenology to the liberal individual as a subject of rights, the taken for granted starting point of modern political and social theory. Here as before, Foucault chose an indirect approach, focusing his reconstructive historical work on the formation of the subject deprived of rights: the criminal.

Discipline and Punish not only returned Foucault’s interest to institutional practices but sharpened his analytical tools concerning the power axis. However, if his work in the 1960s was slightly affected by flirtations with literary criticism and structuralism, his new work showed too much affinity with the contemporary leftist, neo-Marxist vogue, most visibly at the level of rhetorics.

Foucault’s next series of works on sexuality were intended to continue and complement the studies of power-knowledge as started in *Discipline and Punish*. However, when finishing the first, reflexive and methodological volume of the series, *La volonté de savoir*, Foucault again stepped back from his approach and undertook another major reorientation. This meant that while previously, whether focusing on knowledge or institutional practice, Foucault studied how certain human beings shaped other human beings as objects of power and knowledge, now he set out to analyze directly how individual human beings were driven to reflect on their own being as subjects.

This was accompanied by a marked change in style, even a refocusing of his entire project. While Foucault’s aim so far was to render visible the singularity of the idea that the individual subject should be the foundation of knowledge and social order, now he recognized the seriousness of the problem to which this came to be the answer. The shifting of the focus of interest to the individual was not simply an epistemological mutation, nor a particularly cunning turn in the exercise of power, but was an answer to the manner in which being has become problematized at a certain point; the way in which human beings came to pose question marks to ‘what they are, what they do, and the world in which they live’ (Foucault 1984a). While this turn on the one hand rendered visible the Heideggerian inspiration behind his work, on the other it also brought to the surface the close affinities between Foucault and Max Weber.

3. The Foucault ‘Effect’

Foucault has been one of the most influential social theorists of the past decades. His works are widely read in all branches of the humanities and the social

sciences, including philosophy, sociology, history, anthropology, cultural studies, and literary criticism. However, they have always been controversial and most probably will remain so in the future.

Both characteristics can be traced to the same source. Foucault's work developed not simply as a dissent from the then ruling paradigms in philosophy and psychology, phenomenology and positivism, but was similarly critical, after some initial flirtation, of the undertakings of Marx and Freud. Based on his reading of Kierkegaard, Heidegger, and especially Nietzsche, Foucault went into directions that did not easily fit on the contemporary intellectual landscape. This led to a persistent problem of understanding, reception, classification, and (mis)labeling. It was only deepened by some strategic concessions Foucault made when trying to establish bridges between his lonely undertaking and the work of his contemporaries.

Foucault's work first gained wide currency in the mid-1960s when, after the publication of *The Order of Things*, he was considered as a prime exponent of structuralism, together with Barthes, Lacan, and Lévi-Strauss. His archaeological method, aiming at the establishment of discontinuities at the level of discourses and considering texts written in between two breakpoints as belonging to a discursive formation, governed by certain autonomous rules, is still influential in wide circles within philosophy and literary criticism. The work largely followed the footsteps of Foucault's earlier works, relying more on Dumézil, Canguilhem, and Bachelard than on Lacan or Lévi-Strauss. It was also different from the structuralist project and mentality, as it made no attempt at developing formal schemes with claims of general validity. Still, the label 'structuralist' stuck due to the gestures Foucault made in its last chapter to the then rising structuralist tide.

Though Foucault's work of the 1970s on power-knowledge came to be inserted in a different debate and exerted its impact on a different field, the modalities were remarkably similar. As a conscious break with his earlier work, *Discipline and Punish* aimed at direct political relevance and impact. While *The Order of Things* was critical of Marxism and took up no side in the political landscape, the new work was clearly aiming at making contact with the far left and started by using a language heavily indebted to Marxism. As the work progressed, Foucault steadily retrieved his central interest in the formation of the self through 'true' discourse, in this case the shaping of the criminal. However, in its style and rhetoric, the ballast was still present, and the work came to be read and regarded as a prime example of the 'poststructuralist' and 'postmodern' alternative to old-fashioned Marxism. While in the 1960s Sartre considered Foucault's work as the 'last rampart' of the bourgeoisie, his work of the 1970s served as a lifebuoy for many on the sinking boat of the radical left.

Soon Foucault became just as deeply disturbed by this reception of his work as with the structuralist label in the 1960s. His self-reflexive exercises of the late 1970s and the early 1980s not only helped him to regain focus in his own work but also to find at the end, after some flirtation with the thought of the Frankfurt school, the most fruitful intellectual connections with the works of philosophers such as Gadamer and Hadot, and social scientists like Elias and Goffman. Most important of all, however, was his late in-depth encounter with the work of Max Weber, in whom he recognized a kindred spirit and a selected predecessor.

Though this was not lost in some segments of the Foucault reception (Dreyfus and Rabinow 1982, Gordon 1991, Owen 1994), by then the damage had already been done. Foucault thus came to be regarded as a prime exponent of the very intellectual currents whose presuppositions he wanted to question and map in his work. In this, his fate can again be best compared to Max Weber's. As a result, in the burgeoning literature for or against Foucault, the problem he started to pose almost half a century ago concerning the taken for granted, allegedly universal but quite peculiar and questionable character of the way modern Western subjects form their subjectivity (selfhood or identity)—a problem very close to the Weberian interest in inner-worldly asceticism and in the methodical conduct of life, or the interest of Elias in the formation of *habitus* through the civilizing process—became lost. This is demonstrated most paradoxically in the way the new literature on sexual, gender, race, etc., identities takes Foucault as one of its sources, although his work should rather be seen as a most effective way in which this 'politics of identity' could be criticized.

While Foucault, just like Weber, is one of the most widely read social theorists of the century, his work is similarly still more of an unfulfilled promise than a closed, exhausted chapter in the history of thought.

See also: Control: Social; Crime: Sociological Aspects; Crime, Sociology of; Gender Ideology: Cross-cultural Aspects; Heterosexism and Homophobia; Lesbians: Historical Perspectives; Lesbians: Social and Economic Situation; Medicine, History of; Postmodernism in Geography; Postmodernism in Sociology; Postmodernism: Philosophical Aspects; Psychiatry, History of; Rationality and Feminist Thought; Transsexuality, Transvestism, and Transgender

Bibliography

- Dreyfus H L, Rabinow P 1982 *Michel Foucault: Beyond Structuralism and Hermeneutics*. University of Chicago Press, Chicago
- Eribon D 1991 *Michel Foucault*. Harvard University Press, Cambridge, MA

- Foucault M 1961 *Folie et déraison: Histoire de la folie à l'âge classique*. Plon, Paris [1965 *Madness and Civilization*. Harper, New York] {partial version}
- Foucault M 1963 *Naissance de la clinique: Une archéologie du regard médical*. Presses Universitaires de France, Paris [1975 *The Birth of the Clinic*. Vintage, New York]
- Foucault M 1966 *Les mots et les choses: Une archéologie des sciences humaines*. Gallimard, Paris [1973 *The Order of Things*. Vintage, New York]
- Foucault M 1969 *L'archéologie du savoir*. Gallimard, Paris [1976 *The Archaeology of Knowledge*. Harper, New York]
- Foucault M 1975 *Surveiller et punir: Naissance de la prison*. Gallimard, Paris [1979 *Discipline and Punish*. Vintage, New York]
- Foucault M 1976 *La volonté de savoir: Histoire de la sexualité 1*. Gallimard, Paris [1980 *The History of Sexuality, Volume 1: An Introduction*. Vintage, New York]
- Foucault M 1980 *Power/Knowledge: Selected Interviews and Other Writings by Michel Foucault, 1972–1977*, Gordon C (ed.). Harvester Press, Brighton, UK
- Foucault M 1984a *L'usage des plaisirs: Histoire de la sexualité 2*. Gallimard, Paris [1986 *The Use of Pleasure*. Vintage, New York]
- Foucault M 1984b *Le souci de soi: Histoire de la sexualité 3*. Gallimard, Paris [1987 *The Care of the Self*. Vintage, New York]
- Foucault M 1984c *The Foucault Reader*. Rabinow P (ed.). Pantheon, New York
- Foucault M 1991 Politics and the study of discourse. In: Burchell G, Gordon C, Miller P (eds.) *The Foucault Effect: Studies in Governmentality*. Harvester Wheatsheaf, London
- Foucault M 1994 *Dits et écrits*. Defert D, Eward F (eds.). Gallimard, Paris, 4 Vols
- Foucault M 1997 The ethics of the concern of the self as a practice of freedom. In: Rabinow P (ed.) *Ethics: Subjectivity and Truth, The Essential Works of Michel Foucault*. The New Press, New York, Vol. 1
- Gordon C 1991 Introduction. In: Burchell G, Gordon C, Miller P (eds.) *The Foucault Effect: Studies in Governmentality*. Harvester Wheatsheaf, London
- Owen D 1994 *Maturity and Modernity: Nietzsche, Weber, Foucault and the Ambivalence of Reason*. Routledge, London
- Szokolczai A 1998 *Max Weber and Michel Foucault: Parallel Life-Works*. Routledge, London

A. Szokolczai

Fourier Analysis

1. Fourier Series

1.1 Trigonometric Functions

The term Fourier analysis originally refers to the possibility to approximate functions (up to scaling) in the open unit interval $I :=]-1/2, +1/2[$ by linear combinations of trigonometric functions

$$t_0(x) := 1, t_{2n}(x) := \sqrt{2} \cos(2\pi nx), \\ t_{2n-1}(x) := \sqrt{2} \sin(2\pi nx) \quad \text{for } n = 1, 2, 3, \dots \quad (1)$$

and commemorates the role that the French baron Jean-Baptiste-Joseph Fourier played in furthering the development of this method. (Here, the colon marks that side of an equation which is defined by the other side; thus, the symbol ‘:=’ could be read as ‘defined as.’) Fourier’s influential work has its precursors in early studies of the vibrating string by the Swiss mathematicians Daniel Bernoulli and Leonhard Euler. The formulas for what are nowadays called Fourier coefficients date back as early as 1728 and are due to Bernoulli. These formulas are best understood by invoking the geometric metaphors of length and orthogonality, which can be described in terms of a so-called inner product $\langle \cdot | \cdot \rangle$, in our current context specifically given by

$$\langle f | g \rangle := \int f(x)^* g(x) dx \quad (2)$$

for f, g in the set $BC(I)$ of bounded, continuous, complex-valued functions on I (here z^* denotes the complex conjugate of the complex number z ; for real numbers r we have $r^* = r$). As a matter of convenience and simplicity of notation we may consider the integral in Eqn. (2) to extend over the whole real line $\mathbb{R}$ by considering f, g as well as the functions $t_k, k = 0, 1, 2, \dots$, as defined on $\mathbb{R}$ but vanishing outside the interval I .

The inner product induces the concept of a distance between two functions $f, g \in BC(I)$ defined by $\|f - g\| := \sqrt{\langle f - g | f - g \rangle}$. Here, ‘... $\in$...’ should be read as ‘... element(s) of ...’. In particular, the distance of such a function f to the constant function 0 (with value zero for all arguments) $\|f\| = \|f - 0\| = \sqrt{\langle f | f \rangle}$ is called the norm (or length) of f . Functions f with norm $\|f\| = 1$ are called normalized. Two functions f and g are called orthogonal when their inner product vanishes, that is, when $\langle f | g \rangle = 0$. In fact, the specific functions of the family $(t_n)_{n \in \mathbb{N}}$, where $\mathbb{N} := \{0, 1, 2, \dots\}$ denotes the set of natural numbers, are all normalized and pairwise orthogonal if we interpret t_n to be zero outside of the unit interval I . A family of functions such as $(t_n)_{n \in \mathbb{N}}$ is, therefore, also called an orthonormal system. The interest in such orthonormal systems is based on the simplicity with which ‘best approximations’ to given elements can be found in terms of linear combinations of elements of such orthonormal systems. Indeed, let $f \in BC(I)$ be given, then

$$\|f - \sum_{k=0}^N \langle t_k | f \rangle t_k\| \quad (3)$$

is the smallest possible distance between f and any linear combination of the t_k for $k = 0, 1, 2, \dots, N$. The so-called Fourier coefficients $\langle t_k | f \rangle, k = 0, \dots, N$, are obviously independent of $N \in \mathbb{N}$ as N grows. It should be emphasized that this feature is indeed shared by any orthonormal system. The term Fourier coefficients is also used in the general situation. For sake of brevity,

however, we will have to focus here and in the following predominantly on the issue of Fourier analysis in a narrower sense.

1.2 Fourier Series

Letting $N \rightarrow \infty$ one finds that the quantity in Eqn. (3) tends indeed to zero. This property is known as the completeness of the orthonormal system $(t_n)_{n \in \mathbb{N}}$. According to the Fischer–Riesz theorem the convergence of $\sum_{k=0}^{\infty} c_k t_k$ (with respect to the norm $\|\cdot\|$) is determined completely by the convergence of the numerical series $\sum_{k=0}^{\infty} |c_k|^2$. It turns out that there are more functions than just functions in $BC(I)$ which can be approximated by Fourier series, that is, by series of the form $\sum_{k=0}^{\infty} c_k t_k$ (assuming only that the Fourier coefficients are such that $\sum_{k=0}^{\infty} |c_k|^2$ exists). The set of all such functions is denoted by $L_2(I)$, which is the set of measurable, square-integrable functions. Identifying functions $f, g \in L_2(I)$ if $\|f - g\| = 0$, we obtain a 1-to-1 and onto correspondence between elements of $L_2(I)$ and complex number sequences in $\ell_2(\mathbb{N}) := \{(c_k)_{k \in \mathbb{N}} \mid \sum_{j=0}^{\infty} |c_j|^2 \text{ convergent}\}$. Indeed, one finds

$$\|f\| = \left(\sum_{j=0}^{\infty} |\langle t_j | f \rangle|^2 \right)^{1/2} \quad (4)$$

for all $f \in L_2(I)$ and

$$f = \sum_{j=0}^{\infty} \langle t_j | f \rangle t_j \quad (5)$$

Thus, the mapping $\mathcal{T}$ —defined as the association of $f \in L_2(I)$ with the sequence $(\langle t_k | f \rangle)_{k \in \mathbb{N}}$ —is norm-preserving if we define

$$\|(c_k)_{k \in \mathbb{N}}\|_{\#} := \left(\sum_{j=0}^{\infty} |c_j|^2 \right)^{1/2} \quad (6)$$

for $(c_k)_{k \in \mathbb{N}} \in \ell_2(\mathbb{N})$ as norm in $\ell_2(\mathbb{N})$, that is

$$\|\mathcal{T}f\|_{\#} = \|f\| \quad (7)$$

This means in particular that all the information of f is preserved just in the sequence of the Fourier coefficients.

The advantage of using the functions in Eqn. (1) as an orthonormal system is that real-valued functions result in real Fourier coefficients. However, using the equivalent complex orthonormal system $(\eta_k)_{k \in \mathbb{Z}}$

$$\begin{aligned} \eta_k(x) &:= \exp(2\pi i k x) \\ &= \cos(2\pi k x) + i \sin(2\pi k x) \quad \text{for } x \in I \end{aligned} \quad (8)$$

for $k \in \mathbb{Z} := \{\dots, -3, -2, -1, 0, +1, +2, +3, \dots\}$, yields a simpler, more unified description $(\eta_k)_{k \in \mathbb{Z}} =$

$(\exp(2\pi i k \cdot))_{k \in \mathbb{Z}}$. The relation between the new Fourier coefficient $(\langle \exp(2\pi i j \cdot) | f \rangle)_{j \in \mathbb{Z}}$ and the previous ones is (due to Eqn. (8)) easily established. Similarly, as before we obtain

$$f = \sum_{j=-\infty}^{\infty} \langle \exp(2\pi i j \cdot) | f \rangle \exp(2\pi i j \cdot) \quad \text{in } L_2(I) \quad (9)$$

and the mapping $\mathcal{F}_{\#}$ associating the sequence of Fourier coefficients

$$\begin{aligned} &(\langle \exp(2\pi i k \cdot) | f \rangle)_{k \in \mathbb{Z}} \in \ell_2(\mathbb{Z}) \\ &:= \left\{ (c_k)_{k \in \mathbb{Z}} \mid \sum_{j=-\infty}^{\infty} |c_j|^2 \text{ convergent} \right\} \end{aligned} \quad (10)$$

with the element $f \in L_2(I)$ is also norm-preserving, that is

$$\|\mathcal{F}_{\#} f\|_{\#} = \|f\| \quad (11)$$

2. The Fourier Transform

The Fourier coefficients $(\langle \exp(2\pi i j \cdot) | f \rangle)_{j \in \mathbb{Z}}$ suggest the following definition

$$\hat{f}(p) := \langle \exp(2\pi i p \cdot) | f \rangle \quad \text{for } p \in \mathbb{R} \quad (12)$$

where f should now vanish outside a bounded interval (rather than just the unit interval I). The mapping $\mathcal{F}$ that associates with f the function $\hat{f}$ is called the Fourier transform (see Gaskill 1978, Goodman 1968 as general references). The function $\hat{f} = \mathcal{F}f$ is referred to as the Fourier transform (or the spectrum) of f . In view of the Fourier coefficients, we note that any such function can be forced into $L_2(I)$ by rescaling with a suitably large scaling factor $L \in \mathbb{R}$: $\sigma_L f := \sqrt{L} f(L \cdot)$. The Fourier transform interacts with rescaling in the following way

$$\begin{aligned} \mathcal{F}(\sigma_L f) &= \sqrt{|L|} \mathcal{F}(f(L \cdot)) \\ &= \frac{1}{\sqrt{|L|}} (\mathcal{F}f)(\cdot/L) \\ &= \sigma_{1/L}(\mathcal{F}f) \quad \text{for } L \in \mathbb{R}, L \neq 0 \end{aligned} \quad (13)$$

Any such function may also be considered as an element of $L_2(\mathbb{R})$ (the set of measurable, square-integrable functions on $\mathbb{R}$).

Using rescaling with -1 , that is, reflection at the origin, we get $\mathcal{F}^* := \sigma_{-1} \mathcal{F}$ as the inverse Fourier transform and so we are led to the so-called Fourier inversion formula

$$f = \int \exp(2\pi i p \cdot) \hat{f}(p) dp =: \mathcal{F}^* \hat{f} \quad (14)$$

From Eqn. (14) it can be concluded that $\mathcal{F}$ is norm-preserving (and so also preserves the inner product $\langle \cdot | \cdot \rangle$)

$$\|\mathcal{F}f\| = \|f\| \quad (15)$$

which in turn allows the extension of the Fourier transform (as well as Eqns. (14), (15), (13)) by continuity to all $f \in L_2(\mathbb{R})$ (although then the original integral Eqns. (13) and (14) do not in general make sense any more). Analogously to Eqn. (15), the inverse Fourier transform $\mathcal{F}^*$ also preserves the inner product. We note the following important relation between $\mathcal{F}$ and its inverse $\mathcal{F}^{-1} := \mathcal{F}^*$

$$\langle \mathcal{F}f | g \rangle = \langle f | \mathcal{F}^*g \rangle \quad (16)$$

for all $f, g \in L_2(\mathbb{R})$. Equation (15) shows again that by transforming with $\mathcal{F}$ no information is lost. This information is, however, encoded in such a form that local information in $\hat{f}$ represents global information in the original f . This feature is utilized (actually for the two-dimensional case; see the brief discussion in Sect. 5) in image processing where local information is removed from or added to the Fourier transform $\hat{f}$ of an image f (modeled as an illumination distribution function f) to actually improve the visual appearance of the original f by, for example, removing grid structures or sharpening edges. The visual system appears to perform image processing, at least approximately, in this manner (see Cornsweet et al. 1984, Georgeson 1979, Graham 1981). Another application of the particular properties of Fourier transform encoding can be found in holography (see Stroke 1969). Holographic encoding has been utilized as a model for particular features in memory and perception (Pribram et al. 1974). The resistance of the original signal f against ‘noise’ afflicting the Fourier transformed signal $\hat{f}$ is also utilized in data transmission security.

In concrete applications one has to deal with real quantities (the complex notation just being a matter of formal convenience). In this context the so-called power spectrum $|\hat{f}(\cdot)|^2$ is of particular interest (the inverse Fourier transform of the power spectrum is known as the autocorrelation function). Also note that if f is even, that is, $f(-x) = f(x)$, then $(\mathcal{F}_{\cos} f)(x) := \langle \cos(2\pi x \cdot) | f \rangle = (\frac{1}{2}(\mathcal{F} + \mathcal{F}^*)(f))(x)$ (the Fourier cosine transform) and if f is odd, that is, $f(-x) = -f(x)$, then $(\mathcal{F}_{\sin} f)(x) := \langle \sin(2\pi x \cdot) | f \rangle = (\frac{1}{2i}(\mathcal{F} - \mathcal{F}^*)(f))(x)$ (the Fourier sine transform) all of which are real quantities. The Fourier cosine transform would recover the even part $E(f)(x) := \frac{f(x) + f(-x)}{2}$ of f from $\hat{f}$ and the Fourier sine transform its odd part $O(f)(x) := \frac{f(x) - f(-x)}{2}$. Readers who are uncomfortable with complex numbers may restrict their attention to odd and even functions separately (the less mathematically inclined reader may also find in Weisstein 1980 an interesting and entertaining survey on Fourier

transforms). Their information is then fully preserved by the (real) Fourier cosine and the (real) Fourier sine transform, respectively.

To give at least a flavor of how Fourier transforms act, we give two interesting examples (for a visual approach to Fourier transforms see Harburn et al. 1975). The Gauss distribution function γ —well-known from statistics—is given by $\gamma(x) := \exp(-\pi x^2)$ for $x \in \mathbb{R}$. It has the remarkable property that

$$\mathcal{F}_{\cos} \gamma = \mathcal{F} \gamma = \gamma \quad (17)$$

Let χ_M denote the characteristic function of a set M (given by

$$\chi_M(x) := \begin{cases} 1 & \text{for } x \in M \\ 0 & \text{for } x \notin M \end{cases} \quad (18)$$

for $x \in \mathbb{R}$). The characteristic function $\Pi := \chi_I$ of the above unit interval I has the so-called sinc function as Fourier transform

$$\mathcal{F}_{\cos} \Pi = \mathcal{F} \Pi = \text{sinc} \quad (19)$$

where

$$\text{sinc}(x) := \begin{cases} \frac{\sin(\pi x)}{\pi x} & \text{for } x \neq 0 \\ 1 & \text{for } x = 0 \end{cases} \quad (20)$$

for all real x . The sinc function is important in connection with the Shannon sampling theorem.

A discretization of Eqn. (14) yields the finite sum

$$\exp(-2\pi i(N/(2L)) \cdot) \sum_{j=0}^{N-1} \hat{f}((j-N/2)/L) \times \exp(2\pi i(j/L) \cdot) \frac{1}{L} \quad (21)$$

as an approximation for f as N and L become large. Representing f by the samples of

$$f((\cdot - N/2)/L) \quad (22)$$

at the points $j = 0, 1, 2, \dots, N-1, N$ and letting $L = \sqrt{N}$ we may consider Eqn. (21) as a product of a column vector by the matrix $\frac{1}{\sqrt{N}}[\exp(2\pi i(jk/N))]_{j,k=0,1,\dots,N-1}$ approximating f . This matrix is unitary and known as the inverse discrete Fourier transform (IDFT). Consequently, its inverse, the unitary matrix

$$\frac{1}{\sqrt{N}}[\exp(-2\pi i(jk/N))]_{j,k=0,1,\dots,N-1} \quad (23)$$

is called the discrete Fourier transform (DFT). For practical purposes it is worth noting that the number

of numerical operations required to calculate the discrete Fourier transform (or its inverse) can be dramatically reduced by a factorization technique known as the fast Fourier transform (FFT).

The connection to the Fourier series transform $\mathcal{F}_\#$ is apparent. However, since the transform $\mathcal{F}_\#$ can be recovered from the Fourier transform $\mathcal{F}$ (see Eqn. (45) below), we will here and from now on focus on the latter.

3. The Fourier Transform as a Spectral Representation

3.1 Derivatives

The significance of representing a function in terms of its Fourier transform becomes apparent if we consider how that transform interacts with differentiation ∂ . Introducing for convenience the symbol $D := \frac{1}{2\pi i} \partial$ we find that

$$D = \mathcal{F}^{-1} \mathbf{m} \mathcal{F} \text{ in } L_2(\mathbb{R}) \quad (24)$$

where $\mathbf{m}$ stands for the multiplication-by-argument operation, that is

$$(\mathbf{m}f)(p) := pf(p) \text{ for all } f \in L_2(\mathbb{R}), p \in \mathbb{R} \quad (25)$$

It should be noted that being in the domain of the differential operator D (as naturally characterized by the well-definedness of the right-hand side of Eqn. (24)) implies a generalized concept of differentiability. The elements in the domain turn out to be even in $BC(\mathbb{R})$. Equation (24)—it is called a spectral representation of D —shows the equivalence of differentiation $\frac{1}{2\pi i} \partial$ with multiplication-by-argument $\mathbf{m}$. The obvious symmetry of $\mathbf{m}$, that is

$$\langle \mathbf{m}f | g \rangle = \langle f | \mathbf{m}g \rangle \quad (26)$$

implies the symmetry of D , that is

$$\langle Df | g \rangle = \langle f | Dg \rangle \text{ in } L_2(\mathbb{R}) \quad (27)$$

3.2 Functions of D

The Fourier transform spectral representation in Eqn. (24) gives rise to an extension to polynomials in D and $\mathbf{m}$, respectively. Equation (24) thus suggests to generally define functions of D by letting

$$S(D) := \mathcal{F}^{-1} S(\mathbf{m}) \mathcal{F} \text{ in } L_2(\mathbb{R}) \quad (28)$$

where S can be fairly general (and even matrix-valued, in which case the Fourier transform is to be carried out component-wise). This concept is extremely powerful, since a large class of models discussed in so-called linear systems theory (see Gaskill 1978, Norman 1981)

can be understood as such functions $S(D)$ acting on an input u resulting in an output $S(D)u$. Characterizing properties of such systems are that their input–output relations are linear and translation invariant, that is, sums, multiples, and shifts of input result in sums, in (up to scaling) the same multiple, and equally shifted output. Indeed, translation τ_h defined by its action on functions f as a shift of the argument

$$\tau_h f := f(\cdot + h) \quad (29)$$

is itself a function of D in the above sense

$$\tau_h = \exp(2\pi i h D) = \mathcal{F}^{-1} \exp(2\pi i h \mathbf{m}) \mathcal{F} \quad (30)$$

It may be interesting to note that by substituting the series expansion for the exponential function, this formula (if applied to suitable functions) is recognized as nothing but the Taylor expansion. Such translation-invariant, linear systems occur (usually in their higher-dimensional form, compare Sect. 5) as good approximations of many aspects of visual and auditory perception.

3.3 Convolution

The action of a function $\hat{S}(D)$, where $\hat{S}$ is the Fourier transform of S defines what is called convolution (of S with g)

$$S * g := \hat{S}(D)g \quad (31)$$

which is defined for any g in the domain of $\hat{S}(D)$. In many cases, $S * g$ is actually given as an integral expression, the convolution integral

$$(S * g)(x) = \int S(x - y)g(y) dy \quad (32)$$

In accordance with Eqn. (31) for such convolution integrals the so-called convolution theorem holds

$$\mathcal{F}(S * g) = \hat{S} \hat{g} = \hat{S}(\mathbf{m}) \hat{g} \quad (33)$$

Particularly, the above-mentioned relationship between the autocorrelation and the power spectrum reveals that the autocorrelation is a convolution.

4. Generalized Fourier Transform and Some Applications

Generalizing the meaning of $\langle \cdot | \cdot \rangle$ we introduce generalized functions f (also called distributions) as linear functionals by prescribing the value of

$$f(\varphi) := \langle f | \varphi \rangle \quad (34)$$

for any smooth φ vanishing outside a bounded interval in $\mathbb{R}$ —we shall write $\varphi \in C_0^\infty(\mathbb{R})$ —such that linearity holds, that is

$$\langle f | \varphi + \alpha \psi \rangle = \langle f | \varphi \rangle + \alpha \langle f | \psi \rangle \quad (35)$$

for all complex numbers α and all $\varphi, \psi \in C_0^\infty(\mathbb{R})$. Clearly, any $f \in L_2(\mathbb{R})$ is also such a generalized function, but also $\exp(2\pi i p \cdot)$, although not an element of $L_2(\mathbb{R})$, is such a generalized function, as Eqn. (12) of the Fourier transform shows. A generalized function, which is indeed *not* given by a function, is featured by the distribution $\delta_{\{t\}}$ defined by

$$\delta_{\{t\}}(\varphi) = \langle \delta_{\{t\}} | \varphi \rangle := \varphi(t) \quad \text{for all } \varphi \in C_0^\infty(\mathbb{R}) \quad (36)$$

(this distribution apparently takes a sample of the ‘test function’ φ at point $t \in \mathbb{R}$). The special choice $\delta := \delta_{\{0\}}$ is the so-called Dirac δ -distribution. Since it can only be distinguished from 0 by testing with $\varphi \in C_0^\infty(\mathbb{R})$ satisfying $\varphi(0) \neq 0$ we say $\delta = 0$ on $\mathbb{R} \setminus \{0\}$. However, δ is (by definition) not identically zero as a functional. It is particularly useful in many contexts to have $\delta_{\{t\}}$ ($= \tau_{-t}\delta$) as a model for something very small but nevertheless non-negligible such as a mass point, an elementary charge, a light point, a short impulse, or a molecule at location t (see (Sect. 5) for the particularly interesting higher-dimensional case). With sufficient care a large class of such distributions f (e.g., the so-called tempered distributions) can be established for which the concept of derivative, multiplication by the argument, and Fourier transform can (guided by Eqns. (27) and (16)) be generalized by defining

$$\begin{aligned} \langle Df | \varphi \rangle &:= \langle f | D\varphi \rangle \\ \langle \mathcal{F}f | \varphi \rangle &:= \langle f | \mathcal{F}^*\varphi \rangle \\ \langle \mathcal{F}^*f | \varphi \rangle &:= \langle f | \mathcal{F}\varphi \rangle \end{aligned} \quad (37)$$

for all $\varphi \in C_0^\infty(\mathbb{R})$. As an example for this generalized differentiation we find for the characteristic function $\chi_{]0, \infty[}$ of the interval $]0, \infty[$ of all positive real numbers

$$D\chi_{]0, \infty[} = \frac{1}{2\pi i} \delta \quad (38)$$

The Fourier transform of the distribution $\delta_{\{t\}}$ is according to Eqn. (37)

$$\mathcal{F}\delta_{\{t\}} = \exp(-2\pi i t \cdot) \quad (39)$$

and conversely we have

$$\mathcal{F}\exp(2\pi i t \cdot) = \delta_{\{t\}} \quad (40)$$

The latter fact can be used to detect oscillatory behavior of frequency t which may be hiding in seemingly random data. The oscillation would show up in the Fourier transformed data as a ‘peak’ at point t .

The concept of functions of D and in particular the concept of convolution can also be carried over to tempered distributions. This will be illustrated by a few applications.

4.1 Impulse Response

The importance of the Dirac δ -distribution is that a translation-invariant, linear system $S(D)$ is completely described by its impulse response $S(D)\delta$. Indeed, the general response $S(D)f$ is given by convolution with the impulse response

$$S(D)f = S(D)(\delta * f) = (S(D)\delta) * f \quad (41)$$

In the case that $S(D)$ is the solution operator $S(D) = P(D)^{-1}$ of the differential equation $P(D)u = f$, P being a polynomial, the impulse response is also known as a fundamental solution (or Green’s function).

4.2 Shannon’s Sampling Theorem

Another distribution of particular interest is the (sampling or) comb distribution $\text{III} := \sum_{n=-\infty}^{+\infty} \delta_{\{n\}}$ which takes samples at every integer point. This distribution is reproduced by the Fourier transform, that is

$$\mathcal{F}\text{III} = \text{III} \quad (42)$$

(which is another way of stating the Poisson summation formula). If $f \in L_2(I)$ then $\text{III} * f$ is simply the periodic extension of f to all of $\mathbb{R}$. Therefore, by ‘cutting-off’ with the characteristic function Π we recover

$$f = \Pi(\text{III} * f) \quad (43)$$

As we now demonstrate, the so-called Shannon sampling theorem follows from Eqn. (42). In fact, applying Eqn. (43) to $\hat{f} \in L_2(I)$ we get after the inverse Fourier transformation has been applied that

$$\begin{aligned} f &= \mathcal{F}^*(\Pi(\text{III} * \hat{f})) \\ &= \text{sinc} * \left(\sum_{n=-\infty}^{+\infty} f(n) \delta_{\{n\}} \right) \\ &= \sum_{n=-\infty}^{+\infty} f(n) \text{sinc}(\cdot - n) \end{aligned} \quad (44)$$

Any function f with $\hat{f}$ vanishing outside a bounded interval is called band limited. Since any such function can be easily rescaled to have $\hat{f} \in L_2(I)$, Eqn. (44) shows that any band-limited function f can be completely recovered from equidistant sampling. It is worth noting that by interchanging the role of f and $\hat{f}$ in the Shannon sampling theorem we obtain

$$\hat{f} = \mathcal{F}(\Pi(\text{III} * f)) = \sum_{n=-\infty}^{+\infty} \hat{f}(n) \text{sinc}(\cdot - n) \quad (45)$$

and after applying the inverse Fourier transform the Fourier series Eqn. (9) is recovered. The sampling theorem has many applications, in particular in the analysis of auditory and visual perception (see, e.g., Yellot 1990).

4.3 Wavelets

It is interesting to note that $(\tau_{-n}\text{sinc})_{n \in \mathbb{Z}}$ is also an orthonormal system in $L_2(\mathbb{R})$. By virtue of $(\tau_{-n}\text{sinc})_{n \in \mathbb{Z}}$ being an orthonormal system we get (as for the trigonometric series)

$$\sum_{n=-\infty}^{+\infty} \langle \tau_{-n}\text{sinc} | g \rangle \tau_{-n}\text{sinc} \quad (46)$$

as the best approximation of g in these terms. According to Eqn. (44) we have that for functions with bandwidth less than or equal to 1 (i.e., $\hat{f} \in L_2(I)$) g actually is represented by this sum. To cover all band-limited functions one would have to allow for sufficiently large scaling factors L . Letting, for example, $L = 2^p, p \in \mathbb{Z}$, then $(\sigma_{2^p}\tau_{-n}\text{sinc})_{p,n \in \mathbb{Z}}$ can be used to approximate any element $f \in L_2(\mathbb{R})$. All elements of $(\sigma_{2^p}\tau_{-n}\text{sinc})_{p,n \in \mathbb{Z}}$ are generated by the combination $\sigma_{2^p}\tau_{-n}$ of scaling and translation from the single function sinc , the so-called scaling function of this particular construction. Unfortunately, $(\sigma_{2^p}\tau_{-n}\text{sinc})_{p,n \in \mathbb{Z}}$ is only orthonormal for fixed $p \in \mathbb{Z}$. By a suitable orthonormalization procedure (multiresolution analysis) one can obtain, however, that $(\sigma_{2^p}\tau_{-n}(\text{sinc} - \sqrt{2}\sigma_2(\text{sinc})))_{p,n \in \mathbb{Z}}$ is an—even complete—orthonormal system in $L_2(\mathbb{R})$. Any orthonormal system generated by $\sigma_{2^p}\tau_{-n}, p, n \in \mathbb{Z}$, from a so-called mother wavelet—here $\text{sinc} - \sqrt{2}\sigma_2(\text{sinc})$ —is called an orthonormal wavelet basis. Wavelets derive their utility largely from the fact that they do not have a fixed base scale like, for example, the Fourier system $(\exp(2\pi i n \cdot))_{n \in \mathbb{Z}}$, which by allowing for translations would also yield an orthonormal system $(\tau_r(\exp(2\pi i n \cdot)))_{n,r \in \mathbb{Z}}$ for $L_2(\mathbb{R})$. Wavelets naturally allow for the ‘right’ scale in $L = 2^p, p \in \mathbb{Z}$, to emerge for any given f according to their weight in its orthonormal expansion. As the Fourier transform may be considered to be a continuously scaled Fourier series, there are also continuously scaled versions of wavelet expansions known as wavelet transforms.

5. Higher-dimensional Fourier Transform

Much of what has been explored above can be generalized to higher dimensions. This is due to the particular product structure of the higher-dimensional exponential

$$\begin{aligned} & \exp(2\pi i(x_1 p_1 + x_2 p_2 + \dots + x_n p_n)) \\ &= \exp(2\pi i x_1 p_1) \exp(2\pi i x_2 p_2) \dots \exp(2\pi i x_n p_n) \end{aligned} \quad (47)$$

for $x_i, p_i \in \mathbb{R}, i = 1, 2, \dots, n$, that is, $(x_1, x_2, \dots, x_n), (p_1,$

$p_2, \dots, p_n) \in \mathbb{R}^n$. Thus, the n -dimensional Fourier transform, $n \in \mathbb{N}, n \geq 2$, can be understood as a repeated one-dimensional Fourier transform, that is

$$\begin{aligned} (\mathcal{F}f)(p) &= \iint \dots \int \exp(-2\pi i x p) f(x) dx_1 dx_2 \dots dx_n \\ &= \int \exp(2\pi i x_1 p_1) \int \exp(-2\pi i x_2 p_2) \dots \\ &\dots \int \exp(-2\pi i x_n p_n) f(x) dx_n \dots dx_2 dx_1 \end{aligned} \quad (48)$$

where $x = (x_1, x_2, \dots, x_n), p = (p_1, p_2, \dots, p_n)$, and $xp = \sum_{i=1}^n x_i p_i$. As a consequence, many of the above considerations can be carried over to the higher-dimensional case. For example, with $x^2 := xx, x \in \mathbb{R}^n$, the definition of the Gauss distribution function carries over to the n -dimensional case and Eqn. (17) remains true. The n -dimensional Dirac δ -distribution $\delta_{\{x\}}$ simply takes a single sample at $x \in \mathbb{R}^n$ and still has Fourier transform $\exp(-2\pi i x \cdot)$. Also, $\text{III} := \sum_{x \in \mathbb{Z}^n} \delta_{\{x\}}$, which takes samples at every point of $\mathbb{R}^n$ with integer coordinates, still satisfies

$$\mathcal{F}\text{III} = \text{III} \quad (49)$$

Thus, in particular, corresponding variants of the Shannon sampling theorem and of the Fourier series expansion hold. The scaling, however, has to be generalized to $\sigma_L f := |L|^{n/2} f(L \cdot)$, but then Eqn. (13) also carries over to the n -dimensional case. The higher-dimensional Fourier transform finds its applications in higher-dimensional, translation-invariant linear systems (Norman 1981). A particularly prominent application is in optics (see, e.g., Gaskill 1978, Goodman 1968, Taylor 1978). As pointed out earlier, the visual system may be considered at least approximately to be a two- or three-dimensional example and may be analyzed with Fourier techniques (see e.g., Cornsweet et al. 1984, Georgeson 1979, Graham 1981, Julesz 1980, Yellot 1990).

See also: Bernoulli, Jacob I (1654–1705); Computational Imagery; Pattern Matching: Methodology; Spatial Pattern, Analysis of; Spectral Analysis; Statistical Pattern Recognition

Bibliography

- Cornsweet T N, Wandell B, Yellot J I 1984 The beginnings of visual perception: The retinal image and its initial encoding. In: Darian-Smith I (ed.) *Handbook of Physiology, Sect. 1. The Nervous System. Vol. 3: Sensory Processes*. American Physiological Society, Bethesda, MD pp. 257–316
- Gaskill J D 1978 *Linear Systems, Fourier Transforms and Optics*. Wiley, New York
- Georgeson M 1979 Spatial Fourier analysis and human vision. In: Sutherland N S (ed.) *Tutorial Essays in Psychology*. Erlbaum, Hillsdale, NJ, Vol. 2, pp. 39–88
- Goodman J W 1968 *Introduction to Fourier Optics*. McGraw-Hill, San Francisco
- Graham N 1981 The visual system does a crude Fourier analysis of patterns. *SIAM-AMS Proceedings*, 13, pp. 1–16
- Harburn G, Taylor C A, Welberry T R 1975 *Atlas of Optical Transforms*. Bell & Sons, London

- Julesz B 1980 Spatial-frequency channels in one-, two-, and three-dimensional visions: Variations on an auditory theme by Bekesy. In: Harris C S (ed.) *Visual Coding and Adaptability*. Lawrence Erlbaum Associates, Hillsdale, NJ
- Norman F M 1981 Lectures on linear systems theory. *Journal of Mathematical Psychology* **23**: 1–89
- Pribram K, Nuwer M, Baron R 1974 The holographic hypothesis of memory structure in brain function and perception. In: Krantz D H, Atkinson R C, Luce R D, Suppes P (eds.) *Contemporary Developments in Mathematical Psychology*. W. H. Freeman, San Francisco, Vol. 2, pp. 46–57
- Stroke G W 1969 *An Introduction to Coherent Optics and Holography*. Academic Press, New York
- Taylor C A 1978 *Images: A Unified View of Diffraction and Image Formation with All Kinds of Radiation*. Wykenam Publications, London
- Weisstein N 1980 The joy of Fourier analysis. In: Harris C S (ed.) *Visual Coding and Adaptability*. Lawrence Erlbaum Associates, Hillsdale, NJ
- Yellot J I 1990 The photoreceptor mosaic as an image sampling device. In: *Advances in Perception. Proc. Symp. on Frontiers of Visual Science. National Research Council*. National Academy Press, Washington, DC

R. Picard

Fourth World

The Fourth World is both an imagined community and a political coalition of indigenous peoples, transcending the borders of nation states. Indigenous identity is grounded in a people's claim to be descended from the original inhabitants of a country, vis-à-vis ethnic groups which came later as conquerors and colonists; it also implies a radical difference from other life ways. Thus, the Fourth World aligns indigenous peoples with the Third World (see *Third World*) in opposition to the neocolonialism of the First (capitalist) World, but opposes them to both and to the Second (communist) World, in rejecting modernity in favor of spirituality and closeness to nature. Notwithstanding their weakness in terms of conventional political resources, and their often minuscule numbers, Fourth World peoples were able to achieve substantial recognition in the second half of the twentieth century, by linking local with global mobilization.

1. Recognizing Indigenous Peoples

The current use of the term 'indigenous' has its origins in the European colonization of the New World, and the settlers' attempt to make their homes in places that were already inhabited. It can be dated conveniently from Columbus's landfall in the Americas in 1492, though events subsequently followed a similar course in Australasia, parts of Africa, and some Pacific

islands. Settler societies commonly identified indigenous populations by the use of generic terms such as Indians, natives, aborigines, and indigenes. 'Indigenous' has become the preferred term, particularly in Fourth World discourse.

People identifying as indigenous live alongside the descendants of settlers and also recent immigrants in the former colonies of the Americas and Australasia, and most of the governments of these states recognize their presence. Indigenous groups also exist in Arctic Scandinavia and Russia, and in Japan, following the extension of central government into marginal areas, and incursions by the majority populations. These peoples survive either as internal colonies, or integrated with lower strata of the majority society. Some minorities in the recently decolonised nations of Africa and Asia have also identified themselves as indigenous, in resisting the encroachments and intrusions of a dominant population and government. There are other people who might be regarded as indigenous but who have been unable to participate in the Fourth World coalition, or who prefer some other kind of identification.

2. Indigenous Peoples and National Governments

Official criteria of indigenous status vary; thus, some Latin American countries recognize only those who speak an indigenous language, regarding others as assimilated; Australia accepts a combination of self identification and community recognition, while in the US, status derives from treaties made between the government and certain tribes. Taking official estimates from a selection of countries, the indigenous population exceeds 40 percent in the Andean republics and Guatemala; ranges between five percent and 20 percent in New Zealand, Chile, Mexico, Panama, and Nicaragua; between one and four percent in Australia, the United States, Canada, Colombia, Norway, Malaysia, and the Philippines. The percentages are even smaller in Brazil, Sweden, Russia, Japan, and Botswana. However, the size of an indigenous population is no guide to its political visibility on the global arena. From the metropolitan perspective, the position of indigenous peoples is more or less marginal. From their own perspective, keeping the majority at arm's length enables them to preserve an indigenous space, perhaps grounded in a distinctive mode of livelihood, such as fishing or herding or subsistence cultivation. However, the history of indigenous peoples has been characterized by encroachments on this space, and if they can no longer disengage, their lives become hostage to the schemes of the majority. In the second half of the twentieth century indigenous spaces all around the world were subjected to a new round of incursions, in such forms as logging, agribusiness, mining, commercial fishing, and tourism. Resistance to previous incursions had been localized, and often

violently suppressed, but from the 1960s, changes in international relations provided indigenous peoples with new freedoms and opportunities.

3. *International Politics and Transnational Coalitions*

In the context of the Cold War and the replacement of the old imperial domains by nonaligned 'new nations', colonialism and racial discrimination became issues in international politics and reproaches that most governments preferred to avoid. The new medium of television transformed domestic issues such as the African-American civil rights movement in the US and the Bantu struggle against *apartheid* in South Africa, into global dramas, on which people elsewhere could make judgments and relate to their own situations. The eruption of the Cold War in Latin America and Southeast Asia created a climate of dissent in the metropolitan countries involved, as well as others that were not directly involved, for example, in Western Europe.

From the 1970s the world became more receptive to indigenous demands, and indigenous mobilization took on an international dimension, both through coalitions of indigenous peoples across national borders, and the involvement of nonindigenous organizations. Indigenous Americans were the first to mobilize across national borders, with the formation of the International Indian Treaty Council 1974 (Ortiz 1984). In 1975 indigenous Americans formed the World Council of Indigenous Peoples (WCIP), presently opening their ranks to Aboriginal Australians, New Zealand Maori, and Norwegian Saami; others followed. Sympathetic governments facilitated the travel of indigenous spokespersons and several hosted the world conferences that have been held annually ever since. The United Nations recognized the WCIP and its participant organizations in 1977.

As early as the 1950s, the International Labor Organization had drafted a resolution on discrimination against indigenous peoples (Ortiz 1984); but it was not until 1972 that the UN, in response to representations from various nongovernmental bodies, set up the United Nations Sub-Committee on the Prevention of Discrimination and the Protection of Minorities and commissioned a study of the situation of indigenous peoples (Van Cott 1994, Pritchard 1998). In 1977, after further lobbying from nongovernment organizations, the UN convened a conference in Geneva on discrimination against indigenous populations in the Americas, in which the main participants were indigenous. In 1982, it set up the Working Group on Indigenous Populations, bringing to Geneva indigenous peoples not only from the Americas but Australasia and Europe.

Following the Columbus quincentennial, the UN proclaimed an International Year and Decade of

Indigenous Peoples. Since 1995, its Commission on Human Rights has been discussing a draft declaration on the Rights of Indigenous Peoples, to be ratified by the participating governments. These developments at the international level have coincided with the emergence, in Europe as well as North America, of nongovernment organizations specifically concerned with indigenous peoples. The World Council of Churches has also maintained a continuous interest since 1972 and organizations with broader concerns, such as the environmentalists, have discovered interests in common with indigenous groups. As a result of these developments, governments must expect not only to have their own indigenous minorities subject to international scrutiny, but also to take a position on the practices of other governments.

4. *Political Process and Indigenous Mobilization*

Indigenous groups have finally to settle matters with the governments of the countries in which they live, but appealing to international opinion now offers a means of applying a pressure that is not otherwise available. Through the Internet, indigenous organizations are able to disseminate information instantly to sympathisers around the world, and through them to national media and political bodies; where indigenous activists are at risk of reprisals, international observers can be on the scene at short notice. In less critical circumstances, indigenous protesters can take advantage of a national festivity, when the international media are watching, to publicize their grievances through what has become the politics of embarrassment. Otherwise powerless indigenous groups have become skilled in the performance of protest, deploying the signs of their identity in colorful ethnodramas which draw the television cameras (Paine 1982).

The outcome of indigenous mobilization depends on the relative strength of the opposing interests, and—since indigenous peoples live under a variety of regimes—the forms of political process available to them. Even in democracies, a small, low status, or inaccessible minority may have difficulty in attracting the attention of voters. International pressure has induced governments to be more responsive to indigenous claims than their electorates. In some countries, legal process has proved more productive than political process and there has been a substantial development in the law of indigenous rights, with the sharing of experience and insights across national boundaries (Pritchard 1998).

5. *The Politics of Indigenous Difference*

It took some time for a distinctively indigenous politics to take shape. At first indigenous people might mobilize within a framework of agrarian struggle or

civil rights campaigns, with class or race as the basis for alliance. With the emergence of difference politics in the 1970s, the indigenous movement began to articulate its own special needs and aspirations. Thus, a Bolivian Aymara leader argued the need to progress from 'democratic and political reductionism, to an ethnic democracy' (Van Cott 1994).

Despite the diversity among indigenous peoples, certain issues are now regarded as characteristic of all Fourth World peoples, being woven into a transnational discourse of indigeneity. Definitive is the trope of survival, referring to a past of resistance to domination and assimilation, and a future in which indigenous peoples will regain the space, physical and cultural, in which to practice their unique kind of difference. This difference often is grounded in subsistence economic practices that use resources in ways quite different from those of the mainstream economy. Land is not only an economic resource but also the site of the group's mythology and history. Where the land has been expropriated, there is a demand for its return to the traditional owners, and the legislation of some kind of native title. Where recuperation is not a possibility, there may be a call for compensation, to acquire other lands, or for royalties that can be used to underwrite some other basis for indigenous revival.

Indigenous peoples also demand the rehabilitation of their cultural heritage, including language. However, a return to a precolonial existence is rarely an option. Revival of indigenous religion raises particular difficulties, since the groups in question may not only have been practicing Christians for generations, but be indebted to the church for local support and international linkages in its political struggles. Adherents of New Age religions may welcome the return of the old religions, but more common is a negotiation with the churches over the indigenization of clergy and forms of worship. Whether in Christian or other form, the Fourth World proclaims its spirituality in the face of the materialism of developed and developing worlds.

The demand of Fourth World peoples for the restoration of sovereignty constitutes the most direct challenge to the state, just as the claim to be 'first nations' challenges the self image of the nations which encompass them. Many indigenous peoples are too small, too scattered, or too dependent economically, for sovereignty to be realizable. A number have been able to negotiate a degree of self determination within the framework of the state, in respect of a particular territory; the Nunavut agreement in Northern Canada is the most far reaching arrangement of this kind. Peoples who live interspersed with the majority population are harder to accommodate, but the Saami Parliaments in Scandinavia and the Aboriginal and Torres Strait Islander Commission in Australia have assumed responsibility for some of the special needs of their constituents. As with land ownership, it may not be possible or appropriate to try to recreate pre-

existing institutions of government, so that new forms have to be devised. As with the Aboriginal and Torres Strait Islander Commission, power is more centralized than formerly. In the US, by contrast, power continues to lie mainly with the tribes, constituted as a result of treaties drawn up in the nineteenth century.

The Fourth World can be described as an imagined community (Anderson 1991) but it is not one that aspires to nationhood; the aim of coalitions such as the WCIP is to enhance rather than assume the sovereignty of its constituent peoples and to enable them to proclaim their particular kinds of difference to their world. It has, nevertheless, provided the milieu in which a Fourth World discourse could be articulated.

6. The Fourth World and its Constituents

Fourth World discourse represents indigenous peoples as the custodians of a unique wisdom which, if lost, will be lost to all humankind. Cultural survival, then, is not just a matter of minority rights but of global concern. The qualities that make up indigeneity are more or less explicitly alternative to modernity including a spiritual rather than exploitative attitude towards the environment and a focus upon essential human needs over material accumulation and economic development. It is because of these shared qualities that indigenous people are able to talk to one another and imagine people like themselves, even though they have not met them.

Fourth World discourse necessarily presents the qualities of indigeneity as generalities, but this is on the understanding that their value is only realized through their embeddedness in particular local life ways. While it celebrates the special qualities that indigenous peoples share, any tendency to standardize or homogenize them runs up against the countervailing emphasis on the cultural diversity of the constituency. Consistent with this, Fourth World gatherings are celebrations of particularity, in which the various groups perform their cultures in such accessible forms as music and dance, and cuisine. Removed from their sociocultural settings, these practices cannot be apprehended in their full complexity; they are rather the signs of identity, cultural capital which is deployed to achieve a sense of equivalence among Fourth World peoples, and to display difference to others.

While Fourth World leaders speak as members of the particular peoples from which they come, participation in the transnational arena takes them away from their communities for long periods; some may be exiled from their homelands. Thus, paradoxically it is possible to speak of cosmopolitan indigenes, whose time is spent working with other indigenous peoples, and representing the Fourth World to nonindigenous people, leaving little time to keep contact with their grass roots. For the present, however, the Fourth

World is a culture of negotiation and representation; it is not an identity in itself.

7. Scholarly Studies of the Fourth World

The discipline of anthropology emerged in the wake of colonization as a study of the peoples whose descendants make up the Fourth World. The original aim of this research was to reconstruct the early history of humanity. However, from early in the twentieth century anthropologists along with ethnohistorians studied the effects of European settlement and colonization on indigenous peoples and the ways they responded to these influences (Wolf 1982).

Since the 1960s, anthropologists and ethnohistorians have also been involved in the indigenous rights movements, coming forward as experts in court cases, and advising community leaders. Anthropologists tend to work from the perspective of the local community but there have been comparisons of conditions within both North and South America (Kroeber 1994; Urban and Sherzer 1991), and between continents (Dyck 1985). Studies at the level of the nation state have been rather the province of historians, political scientists, and lawyers, some of whom have also worked comparatively, and on the relation between indigenous peoples and the UN (Ortiz 1984, Van Cott 1994, Pritchard 1998). Nongovernmental organizations, such as Cultural Survival, Survival International, and the International Work Group for Indigenous Affairs, are concerned specifically with indigenous peoples and also sponsor and publish research in these areas. What remains unresearched is the actual working of Fourth World conferences and delegations and the development of a meta-indigenous culture among those indigenous people who work at this level.

See also: Colonization and Colonialism, History of; Development, Economics of; Development: Social-anthropological Aspects; Development: Socioeconomic Aspects; Globalization: Geographical Aspects; Globalization: Political Aspects; Industrialization; Inequality

Bibliography

- Anderson B 1991 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, New York
 Dyck N (ed.) 1985 *Indigenous Peoples and the Nation-State: 'Fourth World' Politics in Canada, Australia and Norway*. Institute of Social and Economic Research, Memorial University of Newfoundland, St Johns, Canada
 Kroeber K (ed.) 1994 *American Indian Persistence and Resurgence*. Duke University Press, Durham and London
 Ortiz R D 1984 *Indians of the Americas: Human Rights and Self-Determination*. Zed Books, London

- Paine R 1982 *Dam a River, Damn a People? Saami (Lapp) Livelihood and the Alta/Kautokeino Hydro-electric Project and the Norwegian Parliament*. International Work Group for Indigenous Affairs, Copenhagen, Denmark
 Pritchard S (ed.) 1998 *Indigenous Peoples, the United Nations and Human Rights*. Zed Books, London
 Urban G, Sherzer J (eds.) 1991 *Nation States and Indians in Latin America*, 1st edn. University of Texas Press, Austin, TX
 Van Cott D L (ed.) 1994 *Indigenous Peoples and Democracy in Latin America*. St Martin's Press, New York
 Wolf E R 1982 *Europe and the People without History*. University of California Press, Berkeley, CA

J. Beckett

Free Will and Action

According to the standard account of action, an agent's acts are those of his bodily movements that spring in an appropriate way from his will. Bodily movements that are caused in some other way—the bodily movement of someone who gets blown over by the wind, say—are not among the actions agents perform. More precisely, an agent's actions are those of his bodily movements that occur because he wills them to, where an agent's will is, in turn, identified with his desires and means-end beliefs: actions are bodily movements that occur because an agent desires some outcome and believes that the outcome can be produced by his moving his body in the relevant way. Bodily movements that trace their causal history to other states or events in an agent's psychology, or to nonpsychological states or events, are not actions (Davidson 1963, 1980).

The standard account of action suggests that all animals that are capable of desire and belief, whether those animals are human or nonhuman, are capable of action. The bodily movement of a mouse that occurs because it wants some cheese and believes that it can get some by running in the direction of the crumbs on the floor is an action of the mouse, just as much as the intricate performance of a gymnast or a ballet dancer. By contrast, the movement that the mouse makes when it is picked up by a cat and flung across the room is not an action, any more than the movement of a ballet dancer who is unceremoniously dropped to the floor by her partner.

Not all of an agent's actions, whether the agent is human or nonhuman, are free, however, notwithstanding the fact that they all spring from his will. When a mouse is caught in a trap by its tail, it may give up the struggle and sit down near the trap. Its body remains under its control, not under the control of some alien force, but staying near the trap is not something that the mouse does freely. Similarly, when

a man is locked in a prison cell he may give up the struggle and sit on his chair behind bars. But staying behind bars is not something that he does freely. What, then, within the class of actions, is the mark of those of an agent's actions that are free, as opposed to those that are unfree?

Some hold that free actions are all and only those that are done intentionally: that is, they are those of an agent's actions which, suitably described, can be seen to have been rationalized by his desires and beliefs (Davidson 1973, 1980). This idea is supposedly supported by paradigm cases of unfree action, such as that of the man and the mouse. For though the mouse stays near the trap and the man stays behind bars, neither does what they do intentionally under these descriptions. The man, for example, intentionally sits on a chair, as opposed to on the ground—that is something that is rationalized by his desires and beliefs—and one consequence of his sitting on the chair is that he stays behind bars. But he does not have any desires and beliefs that rationalize his staying behind bars. That is merely a consequence of what he does intentionally, not itself something that he does intentionally.

But though this certainly captures one sense in which we might say of actions that they are under an agent's control, it does not capture the idea of a free action. It entails, for example, that whenever someone who acts in a certain way knowingly, but not intentionally, he does not act freely, and this is surely quite wrong. Suppose that someone, A, desires to convey to another, B, the low opinion he has of a third person, C. But let's suppose further that he goes on to tell B in full knowledge of the fact that C is standing within earshot, and so causes C to become embarrassed. He could have waited until C left, but he didn't because he didn't care. A then knowingly causes C to become embarrassed, but he does not do so intentionally, as there is no desire of A's such that he believes he can satisfy that desire by embarrassing C. C's embarrassment is merely a foreseen consequence of what he does do intentionally, which is to tell B his low opinion of C. It would, however, be extremely implausible to suppose that the fact that C's embarrassment was merely foreseen, rather than intentionally produced, makes it the case that A didn't cause C's embarrassment freely. It was, after all, entirely up to A whether or not C would be embarrassed by what he said. He wasn't forced to embarrass C.

The more standard idea has thus been that an agent's acts are free just in case he could have done otherwise than he in fact did when he acted (Wiggins 1973). The question 'Did the agent act freely?' is in this way supposed to reduce to the question 'Did the agent have any alternative option?' A's causing C to become embarrassed is therefore indeed a free action, on this way of seeing things, because A had the option of acting in some other way instead. He could have waited until C was out of earshot. This is the relevant

sense in which free actions are under an agent's control. The reason that a mouse caught in a trap does not stay in the vicinity of the trap freely is thus that it has no other option: it cannot do otherwise but stay near the trap. The reason a man who sits on a chair in a locked prison cell does not stay behind bars freely is that he, too, has no other option: he cannot do otherwise than stay behind bars.

However, this more standard account of free action immediately raises a difficult question, for without a special explanation of what the 'could' means, it seems to imply that freedom of action is impossible in a deterministic world. Every action performed by an agent in a deterministic world is fixed by the way the world was long before the agent was born together with the operation of causal laws. In a very familiar sense of 'can,' no agent can ever do otherwise than he in fact does. Agents are therefore always forced to do what they do by their past and the forces of nature.

Incompatibilists—those who think that freedom is impossible in a deterministic world—happily accept this conclusion (van Inwagen 1975). But they in turn divide over the substantive question whether the world is or is not deterministic. Those who think that it is—the so-called 'hard' determinists—draw the conclusion that there are no free actions. Those who think that there are free actions—the 'libertarians'—draw the conclusion that the world is not deterministic. But not everyone is an incompatibilist. Compatibilists find it difficult to see how the mere fact of causal determinism is supposed to undermine freedom of action. If anything, they insist, things are the other way around. Events that cannot be explained causally are completely random: this is what it means to say that an event eludes causal explanation. But free actions can hardly be randomly produced. Free actions, like all actions, are produced in a very reliable fashion by the desires and beliefs of their agents. They therefore conclude that free actions, if there are any, are caused by desires and beliefs, and hence that the sense of 'can' at issue must be different from the one that so impresses the incompatibilists (Ayer 1954, 1982).

Compatibilists have therefore coupled the suggestion that an agent's acts are free just in case he could have done otherwise with a special explanation of what this 'could' means. To say of an agent that he could have acted otherwise, in the sense relevant to establish freedom of action, is, they suggest—or, at any rate, this is the way the suggestion was originally put—to say of him that he would have done otherwise if he had chosen to (Moore 1911). In other words, freedom of action turns on whether or not, in the nearby possible worlds in which agents choose to do otherwise than they do in actuality—these are possible worlds in which the past is ever so slightly different—they succeed in doing what they choose to do. The truth of falsity of determinism is irrelevant.

The intuitive idea here should be clear enough, for the suggestion amounts to the claim that agents enjoy freedom of action if the way the external world presents no barrier to their efforts to act as they will: freedom is a matter of the absence of such external barriers or constraints. To see how this allows us to sort cases of free from cases of unfree action, consider the difference between two men who stay in a prison cell, one of whom is locked in, and one of whom isn't. The difference is that the one, but not the other, would have walked out of his prison cell if he had chosen to do so: in the nearby possible worlds in which the agent who is locked in his prison cell chooses to walk out, he fails to do so, as he is still locked in, whereas in the nearby possible worlds in which the agent who is not locked in his prison cell chooses to walk out, he succeeds in doing so, as his cell presents no barrier to him. The one has the option to walk out, whereas the other doesn't. Moreover this account is, apparently, fully general. Of two mice that stay nearby a trap, the one whose tail isn't caught in the trap, and so moves away in the nearby possible worlds in which it chooses to do so, stays nearby the trap freely, whereas the one whose tail is caught in the trap, and so doesn't move away from the trap in the nearby possible worlds in which it chooses to do so, does not stay nearby the trap freely. The external world presents a barrier to the movements of the latter, it serves as a constraint, but it presents no such barrier or constraint to those of the former.

Intuitive though this idea might be, however, the suggestion in the form in which it was originally made is problematic. The problem is that, since choosing is an action, it is something that people can do freely or unfreely (Chisholm 1964, 1982). But now suppose we ask whether a particular action is free. This turns on whether the agent could have done otherwise than he in fact did, and, according to the suggested analysis of 'could,' this turns on what happens in those nearby possible worlds in which the agent chooses to act otherwise. But now note that, whatever answer we give to this question, we have been forced to admit that the status of the original action as free or unfree turns on what happens when another merely possible act is performed, an act whose status as free or unfree in turn needs to be determined. Moreover, the status of this merely possible act turns on whether or not, in those nearby possible worlds in which the agent chooses otherwise than to choose to act otherwise, he succeeds in so choosing. Whatever answer we give to this question, it follows that we have therefore been forced to concede the possibility of yet another act whose status as free or unfree needs to be determined. And so we must go on.

The problem here is the need to posit further acts of choosing to act otherwise, and choosing otherwise than to choose to act otherwise, and so on. The problem is not just that questions about freedom of action are never answered, once and for all—every

question answered leaves another posed but unanswered, and so on *ad infinitum*—the problem is rather that it is implausible to suppose that every creature capable of free action is capable of such ever higher and higher order acts of choosing. In order to raise the mundane question whether a mouse, caught in a trap, sits nearby the trap freely or unfreely, it is surely quite incredible to suppose that we must thereby concede the possibility of its choosing to do otherwise than to sit near the trap, and choosing otherwise than to choose to do otherwise than to sit near the trap, and so on. Mice simply aren't cognitively sophisticated enough to make such choices.

How are we to solve this problem? The suggested analysis of 'could have done otherwise' clearly needs to be rephrased so that the possibility of the freedom of an agent's actions does not entail the possibility of his making ever higher and higher order acts of choice. We should therefore suppose that for it to be true that an agent could have done otherwise it has to be the case that he would have done otherwise, not if he had chosen, but rather *if he had wanted*. This not only removes the problem just encountered—it is indeed possible for agents whose acts are free, men and mice alike, to want other than they actually want, and that possibility does not commit us to the possibility of their making higher order acts of choice—it also better captures the intuitive idea presented at the outset, the idea that an agent's actions are free to the extent that they are dependent on his will. For, as we saw, an agent's will is to be identified with his wants and means-end beliefs. To say of an agent, be he man or mouse, that he acts freely is thus, according to the rephrased suggestion, to say of him that, in the nearby possible worlds in which he wants something different, he acts in that different way. This is what it is for him to have alternative options.

Note two consequences of this account of free action. The first is that certain acts which we would ordinarily suppose to be unfree aren't really unfree. When someone says 'Your money or your life!' and I give him my money, we would ordinarily say that I don't act freely because I have been forced to hand over my money. But, according to the account on offer, I do act freely. I have another option: I refuse in those nearby possible worlds in which I want not to give him my money more than I want him not to kill me. Many are unhappy about this. They think that there must still be a sense in which my freedom is compromised by such a threat. They point out that though, in the nearby possible worlds in which I want to refuse, I do refuse, it isn't the case that wanting to refuse would be at all reasonable, in those circumstances. Indeed, not only would it not be reasonable, it would be completely crazy. There is, they suggest, thus the following sense in which the person who threatens me forces me to act: he puts me in circumstances in which there is no other reasonable option but to comply with his threat. But if this right—that is, if this

is the sense in which threats compromise my freedom—then it seems to follow that people who make me irresistible offers equally compromise my freedom, for they too put me in circumstances in which the only reasonable option is to accept their offer. Whether it is acceptable to suppose that the only sense in which threats compromise my freedom is a sense in which offers also compromise my freedom clearly thus needs much more thought (Frankfurt 1973, 1988).

The second consequence concerns the relationship between free action and moral responsibility. According to this account of free action, freedom is not sufficient for an allocation of moral responsibility. Mice often act freely, but they are not morally responsible for any of their acts. Something else is therefore required. Other more complicated cases tell in favor of the same conclusion. Consider, for example, the plight of a reluctant alcoholic who wants to have a drink, and so has one, knowing he shouldn't. He presumably drinks freely, in the sense provided by the above account of free action, as in the nearby possible worlds in which he wants not to have a drink, he does not drink. Yet it is surely an open question whether or not he is morally responsible for his drinking. If, for example, his desire to drink is literally irresistible, then he is not responsible. Indeed, we would even think that he isn't free, in the important sense of being unable, in the circumstances, to want to do anything else except to drink. This more complicated case thus suggests that in order to understand the conditions of moral responsibility we must look not so much to an account of free action, but rather to an account of freedom of the will. The idea is that, though the alcoholic's actions may be free, his will is not, and that this is therefore the source of our judgements about moral responsibility (Frankfurt 1971, 1988).

Is the performance of a free action a necessary condition of moral responsibility? The answer seems to be that we can and do hold agents responsible for acting when they have no alternative, and that acting freely would thus seem not to be a necessary condition either. Those who act in the false belief that their acts are free are often held morally responsible for what they do, notwithstanding the fact that they don't act freely. Imagine, for example, that when a family visits an old prison they disperse so that they can explore the place individually, but that they agree to meet at a certain time to go home. When the agreed time to leave arrives suppose that one of the children fails to turn up. His parents look for him and find him in a cell. Because he doesn't want to leave, he has hidden from them. But now suppose that the cell door is closed, and that, unbeknown to the child, it is not just closed but also locked. It then follows that the child had no alternative but to be in the prison cell when it came time to leave. He couldn't have met the rest of his family. Is the child off the hook in terms of responsibility? Hardly. His parents would surely quite rightly continue to hold him morally responsible

for failing to turn up. The fact that he couldn't have done otherwise is irrelevant because it doesn't figure, in the right kind of way, in the explanation of the child's failure to meet them. In order to see what is relevant, though, we must once again look at what does figure. This case, too, suggests that we need to look more carefully to an account of freedom of the will (Frankfurt 1969, 1988).

Now that we have seen what it is for an agent's acts to be free, we must ask what it is for his will to be free. It might be thought that we should model our account of freedom of the will on the account we give of freedom of action. Thus, the suggestion might be, just as freedom of action is a matter of being in circumstances in which one can do whatever one wants to do, so freedom of the will is a matter of being in circumstances in which one can want whatever one wants to (Frankfurt 1971, 1988). Where freedom of action is a matter of the absence of external constraints on action, so freedom of the will is a matter of the absence of internal constraints on wanting. An agent's will is free just in case, in those nearby possible worlds in which he wants to want to act in a certain way, he wants to act in that way.

There is clearly something right about this idea. Those who enjoy freedom of the will do seem to have the capacity to control their will, where this includes the capacity to control their first-order desires. The problem with the reluctant alcoholic seems to be that he has no such ability: his desire to drink is beyond his control. What is less clear, however, is that this idea is well captured by the suggestion just made. The desires an agent has about his first-order desires are, after all, just more desires he has, and are hence just another part of his will, a part which may or may not be under the agent's control (Watson 1975, 1982). Thus, even supposing it to be true that in those nearby possible worlds in which an agent wants to want to act otherwise, we can always imagine that the agent is unable to have wanted to want to act otherwise. If the agent were unable to want to want to act otherwise then we must surely ask in what sense he enjoys freedom of the will. Certain of the desires such an agent has in certain possible worlds are controlled by certain others in those worlds, but the agent himself doesn't seem to be in control of any desires at all.

Some theorists have argued that in order to solve this problem we need to postulate a special kind of higher order desire: a desire which captures the agent's own sense of identity (Frankfurt 1987, 1988). When an agent has this sort of higher order desire, they suggest, he identifies himself with the lower order desires that are this higher order desire's objects. Freedom of an agent's will is, the suggestion goes, a matter of his being able to have desires that match this sort of higher order desire: in other words, it is a matter of having desires with which he identifies.

The difficulty with this suggestion, however, is to come up with a suitably independent account of

identification. There certainly is a sense in which an agent who acts of his own free will identifies with the desires on which he acts, as he is himself the originating source of such desires. But this is not an independent sense of identification. It is a sense that we define in terms of the idea of desires originating in a free will. There is also at least the following independently characterizable sense of identification. Some people who take drugs against their will, and who struggle to overcome their addiction, ultimately give in, and so come to identify with their addiction in the sense of building all of their expectations around it. In other words, they come to see themselves as hopeless drug addicts: this becomes their self-conception. But this is hardly a sense of identification that will help us understand the idea of desires originating in a free will.

There is something right about connecting freedom of the will with the possession of desires with which an agent identifies, however. Since agents are the originating source of such desires it does seem that, when they act of their own free will they express their true nature. The desires on which they act are, in some sense, essential to who they are. But though this seems right, it in fact suggests a quite different account of free will, one which does not depend crucially on our having an independent idea of what it is for an agent to have desires with which he identifies.

Since the Enlightenment we have come to conceive of people as being essentially truth-oriented reasoners. They are creatures who have the capacity to recognize the truth and to respond appropriately to it. Insofar as we are deliberators, what this suggests is that we are essentially creatures who have the capacity to recognize and respond to the facts about what we have reason to do. Thus, according to this account, an agent whose will is free is someone who asks himself what there is reason to do and, in answering this question, he is sensitive to the available evidence. This is not to say that he will inevitably acquire true beliefs about what there is reason to do, but it is to say that when he fails to do so, and when the evidence was available to him, his failure will be traceable to a failure to exercise his belief-forming capacities appropriately: he could and should have formed the correct belief, but he failed to do so. Moreover, an agent whose will is free is someone who, in the formation of desires, is sensitive to the beliefs he has about what he has reason to do. Again, this is not to say that he will inevitably form desires that accord with his beliefs about his reasons, but it is to say that when he fails to do so, his failure will be traceable to a failure to properly exercise self-control: he could and should have formed the correct desire, but he failed to do so (Wolf 1990).

As is perhaps clear, this conception of free will differs crucially from the one suggested earlier. What is important for freedom is indeed a capacity to desire otherwise, but the possession of this capacity is not to be analyzed in terms of what happens in nearby possible worlds in which the agent wants to want

otherwise. Instead it is to be analyzed in the same way, whatever that way is, that we analyze the ability rational thinkers possess to respond appropriately to evidence in the formation of their beliefs. Just as a rational thinker who fails to adjust his beliefs to the available evidence makes a mistake he shouldn't have made, and so, we suppose, must have had the capacity to make the adjustment and believe rightly instead, so a rational deliberator who fails to form beliefs about what he has reason to do in accordance with the available evidence, or who fails to desire in accordance with his beliefs about what he has reason to do, makes a mistake he shouldn't have made, and so is supposed to have had the capacity to control himself and acquire the correct belief or desire instead (Pettit and Smith 1996).

Note two consequences of this account of freedom of the will. The first is that it looks tailor-made to provide an account of the conditions of moral responsibility. Those who have no capacity to believe that they have reason to do anything other than they in fact believe they have reason to do, or to desire to do anything other than they in fact desire to do, can hardly be praised for having the right beliefs and desires, if it so happens that they're right, or blamed for having the wrong ones, if it so happens that they're wrong. This explains why we don't hold the alcoholic responsible. The irresistible nature of his desire to drink makes him incapable of an exercise of self-control, despite the fact that he believes, correctly, that he has reason to abstain. It also explains why we don't hold mice morally responsible for what they do. Their lack of cognitive sophistication deprives them of the capacity even to form beliefs about their reasons for doing what they do, let alone the capacity to desire accordingly.

Moreover, nothing other than the capacity to form the right beliefs and desires would seem to be required for people to be responsible either. This is why we hold the child described earlier morally responsible for failing to meet his parents at the appropriate time, notwithstanding the fact that he was locked in a prison cell and so couldn't have met them. We hold him responsible because we assume that his behavior was in fact the product of a desire he had which accorded with his beliefs about what he had reason to do, a belief that we take to be culpably mistaken. It is this that attracts our censure.

The second consequence concerns the possibility of freedom of the will in a deterministic world. To my knowledge, no one yet has a fully worked out theory of what it is for someone to have a rational capacity. It therefore follows that, to this extent at least, a proof that freedom of the will is possible within a deterministic world is yet to be provided. Provision of such a proof would, after all, require a demonstration that individuals in a deterministic world have the where-withal to possess just such a rational capacity. On the face of it, however, it must be said that there doesn't

look to be any particular difficulty in principle involved in providing such a demonstration. We regularly discriminate between people who do and people who don't have rational capacities, and, in so doing, we seem to pay no regard whatsoever to issues of determinism. This is because even if determinism is true, that only shows that people will make the mistakes that they make. It doesn't show that they couldn't help making those mistakes, in the sense of having no rational capacity to get things right.

But, whatever we ultimately come to think about the possibility of rational capacities in a deterministic world, the crucial lesson of this account of freedom of the will is that, if there do turn out to be such difficulties, then it isn't just the idea of freedom of the will that becomes problematic, it is the very idea of a rational agent as someone who has the capacity to recognize and respond appropriately to the truth. Whether freedom of the will is possible in a deterministic world or not, its possibility goes hand in hand with the possibility of agents who have the capacity to form beliefs and desires in a rational manner.

Bibliography

- Ayer A J 1954/1982 Freedom and necessity. In: Watson G (ed.) *Free Will*. Oxford University Press, Oxford, UK
- Chisholm R M 1964/1982 Human freedom and the self. In: Watson G (ed.) *Free Will*. Oxford University Press, Oxford, UK
- Davidson D 1963/1980 Actions, reasons and causes. In: Davidson D (ed.) *Essays on Actions and Events*. Clarendon Press, Oxford, UK
- Davidson D 1973/1980 Freedom to act. In: Davidson D (ed.) *Essays on Actions and Events*. Clarendon Press, Oxford, UK
- Davidson D (ed.) 1980 *Essays on Actions and Events*. Clarendon Press, Oxford, UK
- Frankfurt H G 1969/1988 Alternate possibilities and moral responsibility. In: Frankfurt H G (ed.) *The Importance of What We Care About: Philosophical Essays*. Cambridge University Press, Cambridge, UK
- Frankfurt H G 1971/1988 Freedom of the will and the concept of a person. In: Frankfurt H G (ed.) *The Importance of What We Care About: Philosophical Essays*. Cambridge University Press, Cambridge, UK
- Frankfurt H G 1973/1988 Coercion and moral responsibility. In: Frankfurt H G (ed.) *The Importance of What We Care About: Philosophical Essays*. Cambridge University Press, Cambridge, UK
- Frankfurt H G 1987/1988 Identification and wholeheartedness. In: Frankfurt H G (ed.) *The Importance of What We Care About: Philosophical Essays*. Cambridge University Press, Cambridge, UK
- Honderich T (ed.) 1973 *Essays on Freedom of Action*. Routledge & Kegan Paul, Boston, MA
- Moore G E 1912/1944 *Ethics*. Oxford University Press, Oxford
- Pettit P, Smith S 1996 Freedom in belief and desire. *Journal of Philosophy* 93(9): 429–49
- van Inwagen P 1975 The incompatibility of free will and determinism. In: Watson G (ed.) *Essays on Actions and Events*. Clarendon Press, Oxford, UK

Watson G (ed.) 1982 *Free Will*. Oxford University Press, Oxford, UK

Watson G 1975/1982 Free agency. In: Watson G (ed.) *Free Will*. Oxford University Press, Oxford, UK

Wiggins D 1973 Towards a reasonable libertarianism. In: Honderich T (ed.) *Essays on Freedom of Action*. Routledge & Kegan Paul, London

Wolf S 1990 *Freedom within Reason*. Oxford University Press, New York

M. A. Smith

Freedom/Liberty: Impact on the Social Sciences

The concept of freedom, or in another rendering in the English language: liberty, is one of the most basic concepts in social and political thought. In many respects, it is the basis of all thought about politics and society. The assumption, in some form, of free will in human beings, namely, raises the question of action as making a difference in the world, and the relation of that difference to the course of the world had no action occurred (see *Free Will and Action*). More particularly, the term political freedom describes the possibility of a multitude of human beings deliberating about what they want in common and acting in common (see *Freedom: Political*).

The tension between the freedom of the singular human being and political freedom is at the center of all modern political thought, that is, all political thought that assumes the human ability of giving themselves their own laws. (Sometimes, this difference finds expression in the terminological choice between liberty, more consistently used for individual freedom, and freedom, for political freedom. Most authors, however, do not make any distinction between these two terms, and none will be made here.) It is also in the background of the theoretical *problématique* of agency and structure in the social sciences.

Comparative research on the history of the idea of freedom is but little advanced, partly due to prevailing preconceptions about the Western origins of freedom, partly due to the intrinsic difficulties of historical comparisons of the most basic linguistic concepts. Refraining from any explicit comparison, however, it is possible to state that the 'Western' tradition of political thought is marked by the confluence of both ideas, individual and political freedom (Patterson 1991). The latter originated in ancient Greece with the idea of self-government of the city states, and led into the Republican tradition, first of the Roman Republic, and then refounded by the Italian city republics, emerging from under the imperium, in the course of

what then became called the Renaissance (Skinner 1978). The former had important roots in early Christianity, as a religion that established more direct links between the singular human beings and their conscience and God, not least under conditions in which the declaration of Christian belief could lead to persecution by the political rulers.

The tension between the two concepts of individual freedom and political freedom is widely acknowledged in political philosophy (see, for instance, Berlin 1969, Miller 1991). This is not to say, however, that a resolution of that tension, or a reconciliation between the concerns that stand behind the two ideas, is not possible. Many authors argue, for instance, that republicanism is a political theory that succeeds in combining both exigencies (see *Republicanism: Impact on Social Thought*), or that romanticism is a social philosophy that underpins self-realization while also developing an idea of a viable polity (see *Romanticism: Impact on Social Thought*). Historically, though, it has been the case that a political theory emerged as central for the self-understanding of the European nation-states as well as for the USA that was based on the concept of individual liberty. Increasingly so during the nineteenth and twentieth centuries, republican views of political freedom were marginalized, and romanticist views of the embeddedness of the singular human beings were assigned a merely supportive, even though not insignificant, role in sustaining polities.

Recent research in the history of political concepts has emphasized that this shift in balances between various political languages took place between the middle of the eighteenth and the middle of the nineteenth century. This period is at the same time the period of the emergence of the early social sciences. It is also the period of the 'democratic revolutions' (R. R. Palmer), of the American and the French Revolution, which are often seen as aiming at giving institutional expression to the principle of liberty. It is from this historical constellation that an assessment of the impact of the concept of liberty on the social sciences has to start out.

1. Liberty and the Moment of the Revolutions

The pervasiveness of political issues, and in particular of the issue of liberty, in the early social sciences, that is, in authors like Smith, Condorcet, or Hegel, is evident. In political terms their modes of reasoning were, however, marked by a curious paradox. These authors were working in the shadow of liberal thinking (in a very broad sense), namely accepting the right of human beings to self-determine their individual and social lives, even though not always without important qualifications (see *Liberalism: Impact on Social Science*). At the same time, though, they were holding that the free action of human beings would not actually

generate, as one might innocently assume, a very wide range of possible individual and social forms and orientations of behavior. Given complete autonomy, human beings would reveal themselves not to be free in a radical sense, but driven by a limited number of well intelligible inclinations. One might even infer that it was exactly this predictability and order that gave grounds and reasons—to these authors themselves and to more hesitant others—not to reject the normative project of liberty. And this linkage of freedom and predictability became particularly important at the historical moment when the externally imposed barriers to free deliberation were threatened to be removed, the moment of the American and French Revolutions.

These revolutions gave institutional expression to the political aspect of a broader culture of individual autonomy that is a key element of modernity. In this sense, much of this era can be seen as a liberation of human beings from imposed ties, but this liberation was far from unproblematic. As Lefort (1986, pp. 214–15) once described this feature of modernity: 'When he is defined as independent, the individual does not exchange ... one certainty against another one. ... The new mode of existence of the individual within the horizons of democracy does not merely emerge as the promise to control one's own destiny, but also and not less as the dispossession of the assurance as to one's identity—of the assurance which once appeared to be provided by one's place, by one's social condition, or by the possibility of attaching oneself to a legitimate authority.' Liberation is here interpreted as an increase of contingency and uncertainty in the lives of human beings.

If this view were unequivocally valid, one should expect a philosophy of contingency—in Rorty's style (1989), for instance—linked to a liberal-individualist political theory to dominate the intellectual scene forever after the successful revolutions. However, historically this was not at all the case. In contrast, 'the historical moment, about which we speak, emerges in such a way that the real rising of the political instance entails its theoretical abatement' (Manent 1994, p. 123). Instead, the historical moment of liberty coincided with the rise of the social sciences. 'Society' as the object of the social sciences has been a 'post-revolutionary discovery'; or, to put it even more succinctly, 'the sociological point of view constitutes itself in the moment when the notion of liberty becomes the principal articulation of the human world' (Manent 1994, p. 75, 113, Therborn 1976, for an earlier analysis of the emergence of sociology 'between two revolutions'). Such apparent paradox reveals the aporia of political thought after liberation. Very generally speaking, the social sciences are exactly a part of the response human beings gave to their new condition of—self-incurred, one might say—contingency and principled uncertainty. Being unable to rely any longer on externally defined certainties,

political thinkers started searching for regularities and continuities which exist without being commanded. The social sciences have been a means to decrease contingency.

2. Concepts of Liberty before the Revolutions

Earlier political thought had already recognized the inevitable circularity of a reasoning in which unity should rise from diversity as well as, at the same time, impose itself on diversity. At that time, however, there was room to try to approach the issue by mere conceptual construction and theoretical determination. As long as there were barriers to the practice of liberty, every conceivable view on its outcomes and consequences for the polity could be held. The mode of reasoning was bound to change when those barriers were removed and experiences were made and when practical issues, such as the founding of political institutions, had to be tackled.

Political philosophy, in its prerevolutionary versions from Machiavelli and Hobbes onwards, gradually had developed an understanding of that which people have to regulate in common on the basis of the assumption that they do this on their own, without any external demands. The line of thinking that leads from Hobbes to Locke and Rousseau is often seen as the source of Western liberalism. It is most often presented as the theory of the social contract in which free individuals relinquish some of their rights in exchange for security and the regulation of the common realm provided by the state. In recent years, however, it has become common to distinguish the Hobbesian from the Machiavellian tradition. In this view, liberals define liberty as non-interference. The Hobbes–Lockean state, founded by free contract, dominates over the individuals, but it interferes with their liberties only to the degree in which the maintenance of order makes this necessary. The liberal tradition needs to draw a strong boundary between the public and the private, and since non-interference of the public into the private is the supreme principle, it can only have a ‘thin’ concept of membership in a polity. Significantly, the concepts of the state and of membership can be developed in the abstract, since they are based on a reasoning each rational individual is seen to arrive at, regardless of the situation.

In contrast, republicans define liberty as nondomination. Drawing via Machiavelli on Roman political thought, nondomination is conceptualized in stronger terms than noninterference; it requires security against interference. Such security stems at least in part from the ways in which citizens relate to each other, so that there is a less sharp divide between the private and the public and a ‘thicker’ concept of membership than in liberalism (see Pettit 1997, Skinner 1998 for recent accounts). As a consequence, the shape and substance

of each republican polity is specific, since it depends on its members and their deliberations. Most contributors to this recent debate emphasize the distinction between liberalism and republicanism to underline that individualist liberalism was not the only alternative available to provide the discursive underpinning of political modernity at its onset. However, they also agree that republicanism was by and large abandoned around the turn of the eighteenth century, and liberalism, though it neither appeared particularly powerful nor coherent before, very soon emerged as the pivotal political theory in postrevolutionary polities.

3. The Impact of the Revolutions on the Concept of Liberty

At a closer look, it is not too difficult to understand such intellectual shift through a look at the deep shock the revolutions meant to political thinking (Wagner 1998). The revolutions were certainly inspired by republican thinking. Looking at them in the somewhat stylized way that an emphasis on the French experience has long suggested, one recognizes, however, that they aimed at combining two objectives that proved to be practically impossible to hold together. On the one hand, they aimed at transforming state sovereignty in the hands of the monarch into popular sovereignty, that is, they worked with extended notions of citizenship and liberty. On the other hand, they held such a transformation of the polity to be conceivable only in the form of the existing territorial state and within its dimensions.

Such double transformation entailed, first, that the existing relations between the people, tainted with the suspicion of domination and privilege characteristic of feudal society, had to be weakened or abolished. Thus, however, a major available resource for a substantive, ‘thick’ grounding of a modern republic was rejected. Second, the idea of extending political rights widely cast doubts about the viability of a demanding concept of liberty such as the one upheld in the republican tradition. Caution seemed to demand, not least for some more conservative observers, to limit the substance of the concept at the moment at which its reach was extended. As a consequence, the public realm, the state, was robbed of most of its substance and the formal process through which common deliberations were reached was emphasized instead. About such processes, though, there is much less to reason in politico-philosophical terms. The liberal conception that things regulate themselves as soon as atomistic individuals are endowed with the right and the capacity to follow their own interests and passions makes political thought almost superfluous. It suggests that nothing beyond some basic assumptions and rules as to how to set up such processes is needed (though

such rules are much less innocuous than it may have appeared to some of the revolutionaries).

This change is most visible in the context of the foundation of the United States of America, at least in a reading of it, inspired by Gordon Wood, that tries to go beyond the debate over the republican or individualist beginnings of the American polity. In their reaction against obtrusive and illegitimate government, the Americans consciously tried to build institutions so that 'power may not actualize itself anywhere.' The 'disembodiment of government ... goes along with a disentangling of power, of the law, and of the knowledge of the ultimate ends of a society' (Lefort 1991, p. 27). Many of the most important themes of earlier political philosophy were simply ejected as illegitimate from reasoning about political institutions. 'For the first time, the place of power is easily recognized as being empty' (Lefort 1991, p. 27). The Americans of the revolutionary generation had 'broken through the conceptions of political theory that had imprisoned men's minds for centuries' through the renunciation of any substantive foundation of the polity and 'a total grounding of government in self-interest and consent' (Wood 1998, p. 614, 612).

Even in the American context, this deliberate political construction could not exorcise all concerns about the substantive foundations of the public realm. Beyond the calls for republican virtue during the early period, the search for a preordained unity emerged in the nineteenth century, pursued often by Americans with some German intellectual background (Gunnell 1991). However, such political science could not establish itself as a dominant paradigm and was mostly abandoned after the end of the nineteenth century. In Europe with its long-established statist institutions, the idea of unity was not abandoned, but it was transformed rather radically. As in the US, the observation seemed undeniable that, once the reasonable will of the human beings had been cast into institutions, the political order must be seen as intrinsically satisfactory (Manent 1994, pp. 228–9). In this organized context, however, the fear that such new order based on human liberty might turn out unsustainable was much greater than in North America. 'The effect of liberty to individuals is, that they may do what they please: We ought to see what it will please them to do, before we risque congratulations,' as Burke (1993 [1790], pp. 8–9) put it in his reflections on the French Revolution.

4. Liberty and the Rise of the Social Sciences

In this situation, individualist liberalism provides the less demanding concept of ('negative') liberty as noninterference that may live up to Burke's requirements under the proviso that the state is capable of maintaining order for and above the individuals.

However, there is another way out of the aporia of liberty, namely the attempt to arrive at knowing 'what it will please them to do' by other means. In Europe, indeed, the idea of an emptiness of the political space did not fully emerge. In the Enlightenment combination of freedom and reason, the state was rather reinterpreted as the incarnation of reason once the expression of human freedom fed into its construction. This is one of the central themes of Hegel's writings, but it can be found in many other contemporary works as well (Wuthnow 1989, p. 315). This linkage of knowledge and politics, the idea of 'legislative reason' (Bauman 1987), was characteristic of much of the early social science thinking during the revolutionary period in Europe. Reason, rather than merely serving as an ideal point of philosophical orientation (which it also continued to be), was to be made practical in the elaboration and application of the rules of politics. The movement to transform political reason into a science of politics found its most ardent spokespersons in France where politics had most radically been put onto new foundations. Beyond the voluntary commitment to construct a new social order, there was also a feeling of dire necessity to do so on well-informed grounds, not least motivated by the shock of revolutionary terror. The writings of Saint-Simon and Comte mark one major line of such thinking.

The American and French revolutions, thus, strongly suggested to study what held human beings together, how they would actually organize their lives—individually, in 'associations' or 'social movements,' and in the polity and the 'nation'—and what kinds of regularities and orders could be expected, if people were permitted to do so on their own, without imposed restrictions. This is the search for social ties that is one major root of the social sciences, and it is in this sense a politically motivated search.

The hope and aspiration was that the moral and political sciences should and could now achieve 'the same certainty' as the physical sciences (Baker 1975, p. 197). Certainty was a requirement of some urgency, since the new political order needed assurances of its viability. But it was also regarded as an historically new possibility, since political action was liberated of the arbitrariness of decisions by rulers of doubtful legitimacy and given into the hands of the multitude of reason-endowed human beings. The 'blend of liberalism and rationalism' (Baker 1975, p. 385, about Condorcet) can thus be explained as stemming from the same source, the Enlightenment linkage of freedom and reason—though possibly in its most optimistic version.

This linkage, if it could be sustained, made a science possible, where according to the classical view it would have been unthinkable, in the realm of the political. Condorcet subscribed to the new concept of free expression of political wills, but for him there was no contradiction between that view and the possibility of scientific analysis. 'Societal choice' could well be

transformed 'into the rational decision-making of the idealized republic of science' (Baker 1975, p. IX, 193).

5. Varieties of Social Science Responses to Liberty

The rights-endowed individual became thus, in such views, the only conceivable ontological as well as the methodological foundation of a science of political matters after the revolutions. Once the rights of man had been generally accepted as self-evident and unalienable, it seemed obvious, to Turgot and Condorcet for instance, that they were also 'the logical foundation of the science of society' (Baker 1975, p. 218). In rights-based liberalism, the individual is the only category that need not, often in fact: cannot, be debated. The individual is simply there, whereas everything else—for instance, what criteria of justification are to be applied when determining the collective good—needs to be argued about. Substantive aspects of human interaction are subject to communication and consensus. And, to make the issue even more complicated, with whom one should enter into communication, that is, the boundaries of the community, is itself not given, but subject to agreement.

Once this assumption was accepted, basically two avenues of constructing a science of the political had opened. Both these forms of social theorizing connect modernist political philosophy, that is, individualist liberalism, to a science of the political. One possibility was to try to identify by theoretical reasoning the basic features of this unit of analysis, the individual human being, and its actions. Since this unit was conceived as an ontological starting-point, devoid of all specific, historical and concrete, ties to the world, its characterization was to proceed from some inherent features. From earlier debates, those features had often been conceived as twofold, as passions and as interests. In the late Enlightenment context, the rational side of this dichotomy was regarded as the one amenable to systematic reasoning. It thus allowed the building of a scientific approach to the study of at least one aspect of human interaction with the world, namely the production and distribution of material wealth. This approach inaugurated the tradition of political economy, later to be transformed into neoclassical economics and, still later, into rational choice theory. The moral and political philosophy of the so-called early modern period split into a political theory based on the idea of the social contract and a rationalized moral theory based on the idea of exchange. In both cases, the individual is the starting-point and unit of the analysis.

While political economy was based on a highly abstract, but for the same reasons extremely powerful, assumption of human rationality, the other conclusion

from the individualist foundational principle was arguably even more reductionist but much more cautious. Avoiding any substantive assumptions on the driving forces in human beings at all, the statistical approach, often under the label of political arithmetic, resorted to the collection of numerically treatable information about human behavior. The space of substantive presuppositions was radically emptied in this thinking, but the methodological confidence in mathematics increased in inverse proportion (Brian 1994, Desrosières 1993).

Thus, two strands of political thought that had been proposed and elaborated for some time rose to new and greater prominence, political economy and political arithmetic. The denominations these approaches were known by in the late eighteenth century referred explicitly to political matters. Both were to lose these attributes in the nineteenth century when they had consolidated their ways of proceeding and when the application of these cognitive forms had established predominance over political deliberation in decision on common matters, at least in the view of many economists and statisticians. Mostly, this terminological change has been interpreted as an autonomization of cognitive approaches and as a differentiation of the sciences into disciplines. However, it is not exactly appropriate to say that economics and statistics separated from politics. Once the approaches of the former two are accepted as valid, there is nothing political left to study.

The acceptance of the economic and the statistic ways of conceiving of the social world did not go without criticism; and they were never accepted as the only possible ways anywhere. However, the critiques and alternatives that were proposed most often accepted the fundamental change in political reasoning after the construction of a polity based on the assumption of free individuals. (Since nineteenth-century polities all worked with restrictions that are incompatible with a full-fledged individualist liberalism, one should say more precisely that such a polity was put on the horizon of political debate through the revolutions, rather than made actual in institutional form.) After such a polity had begun to come into existence, new problems were identified. These were now essentially liberal problems; they resulted, one might say, from the observation that not everything that was needed for organizing a liberal polity could indeed be derived from an 'original position' (John Rawls). Two main types of problems may be distinguished by reference to the hypothetical original position in which individuals meet under a veil of ignorance.

On the one hand, the range of conclusions that could be drawn from the assumption of free and equal individuals was too limited. These individuals' relations were structured by the existence of politically important 'prepolitical' social facts, of orientations and links between human beings that were seen to

already exist before individuals entered into political communication and deliberation. On the other hand, the working of the liberal rules would themselves produce new kinds of social relations, 'postpolitical' relations, which would have a structuring impact on the polity in turn.

Attempts to theorize prepolitical relations all start out from the critical observation that the human being who enters into political relations is not such a kind of individual as liberal political theory described it, and that the hypothesis of any original position would lead to serious flaws in the conclusions. This thinking emphasizes the rootedness of the singular human beings in contexts from which their ways of giving meaning to the world stem. The broadest intellectual movement of this kind has been the cultural-linguistic theory of the boundaries of the polity, which inaugurated culturalist thinking in the social sciences and also became one source of later nationalism.

The other main line of 'postrevolutionary' social thought started out from the insights, first, that the basic liberal assumptions, once they were cast into effective rules, would have durable and important effects on what social scientists would soon call the 'structure' of social relations. In this sense we can refer to those relations as being conceived as 'postpolitical.' And second, the question of such relations was forced onto the agenda of social and political thought by the fact that the liberal assumptions on their own did not suffice to create and justify a political order. The observation of structures of representation served the interest of enhancing some stability and certainty in political procedures that could appear to be opened to all contingencies by the abolition of any legitimacy of preordained orders (see *Representation: History of the Problem*). We find here again the two main strategies for rediscovering certainties, systematic observation and reflective conceptualization. These two intellectual responses to the political *problématique* inaugurate two further modes of social theorizing, the behaviorist and the structural-functional one. Unlike economics and statistics, they do not make individualistic assumptions but aim at grounding sociopolitical life in purely social forms.

After the claim for a social science had been made by authors like Saint-Simon or Comte, the themes of the emerging social sciences formed around the observation of this internal structuration of societies. In Germany, Lorenz von Stein reported about the 'social movements' in France that announced a major change in the social order. At mid-century, Robert von Mohl diagnosed a transitory situation after the *Ancien Régime* had disassembled and no liberal order had been able to assert itself in the German states, while beginning industrialization and urbanization placed new demands on the political orders (Maier 1980, Wagner 2000). Karl Marx gave a central place in his social theory (and philosophy of history) to the newly forming social phenomenon, the 'working class.' And

Émile Durkheim provided a representation of society in which the elements of the social order were defined according to their position in the division of social labor and their relations regarded as interlocking in the form of 'organic solidarity.' Steps like these marked the construction of a sociological view-point which was clearly identifiable around the turn of the twentieth century.

6. Liberty after the Social Sciences

All these approaches to the social sciences can, under one aspect, be regarded as ways of dealing with the problem of contingency after the assertion of human freedom. It is on those grounds that sciences of the social and the political are proposed. If, in contrast, politics is seen as that human activity that by its nature is open, plural, and diverse (Arendt 1958), then any cognitive linkage of free action and predictable outcome is inconceivable. Orderly outcomes can only result from planned or routine activities, work, and labor in Arendt's terminology, over which certainty can be established before they are started. In contrast, political action in a context of liberty must go along with contingency of outcomes. From an Arendtian viewpoint, thus, those early social scientists established an impossible connection. If they were heading for a 'social science' trying to identify laws and regularities of human action and societal development, then they necessarily abandoned the heritage of political philosophy, the emphasis on creative agency, irreducible diversity, and the permanent possibility of unpredictable beginnings. It is in the light of such considerations that the closing decades of the twentieth century have witnessed a revival of political philosophies of freedom, often going beyond concepts of liberty as held in liberal political theory. These works, by authors such as Claude Lefort, Pierre Manent or, more historically oriented, Quentin Skinner, are not merely contributions to political philosophy or its history. Rather, they challenge the very separation of the social sciences from political philosophy.

See also: Citizenship: Political; Civil Liberties and Human Rights; Civil Rights; Civil Society, Concept and History of; Freedom: Political; Human Rights, History of; Methodological Individualism: Philosophical Aspects; Political Economy, History of; Rational Choice Theory: Cultural Concerns; Revolutions, History of

Bibliography

- Arendt H 1958 *The Human Condition*. The University of Chicago Press, Chicago
- Baker K M 1975 *Condorcet: From Natural Philosophy to Social Mathematics*. The University of Chicago Press, Chicago

- Bauman Z 1987 *Legislators and Interpreters*. Polity, Cambridge, MA
- Berlin I 1969 Two concepts of liberty. In: Berlin I (ed.) *Four Essays on Liberty*. Oxford University Press, Oxford, UK
- Brian E 1994 *La mesure de l'Etat. Administrateurs et géomètres au XVIIIe siècle*. Michel, Paris
- Burke E 1993 [1790] *Reflections on the Revolution in France*. Oxford University Press, Oxford, UK
- Desrosières A 1993 *La politique des grands nombres. Histoire de la raison statistique*. La découverte, Paris
- Gunnell J G 1991 In search of the state: Political science as an emerging discipline in the US. In: Wagner P, Wittrock B, Whitley R (eds.) *Discourses on Society. The Shaping of the Social Science Disciplines*. Kluwer, Dordrecht, The Netherlands, pp. 123–61
- Lefort C 1986 *Réversibilité: Liberté politique et liberté de l'individu. Essais sur le politique. XIXe–XXe siècles*. Seuil, Paris, pp. 197–216
- Lefort C 1991 Introduction. In: Wood G S (ed.) *La création de la république américaine 1776–1787*. Belin, Paris
- Maier H 1980 [1965] *Die ältere deutsche Staats- und Verwaltungslehre*. Beck, Munich
- Manent P 1994 *La cité de l'homme*. Fayard, Paris
- Miller D (ed.) 1991 *Liberty*. Oxford University Press, Oxford, UK
- Patterson O 1991 *Freedom in the Making of Western Culture*. Tauris, London
- Pettit P 1997 *Republicanism. A Theory of Freedom and Government*. Clarendon, Oxford, UK
- Rorty R 1989 *Contingency, Irony, Solidarity*. Cambridge University Press, Cambridge, UK
- Skinner Q 1978 *The Foundations of Modern Political Thought, Vol. 1, The Renaissance*. Cambridge University Press, Cambridge, UK
- Skinner Q 1998 *Liberty before Liberalism*. Cambridge University Press, Cambridge, UK
- Therborn G 1976 *Science, Class and Society. On the Formation of Sociology and Historical Materialism*. New Left Books, London
- Wagner P 1998 Certainty and order, liberty and contingency; the birth of social science as empirical political philosophy. In: Heilbron J, Magnusson L, Wittrock B (eds.) *The Rise of the Social Sciences and the Formation of Modernity*. Kluwer, Dordrecht, The Netherlands, pp. 241–63
- Wagner P 2000 An entirely new object of consciousness, of volition, of thought. The coming into being and (almost) passing away of 'society' as an object of sociology. In: Daston L (ed.) *Biographies of Scientific Objects*. The University of Chicago Press, Chicago, pp. 132–57
- Wood G S 1998 [1969] *The Creation of the American Republic 1776–1787*. University of North Carolina Press, Durham, NC
- Wuthnow R 1989 *Communities of Discourse. Ideology and Social Structure in the Reformation, the Enlightenment, and European Socialism*. Harvard University Press, Cambridge, MA

P. Wagner

Freedom of the Press

No press system is absolutely free from various social, political, economic, or legal controls of its society.

Most discussions of freedom of the press center on the relationship of government and the press, and here the basic question is not whether government controls the press, but how and to what extent. Freedom of the press refers to negative freedom—freedom from governmental constraint, or it may refer to positive freedom—the capacity to do or achieve certain ends. Most issues concerning press freedom have revolved around negative freedom, understood as the independence of the news media from governmental restraint.

1. The Term 'Press'

The 'press' in 'freedom of the press' is elusive. The general public considers the word press as the conventional institutional news media including newspapers, magazines, and radio and television broadcasters. But the popular image of the term press is too limiting to be accepted in US constitutional law. The press goes beyond mainstream newspapers and magazines to 'every sort of publication which affords a vehicle of information and opinion' (Lovell v. Griffin 1938).

Can the term press be distinguished from 'speech' in 'freedom of speech?' Thus far, the answer to this press–speech dichotomy question is 'No' in US law. The now well-established US rejection of the distinction between press and nonpress communicators presents a striking contrast to the clear-cut statutory recognition of the press in continental Europe and some Asian countries. It should be emphasized, however, that the presence (or absence) of a direct press law does not necessarily ensure a privileged (or unprivileged) press in a country: 'The theory that a special press law in itself constitutes a threat to press freedom has merit only if a special press regime is identified with an authoritarian press regime' (Lahav 1985).

2. Theories of Press Freedom

The key values sought by society in protecting freedom of expression are that freedom of expression assures individual self-fulfillment, helps to discover the truth, secures participation by citizens in social and political and economic decision-making, and promotes a stable change in society (Emerson 1966). The press is protected as the surrogate of the public when performing its informing and checking functions as a Fourth Branch, supplementing the executive, legislative, and judicial branches of government.

The foremost value of press freedom comes from the premise that a press free from government control is essential to a political system based on the will of the people. The 'checking value' of a free press posits that the press should be guaranteed freedom to report

abuses of power by public officials. This compels a logical conclusion that the government should not prohibit the people from receiving news critical of the government. *New York Times* columnist Anthony Lewis succinctly discusses the relationship between the news media's checking function and the media's quest for information:

If big government is to be effectively criticized and controlled, it will take more than the exhortations of an orator. It will take the countervailing force of big newspapers and broadcast networks and public-interest groups and lobbying organizations of all kinds. And they cannot succeed without information. In our society information is power. Officials struggle to control it and, in that struggle, the *citizen-critic* needs constitutional support (Lewis 1980).

The libertarian concept of press freedom, most fully developed and most ardently supported in the US, focuses exclusively on negative freedom. Since the 1940s, it has been challenged by views that emphasize 'freedom for' as much as 'freedom from.' Concerned about the inexorably growing influence of the mass media on society, so-called 'social responsibility' theories of the press have emerged as a significant counterpoise to the *laissez-faire* libertarian view. The increasing significance of broadcasting, regulated by government 'in the public interest' from early on, has no doubt contributed to this emphasis.

The criticism of a libertarian concept of press freedom stems from an innovative proposition that freedom of the press carries little meaningful value unless it is supported by an independence of editorial decision-making from both internal and external pressures. In this context, 'journalistic freedom' is distinguished from press freedom. Journalistic freedom is concerned with a relationship between the working journalists of a news medium and the managing executives and editors of the news medium. It is an individualized, not institutional, type of press freedom (Merrill 1989, Baker 1989).

In the US, press freedom is to a large degree guaranteed and practiced institutionally, but journalistic freedom is rarely accepted. But several free-press countries in Europe and Asia try to accommodate the conflict between the institutional press freedom and the internal journalistic freedom. A case in point is South Korea. To ensure separation between the editorial department and management, the *Hankyoreh Shinmun* (One Nation Newspaper) and several Korean newspapers have reporters directly involved in deciding on the overall editorial policy of their newspapers, with managing editors elected by reporters or selected from those recommended by them. The stronger voice of the editorial departments of an increasing number of Korean newspapers has resulted from reactivated labor unions of journalists since 1987, when the Korean government embraced a series of sweeping sociopolitical democratic reforms.

3. Censorship of the Press

In connection with press freedom, both institutional and journalistic, censorship is an unending concern to the print and broadcast media around the world. Censorship is 'the systematic control of the content of any communications medium, or of several or all of the media, by means of constitutional, judicial, administrative, financial or purely physical measures imposed directly by, or with the connivance of, the ruling power or a ruling elite. It may or may not be accompanied by violence, it may or may not be total, it may or may not include propaganda. But the concept of a *systematic* imposition of controls is...essential to determining whether or not censorship is being practised in a given country, just as is the concept of intent' (Scammell 1988).

Censorship is intertwined with power so inextricably that it is referred to as the 'handmaiden of power.' Censorship in the sense of prior restraint on the press is more effective than subsequent punishment because of its broader scope and lack of countervailing mechanism such as public appraisal of its rationale. And it has a propensity to decide against publication. 'The function of the censor is to censor. He has a professional interest in finding things to suppress' (Emerson 1955).

Censorship can be 'total' when all the media are systematically controlled, 'partial or selective' when the censorship is selectively enforced, or 'random' when it is capricious or arbitrary. Further, the methods of censorship are multifarious, ranging from prior restraint to fines and prison sentences for violations to suspensions of a publication to closing down a publication to the many varieties of self-censorship journalists exercise to stay out of trouble (Scammell 1988).

Does freedom of the press start to flourish when control of the media through state censorship ends? It is not as simple as that. An enterprising press is not the rule but the exception in societies in transition to democracy. Where journalists have been accustomed to being told what to report and what not to report, they have been trained and tamed by censorship and do not necessarily take advantage of their new-found freedom easily (Williams 2000).

3.1 Censorship Checklist

Censorship is a crucial criterion against which press freedom is measured. The World Press Freedom Center (WPFC) in the US asks 20 questions to determine whether the news media in a particular country are free:

- (a) Are there restrictive press laws?
- (b) Who owns the media, print, and broadcast? Private, government, or both?
- (c) Are journalists prosecuted for what they write?

- (d) Are journalists in jail? Reasons.
- (e) Is libel a civil or criminal offense?
- (f) Are journalists required to have a government enforced license to work?
- (g) Do journalists have unrestricted access to government proceedings?
- (h) Are journalists harassed while covering the news?
- (i) Does the government pay journalists?
- (j) Are newspapers or broadcasters subsidized?
- (k) Is there government-supported censorship? Is there self-censorship?
- (l) Are there restrictions on the means of production, such as government allocations of paper, control of distribution systems, and ownership of printing facilities?
- (m) Is government advertising allocated fairly?
- (n) Is there a legally mandated right of reply, which includes government officials?
- (o) Are insult laws routinely used to shield officials' conduct from public scrutiny?
- (p) Are courts able to judge news media cases independently?
- (q) To what extent are media outlets owned by political, government-linked entities, or others desiring to control content?
- (r) Are crimes against journalists prosecuted by authorities?
- (s) Are the activities of government—courts, legislature, officials, records—open to the press?
- (t) Do journalists themselves consider themselves free to write or broadcast the news as they find it? (Greene 2000).

3.1.1 Right of reply. The WPFC checklist reflects a US-centric approach to evaluating press freedom against the exercise of censorship in that it emphasizes the negative-freedom view on the press rather than positive freedom. The right of reply is a good illustration. In the US, the mandatory right of reply statutes governing the news media as a whole and the print press in particular are seen as violations of freedom of the press.

The US Supreme Court in 1974 struck down a right of reply statute on the ground that it would infringe upon the editorial rights of the press: 'A newspaper is more than a passive receptacle or conduit for news, comment, and advertising. The choice of material to go into a newspaper, and the decisions made as to limitations on the size and content of the paper, and treatment of public issues and public officials—whether fair or unfair—constitute the exercise of editorial control and judgment. It has yet to be demonstrated how governmental regulation of this crucial process can be exercised consistent with First Amendment guarantees of a free press as they have evolved to this time' (Miami Herald Publishing Co. v. Tornillo 1974)

At least 13 countries in Europe, including Germany and France, however, recognize the right of reply. The right of reply in German law is derived from French law. The policy justifications for the right of reply in Germany are not so much to provide the public with access to the media as to protect individuals from false defamation. This differentiates German law from the South Korean law on the right of reply. The right of reply in Korea is justified partly because it can contribute to the public's discovering truth and forming diverse public opinions by creating a more open marketplace of ideas by enhancing the objectivity of news information. In this vein, the right of reply offers access for the Korean public to the news media in a positive way. Especially noteworthy is the international acceptance of the right of reply. The International Convention on the Right of Correction and the American Convention on Human Rights provide the right of reply.

The explicit recognition of the right of reply in several civil-law countries in Europe and Asia and under international law signifies that there are alternative visions of press freedom, which stand in sharp contrast with the US approach to the negative freedom of the press, that is, access to the press is not entirely antithetical to freedom of the press, but it can be viewed as a plausible 'positive' notion of press freedom in ensuring more opportunity for freedom of expression.

3.1.2 Press subsidies. Do government subsidies for the press create a cause of concern for censorship? Evidence from the Scandinavian press subsidies disproves an assumption underlying the adversarial press-state relations in America that newspapers receiving direct public subsidies will lose their editorial independence and autonomy to the government. The Swedish subsidy program 'has had little effect on the willingness of subsidized publications to criticize government' (Gothberg 1983, Gustafsson 1980)

3.1.3 Licensing of journalists. Licensing of journalists is incompatible with freedom of the press as a basic human right. As the Inter-American Court of Human Rights (IACHR) ruled in 1985: 'Such a law (on licensing of journalists) would contain restrictions to freedom of expression...and would consequently be in violation of not only the right of each individual to seek and impart information and ideas through any means of his choice, but also the right of the public at large to receive information without any interference' (IACHR 1985).

3.1.4 Administrative vs. judicial censorship. The legality and scope of press censorship is sometimes

examined by focusing on who primarily dictates the censorship mechanism. Is censorship entirely in the hands of executive officers with no judicial oversight? On this issue, the US differs from Japan. The US law does not allow even court-issued prior restraint under the First Amendment to the US Constitution. In Japan, however, the anticensorship principle does not extend to judicial restrictions because censorship is limited to administrative imposition of prior restraint on the press.

4. Self-censorship of the Press

Self-censorship is viewed as a widespread and insidious problem among reporters and editors. Journalists resort to self-censorship when they decide to avoid legitimate stories not because of their professional judgments but on nonprofessional grounds. In some cases, self-censorship can be intimately intertwined with state censorship. This is the case in Israel: 'When acting as concerned members of society, editors and journalists are of one mind with the censors concerning the prohibition of public exposure of issues considered crucial to the physical welfare of Israel or to its ideological *raison d'être* (even though the public may not share that view)' (Liebes 1997).

Intrusion of commercial interests into editorial decision-making is another major cause of self-censorship. 'Market forces are seen as the primary reason why worthwhile stories are not pursued, and this factor is especially prevalent in the broadcast arena,' reports a recent survey of American journalists on self-censorship (Kohut 2000). Self-censorship is related closely to threat of lawsuit for defamation, invasion of privacy, breach of confidentiality agreements, and related issues.

4.1 Libel Law

Libel was one of the earliest legal actions available against the press, and it is still the most common legal danger to the media. Libel is an occupational hazard for media professionals. Seventy percent of all libel actions are filed against the mass media; at least two-thirds of those actions are brought against newspapers in the US. Libel law protects the reputational interests of individuals because it involves the expression of opinions about them by others. It also attempts to compensate for economic and emotional injury and promotes human dignity by providing a civilized forum in which a dispute is settled. Most importantly, however, libel law acts as a deterrent on the publication of false and injurious speech through the award of damages. In addition, libel law serves a vital social interest by providing a check on media power. It

opens the newsgathering and decision-making process of the media to public scrutiny and accountability.

4.1.1 Transnational libel litigation. The globalizing media confront a new type of libel threats as off-shore litigation against US and other transnational news media is growing. One much-publicized case involved Ariel Sharon, the former Defense Minister of Israel, who sued against Time, Inc. for a defamatory article published in *Time* magazine. Sharon filed his lawsuits in a US federal court and in an Israeli court in 1983.

Most recently, former Indonesian President Suharto filed a \$27 billion lawsuit against the US-based *Time* magazine in a Jakarta court. Suharto claimed that *Time* published an article alleging that he had transferred billions of dollars from Switzerland to an account in Austria in an effort to stash his multibillion-dollar fortune abroad. The Indonesian court ruled in November 1999 that it had the authority to try Suharto's civil libel suit against *Time*.

Suing American media in foreign courts is one thing; enforcing foreign court judgments in the US is another. Thus far, American courts have consistently refused to recognize and enforce foreign libel judgments, holding that the judgments were based on libel law repugnant to the US Constitution on press freedom.

4.1.2 Internet libel. The 'information superhighway' epitomized by the Internet exerts an enormous impact on libel law all over the world. This is especially the case with jurisdiction and enforcement of libel law. While other countries are still in the process of sorting out a vast array of jurisdictional and other issues, the US Congress in 1996 passed a law immunizing Internet service providers (ISPs) from liability for third-party postings of libelous material. One leading English libel lawyer suggested that the US law would pressure England and other European countries to adopt similar standards for Internet libel 'to reflect the transnational nature of such publications' (Hooper 2000).

5. Preventive Counseling for News Media

Because freedom of the press is litigated in increasing frequency in the US, England, and other countries, media organizations involve lawyers in prebroadcasting and prepublication reviews. Lawyers in newsrooms have now become a fact of life for many, if not all, reporters and editors in dealing with a host of legal problems ranging from libel to access to in-

formation. Some journalists are concerned about the chilling effect of lawyers' presence in and outside the American newsroom. CBS News anchor Dan Rather claimed: 'I'd rather take my chances in front of a jury than take my chances in front of corporate lawyers' (Grossman 1996). Nonetheless, the journalist-lawyer relationship is not necessarily harmful to freedom of the press insofar as lawyers are sufficiently 'sensitive to the murky line between legal advice and journalistic or artistic judgment' (Smolla 2000). No less important is the fact that journalists, the integral part of the media system, are more often responsible than lawyers for determining what to broadcast or publish or what not to broadcast or publish in the US. Likewise, a 1997 study of the impact of libel law on the media both in England and in Scotland concluded, '(A) solicitor's advice is just that, and ... on occasion editors do decide to ignore it' (Barendt et al. 1997).

6. Access to Information and Free Press vs. Fair Trial

The press's role of serving as a 'public watchdog' on the government and of imparting information on matters of public interest is difficult, if not impossible, to fulfill without access to public records. Without some protection for seeking out the news, freedom of the press could be a farce. Now freedom of information (FOI) is accepted more widely around the world than ever. This is a profound development in freedom of the press as a positive concept. In this connection, it should come as no surprise that the Scandinavian countries that embrace a 'proactive' concept of press freedom has the most comprehensive FOI laws. For example, Sweden, whose government information statute is considered the oldest in the world, has adopted the policy on access to official records as a fundamental foundation of its democracy since 1766. The presumption of openness in government documents in Sweden is unambiguous when the Swedish FOI act states that all official records are public if there is no specific statute restricting the access to the records.

Traditionally, freedom of the press has been confined to the right of news reporters and editors to disseminate any information in their possession, which does not necessarily extend to the legal right for the media to gather information. But newsgathering has now emerged as a vital issue for most journalists and news media organizations. In the US, how to deal with access to government records, meetings, institutions, courtrooms, and the like constitutes by far the largest body of mass media law.

Freedom of the press is often in conflict with the accused's right to a fair trial in that court proceedings and judicial records are an important source of news information, and trials are government proceedings for the public to know. Some countries such as

Australia, Sweden, and the US favor open proceedings. Other countries including the UK are more restrictive of the press coverage of trial proceedings.

7. Journalist's Privilege to Shield Confidential Sources

Equally important is the issue of whether journalists should be privileged to protect their sources' confidentiality from forced disclosure. In the US, a journalist's privilege is qualified and lags far behind Australia, France, Germany, and Sweden, where the strongest legal protection is provided to the confidentiality of sources. In the European countries, the protection of confidential sources is premised on the conclusion that 'society is better served by encouraging people to disclose matters of public interest to the press, even when the information may include evidence of their own wrongdoing, than by identifying and possibly convicting a particular wrongdoer,' and 'journalists rarely divulge information that would compromise sources, courts rarely compel disclosure, and even more rarely do courts enforce a disclosure order through imprisonment' (Coliver 1993).

8. Freedom of the US Press in the Globalizing Era

Freedom of speech and the press has been the international buzzword since the end of the Cold War in the late 1980s. And the positive concept of press freedom has crystallized the undisputed paradigm throughout the world and it inspires the standard for quality journalism. The growing impact of the now modified libertarian press theory, with an emphasis on the affirmative role of the state in expanding press freedom, on many newly independent countries and emerging democracies in Asia, former eastern Europe, and Latin Africa is credited partly to the American successful experiment with a free press over the past 200 years. Professor Owen M. Fiss of Yale Law School offers a cogent note: 'In building a free press, the reformers [in formerly Soviet bloc nations] should look to the American experience, but only selectively. They must create for the press a measure of autonomy from the state without delivering the press totally and completely to the vicissitudes of the market' (Fiss 1996).

See also: Censorship and Secrecy: Legal Perspectives; Censorship in Research and Scholarship; Control: Social; Cultural Expression and Action; Freedom/Liberty: Impact on the Social Sciences; Freedom: Political; Media and History: Cultural Concerns; Media, Uses of; Political Communication; Political Discourse

Bibliography

- Baker C E 1989 *Human Liberty and Freedom of Speech*. Oxford University Press, New York
- Barendt E, Lustgarten L, Norrie K, Stephenson H 1997 *Libel and the Media*. Clarendon Press, Oxford, UK
- Coliver S 1993 Comparative analysis of press law in European and other democracies. In: Coliver S (ed.) *Press Law and Practice*. Article 19, London
- Emerson T I 1955 The doctrine of prior restraint. *Law & Contemporary Problems* 20: 648–71
- Emerson T I 1966 *Toward a General Theory of the First Amendment*. Random House, New York
- Fiss O M 1996 *Liberalism Divided*. Westview Press, Boulder, CO
- Gothberg J A 1983 Newspaper subsidies in Sweden pose no dangers. *Journalism Quarterly* 60: 629–34
- Greene M J 2000 *New Code Words for Censorship*. World Press Freedom Committee, Reston, VA
- Grossman L 1996 Lessons of the 60-minute Cave-In. *Columbia Journalism Review* Jan/Feb: 39–51
- Gustafsson K 1980 The press subsidies of Sweden: A decade of experiment. In: Smith A (ed.) *Newspapers and Democracy*. MIT Press, Cambridge, MA
- Hooper D 2000 *Reputations under Fire*. Little, Brown, London
- Inter-American Court of Human Rights (IACHR), San Jose, Advisory Opinion of November 13, 1985, No. OC-5/85—Compulsory membership in an association prescribed by law for the practice of journalism (Arts. 13 and 29 American Convention on Human rights)
- Kohut A 2000 Self-censorship: Counting the ways. *Columbia Journalism Review* May/June: 42
- Lahav P 1985 Conclusion: An outline for a general theory of press law in a democracy. In: Lahav P (ed.) *Press Law in Modern Democracies*. Longman, New York
- Lewis A 1980 A public right to know about public institutions: The first amendment as a sword. *Supreme Court Review*, pp. 1–25
- Liebes T 1997 *Reporting the Arab–Israeli Conflict*. Routledge, New York
- Lovell v. Griffith, 1938 303 U.S. 444
- Miami Herald Publishing Co. v. Tornillo 1974 418 U.S. 241
- Merrill J C 1989 *The Dialectic in Journalism*. Louisiana State University Press, Baton Rouge, LA
- Scammell M 1988 Censorship and its history: A personal view. In: Boyle K (ed.) *Article 19 World Report 1988*. Times Books, New York
- Smolla R 1999 *Law of Defamation*, 2nd edn. West Group, St. Paul, MN
- Williams L 2000 Censors at work, censors out of work. In: Williams L, Rich R (eds.) *Losing Control: Freedom of the Press in Asia*. Asia Pacific Press, Canberra, Australia

K. H. Youm

Freedom: Political

1. Introduction: Liberty, Equality, Fraternity;
Two Concepts of Liberty

Since the bourgeois revolutions, modern states have been committed, at least formally, to the ideals of liberty, equality, and fraternity. From the start there

have been tensions between those three aspirations. Attempts to resolve these tensions have taken different forms, from rampant *laissez-faire* libertarianism to interventionist, redistributivist socialism. The theories examined in this article focus mainly on the relation between liberty and equality, with an occasional nod towards fraternity. Since political freedom or liberty is an essentially contested concept, it is impossible to start from a shared definition of it. One can only trace the sources of a number of incompatible conceptions and establish an elementary taxonomy of them.

An important moment in the modern debate about political freedom is Berlin's *Two Concepts of Liberty* (1958). It is an unsystematic mixture of different elements: conceptual analysis, political engagement in the postwar antitotalitarian struggle, and a rich, if eclectic, essay in intellectual history. But for all its shortcomings, it made the opposition between positive liberty and negative liberty central to much subsequent discussion.

Berlin's political task was to unmask the perversion of language by Stalinist ideologues, who presented the worst forms of oppression as 'real (socialist)' freedom, as opposed to 'mere bourgeois (formal)' freedoms. Berlin aligned the former with 'positive,' and the latter with 'negative,' liberty, and reasserted the value of the latter against its critics.

Historically, Berlin identified negative liberty with liberalism, the tradition that defended the individual against 'interference,' particularly when exercised by the state. But the idea of 'interference' is hard to tie down. Berlin himself was ambivalent as to whether the interference must be deliberate or not, but he was adamant that an inability caused by the social structure (e.g., the inability of the worker to control the means of production in traditional Marxist theory) is not a case of unfreedom. He agreed that the 'classical English political theorists' were 'prepared to curtail freedom in the interests of other values and indeed of freedom itself.' They held that without law, 'freedom for the pike is death for the minnow.' A natural interpretation of that insight is that the law positively creates freedom for the minnow, allowing it to do things it would otherwise be unable to do. In the words of Locke (a quintessential negative libertarian for Berlin): 'the end of Law is not to abolish or restrain, but to preserve and enlarge freedom.' Berlin could agree that the pike's freedom is restricted for the sake of the minnow's freedom ('curtailed in the interests of liberty itself'), but his polemical drive was against partisans of positive liberty who would maintain that the pike's freedom too is (really) enhanced by the restriction.

Another ambivalence within negative liberty concerns the connexion between freedom and wishes. Berlin began by tying the two together directly, defining oppression as 'the part I believe to be played by other human beings ... in frustrating my wishes.' But later he rejected the definition of freedom as 'the

ability to do what one wants' because it is compatible with a situation in which all one's wants are restricted by brainwashing or the removal of desirable objects. At the heart of negative liberty lies 'the right to be let alone,' but that right turns out to be remarkably elusive.

If negative liberty (NL) is hard to tie down, positive liberty (PL) is even more so. PL contains both (a) political elements (PL-pol) and (b) moral ones (PL-mor). (a) is essentially the freedom to participate in the political process; (b) includes the ability to be the autonomous master of one's own drives, which is associated with the romantic notion of the divided self. Rousseau says that both (a) and (b) are the outcome of entering civil society, but he is careful not to identify them. When Constant (1988) introduced the distinction between the 'liberty of the ancients' and the 'liberty of the moderns,' he was concerned with (a). His polemical claim was that the 'liberty of the ancients' nowadays meant the unbridled exercise of popular sovereignty (PL-pol) to the detriment of NL. For Berlin, the danger of modern totalitarian ideology is that it harnesses (a) to (b). It suggests that, through laws endorsed by 'the people,' irrational empirical selves should be forced to coincide with rational ends (their 'real selves').

Five broad approaches to political freedom can now be distinguished. In each of these, the three fundamental values of liberty, equality, and fraternity are addressed. Writers inspired by Berlin focus on incompatibilities between those values. Others seek to show that when they are more clearly understood, it will be seen either that they are compatible or that the very idea of balancing one against the other must be abandoned. In their analyses, the terminology of PL and NL frequently recurs, but the distinction made above between dimensions (a) and (b) of PL frequently is ignored.

The approaches to be considered are: (a) Welfare liberalism; (b) Libertarianism; (c) Freedom, economics and power; (d) Freedom and community; (e) Civic republicanism.

1.1 Welfare Liberalism

For theorists of welfare liberalism, political freedom is exercised in a context of representative democracy constrained by constitutional checks to protect individual rights and by a fluctuating sense of obligation to redress the inequalities thrown up by the operation of the market. The two outstanding proponents of this approach are John Rawls and Ronald Dworkin.

1.1.1 Rawls. Rawls's *A Theory of Justice* (1971) is the most comprehensive, systematic work of modern political philosophy. In it the author expounds the

moral basis of pluralist welfarism, combining the three elements of liberty, equality, and fraternity into a coherent whole.

At the heart of the book lie the Two Principles of Justice. The First Principle states that 'Each person is to have an equal right to the most extensive total system of basic liberties compatible with a similar system of liberty for all.' 'Basic liberties' are specified as 'roughly speaking, political liberty (the right to vote and to be eligible for public office) together with freedom of speech and assembly; liberty of conscience and freedom of thought; freedom of the person along with the right to hold (personal) property; and freedom from arbitrary arrest and seizure as defined by the concept of the rule of law.' These are what Marxists called 'bourgeois liberties.' The list is not equivalent to NL-pol, since they include certain 'positive' political elements, even if the 'negative' liberties of privacy may be more important than 'positive' ones to modern citizens.

In subsequent work, Rawls (1993) has responded to Hart's criticism that the idea of 'maximizing freedom' is incoherent. He has rewritten the First Principle to read: 'Each person has a claim to a fully adequate scheme of basic rights and liberties, which scheme is compatible with the same scheme for all; and in this scheme the equal political liberties, and only those liberties, are to be guaranteed their fair value.' He has also added the 'integrity' (physical and psychological security) of the person to the list of basic liberties to be protected by the First Principle. But he has not modified the central claims of *A Theory of Justice* that the modern liberal state must (a) guarantee 'basic liberties,' and (b) strive to maximize the welfare of the worst-off members of society by redistributive mechanisms, and that the First Principle must not be sacrificed to the Second, since 'liberty can be restricted only for the sake of liberty.'

Most of the other important revisions have concerned structure and methodology rather than content. Rawls has anchored the Two Principles more explicitly in a conception of citizens of a modern pluralist state, as it has developed since the Reformation. They are 'regarded as free and equal persons,' endowed with 'two moral powers ... namely a capacity for a sense of justice and a capacity for a conception of the good.' Such citizens display their freedom in the following ways. They have 'the moral power to have a conception of the good' and are 'capable of revising and changing this conception.' They 'regard themselves as self-authenticating sources of moral claims.' They are capable of taking responsibility for their 'ends.'

The Two Principles of justice remain unaltered and are still subject to attack from both left and right. Left-wing critics charge that First Principle 'bourgeois' liberties are hollow and worthless for the mass of the population. To them Rawls had already responded in *A Theory of Justice* by distinguishing equal liberty

from equal worth of liberty, acknowledging that it is the latter that is important, and predicting, perhaps optimistically, that it would be enhanced by the application of the Difference Principle. Right-wing 'libertarian' critics reject all redistributive mechanisms, because they infringe the individual's unrestricted right to accumulate property. Rawls' response to them is also contained in *A Theory of Justice*, namely that individuals have a basic right to own *personal* property, but that right does not entail any particular property system, whether capitalist or socialist. In each particular case the choice of system will be determined by its ability to maximize the welfare of the worst-off.

1.1.2 Dworkin. Dworkin's political philosophy is an attempt to 'fine-tune' both the methodology and the political recommendations of Rawls's welfare liberalism. Methodologically, Dworkin denies that liberalism must balance liberty against equality. That image wrongly presupposes that liberty can be quantified, so that 'fundamental' liberties weigh more heavily than trivial ones. Instead, 'the fundamental liberties are important because we value something else that they protect.' And that 'something else' is, unexpectedly, a form of equality, the principle that a government should treat all its citizens with 'equal concern and respect.' The 'fundamental' liberties, then, are entailed by the equality principle itself, according to which 'the government (should) treat all its citizens as free, or as independent, or with equal dignity.' Dworkin accepts the Rawlsian requirement that 'political decisions must be ... independent of any particular conception of the good life ...' Thus governments are bound by the negative requirement to be neutral between lifestyles, and also by the positive requirement to 'respect' the 'dignity and independence' of its citizens.

In terms of political recommendations, Dworkin proceeds from a tension at the heart of Rawls's Two Principles. While the First Principle defends the conditions of basic freedoms, and addresses people as autonomous agents, the Second Principle might undermine people's motivation to exercise their freedom to 'better themselves,' addressing them as passive recipients of benefits. In response to this tension, Dworkin introduces a key modification. He follows Rawls in compensating people for 'irrelevant differences, like differences in talent.' But he argues that people should be 'responsible for the costs of their choices.' The system, in short, must be both 'ambition-sensitive' and 'endowment-insensitive.' But Dworkin's modification may have introduced a new tension of its own, since however finely tuned it is, its two requirements must work against one another: 'the more we try to make the distribution sensitive to people's ambitions, the more likely it is that some

people disadvantaged by circumstances will be undeservedly penalized, and vice versa' (Guest 1997). Everything now hinges on the possibility of making a realistic distinction between ambition and endowment, between 'personality' and 'circumstances.' The fact that this is so difficult suggests that the tension between freedom and equality may still be unresolved.

2. *Libertarianism, Freedom as the Unrestricted Right to Own Private Property: Nozick*

Nozick's *Anarchy, State and Utopia* (1974) was a polemical response to Rawls' *A Theory of Justice*. At the time of its publication, there was a backlash against welfarism and state intervention in the economy. That backlash was driven by the economic theory that managed markets are inefficient, since redistribution through taxation takes away incentives. But although Nozick's book was congenial to that theory, his philosophical position is independent of it. Nozick would defend his position on the grounds that it was the correct one, even if it were inefficient. He would in fact reject appeals to economic efficiency as improperly utilitarian.

Though labeled a 'libertarian,' Nozick's starting point is not the idea of 'liberty' or 'freedom,' but rather the Lockean one of 'self-ownership.' This has polemical value because it allows Nozick to proceed against Rawls' (and other welfarists') programs of redistribution by appealing to a hatred of slavery shared by all across the political spectrum since the 'bourgeois revolutions,' and suggesting, by a rhetorical sleight of hand, that 'taxation of earnings from labor is on a par with forced labor.' The slide has some source in Locke's 'labour theory of property,' the idea that you have a property right to something if you have 'mixed your labour with it.' But Locke did not suggest that that right is absolute in the same sense that your right to your bodily integrity is.

The sleight of hand at the heart of libertarianism is this. We all agree that if A has the right to make use of B's body, of which B is the rightful owner, without B's consent, then B is A's slave. By parity of reason, it is suggested, if A has the right to make use of B's property, of which B is the rightful owner, without B's consent, by taxing B, then B is A's slave. The sleight of hand depends on accepting the extension of self-ownership from ownership of one's own body to ownership of whatever one has acquired legally, an extension which is not evidently justified. Nonetheless, libertarians may be able to deploy a 'slippery slope' argument against welfarism without making this particular sleight of hand. Rawls's Two Principles of justice are intended 'to balance the twin demands of (a) respecting choices and (b) rectifying circumstances.' Rawls intends (a) to be applied to individuals, (b) to groups. But, according to the libertarian, there is

nothing within Rawls's theory to prevent the radical welfarist from extending rectificatory treatment to individuals too. How is one to define what intervention is legitimate in order to rectify circumstances? If the line is to be drawn around the body, it would seem to mean that any nonbodily interference is at least in principle legitimate in order to rectify circumstances. And that sounds sinister even to those who are not extreme 'libertarians.'

3. Freedom, Economics, and Power

In the past, egalitarians, Marxists and non-Marxists alike, derided 'bourgeois freedoms' for masking from the workers their real enslavement. The modern egalitarians considered in this section all take seriously the demand for real equality of freedom. But unlike their predecessors, they all also take seriously the importance of the 'basic' freedoms that lie at the heart of liberalism. The three positions considered in this section are linked by the insight that freedom and equality are linked intimately, and that the conditions of equality and inequality are zero-sum relations of economic power. In each of the three, real freedom is 'positive' in the sense that it requires more than just the 'right to be let alone.' On the other hand, it is not the 'liberty of the ancients,' since it recognizes the importance of private life in the modern world.

3.1 Neo-Marxism: Cohen

In the most sophisticated version of neo-Marxism, that of Cohen, 'there is ... a conflict between social equality and the liberty of some people,' but that should not stand in the way of 'the pursuit of social equality, since a humane concern for liberty must first of all direct itself to the condition of those who enjoy hardly any of it.'

Neo-Marxists thus agree with liberals that freedom of the individual is the most important political value, and that modern capitalism delivers it to all members of society, including proletarians, in several important ways. First, all members of a capitalist society (worker and capitalist alike) are legally free. Second, any particular worker is always legally free, and often economically free, to leave the proletariat and become a petty bourgeois or even a capitalist. But, within a capitalist society, 'although most proletarians are free to escape the proletariat, indeed even if all are, the proletariat is an imprisoned class.' Third, capitalism has delivered 'important freedoms beyond that of buying and selling.' These include 'freedom of speech, assembly, worship, publication, movement, political participation ...'. The neo-Marxist Cohen is agnostic about 'how accidental the connexion between capitalism and those freedoms has been and is,' but is committed to the view that only 'freedom to buy and

sell belongs to capitalism's inmost nature.' But Marxists must still acknowledge that 'bourgeois freedoms' really are freedoms: '... when socialists suggest that there is no real liberty under capitalism, or that socialism promises liberty of a higher and unprecedented kind ... their line is theoretically incorrect and politically disastrous. For liberty under capitalism is, where it exists, just that, liberty; and if socialism will not give us plenty of it, we shall rightly be disappointed.'

In recent writings, Cohen (1995) has begun to investigate the idea of real freedom, understood as 'autonomy, the circumstances of genuine control over one's own life.' With this turn, neo-Marxism returns to the old site of conflict, since we must 'ask what kind and degree of control over external things a person must have to enjoy autonomy, and then to ask whether such control is compatible with socialist equality.' Cohen's project now is to translate the idea of real freedom into terms that are applicable to the real world, and, where possible, quantifiable. His model of 'equal access to advantage' now stands alongside van Parijs's (1995) and Sen's (1992) models. It thus forms part of a rich developing research program. Whether it should still be called Marxist is another question.

3.2 Post-Marxism: van Parijs

In the post-Marxist program of van Parijs, the normative goal of equalizing access to freedom is maintained. At the same time, the Marxian *theory* of exploitation is abandoned, and the capitalist market is now seen as the only vehicle that can deliver 'real freedom for all' in *practice*. The goal is to 'empower' the greatest number of people. The method is to guarantee them an unconditional 'basic income' that will prevent personal dependence and allow individuals to make meaningful choices between life-styles. The 'basic income' project has now entered the real political agenda in western Europe.

The goal of the post-Marxists is to reconcile freedom and equality by the principle to 'maximization of an unconditional income, subject to formal freedom being protected and to no-one's endowment being unanimously found worse than someone else's.' That would guarantee the widest, and fairest, distribution of 'real' freedom that could also be delivered by a 'sustainable' economic system. Van Parijs maintains that that system would be a highly regulated, interventionist form of capitalism. His reasons are fourfold. First, it is (probably) more efficient than socialism. Second, it can (probably) guarantee a higher basic income than socialism can. Third, it can (certainly) guarantee formal freedoms, whereas socialist régimes have still to prove that they can. Last, and this is a decisive consideration, even if there are reforms of international economic affairs, it will be impossible to promote an 'autarkic' socialist régime in one country

without imposing unacceptable restrictions on the movement of its people to other (higher paying) countries.

3.3 *Neo-Aristotelian Freedom as Capability: Sen*

The third egalitarian model combines sophisticated economic theory with a deep philosophical anthropology derived from Aristotle. The leading proponent of this model is Sen.

The aim here is to construct a model of freedom which is sufficiently objective to be used in cross-cultural comparisons, but which is also sensitive to the different weightings given to the elements of a flourishing life in different cultures. Sen (1992) draws up a list of 'functionings' that include 'being adequately nourished' and 'having self-respect,' and defines 'well-being freedom' as the capability to achieve those functionings. But there is also 'agency-freedom,' defined as the capability to achieve whatever one's objectives are. Since many people have objectives other than their own well being, these two freedoms may conflict.

It is agency freedom that has been at the center of most liberal discussion. Sen acknowledges its importance, but weighs it alongside well-being freedom in evaluating the totality of a person's life. Even within agency freedom, Sen advises liberals not to exaggerate the importance of the agent's direct control of an outcome. Instead, he proposes a counterfactual formula, 'I am free when control is exercised in line with what I would choose.' That formula can be generalized to cover all forms of freedom, even when the term 'freedom' is extended to cover 'freedom from hunger,' 'freedom from malaria,' etc.: 'One values living without malaria, desires such a life, and would have chosen it, *given the choice*.'

Society has different obligations to its members concerning their well being and their agency. It has an evident 'responsibility to a person's well-being, especially when it is in some danger of being particularly low' which it does not have for that person's particular agency objectives. In advanced liberal societies at least, 'in which the focus is on the *freedom* that people have,' no patterned outcome can be imposed. For 'even when complete equality of well-being freedom is ... entirely realized, this need not lead to an equality of well-being *achieved*, since different individuals may give different priority to the pursuit of their own well-being.' In rejecting patterned outcomes, Sen has shown that the capability model does not conflict with the central values of liberalism expounded by Rawls and Dworkin.

The capability model inherits from Aristotle the idea that human beings are endowed with specific capacities that can be more or less fully realized in a flourishing life. In his presentation of the model, Sen is concerned that Aristotle's particular 'view of human

nature (with a unique list of functionings for a good human life) may be tremendously overspecified.' He prefers to leave open 'both ... the exact grounds underlying the determination of relative weights, and the actual relative weights chosen.' This gives Sen's version of the 'capability approach' a degree of incompleteness relative to Aristotle's, but also, he thinks, a 'considerable "cutting power"' for use in the real world of cultural diversity.

3.4 *Freedom as Attaining Autonomy Through Community: Raz*

Communitarian critics including Taylor, MacIntyre, and Sandel brought two charges against liberalism. At the level of theory, they claimed that it is based on a methodological individualism which in turn derives from an impoverished, atomistic conception of human nature. At the level of practice, they argued that because the liberal order fails to provide a satisfactory sense of identity for its members, 'intolerance flourishes most where forms of life are dislocated, roots unsettled, traditions undone.'

Along with the communitarians, Raz (1986) rejects methodological individualism, but against them he argues that this does not entail a rejection of liberal values. Raz's liberalism, 'the doctrine of limited government,' is rooted in a 'non-individualist conception of morality.' In opposition to Rawls's antifoundationalism, Raz propounds a 'doctrine of political authority which rests on a perfectionist political defence and promotion of liberty and autonomy.' But this is not a brutish libertarianism, since freedom is 'a distinct value, but one which is intimately intertwined with others, and cannot exist by itself.' Raz in short is a foundationalist, who seeks to establish a 'continuity' between a moral base and a political-legal superstructure.

For someone to attain 'the ideal of autonomy' and live an autonomous life, certain conditions must be satisfied. These are the possession of 'appropriate mental abilities,' of an 'adequate range of options,' and of 'independence.' The ideal is positive, comprehensive. It constitutes the moral base of the system, but is not 'moralistic.' In particular, it does not involve a teleological conception of self-realization, since, as an autonomous person, one is free to choose whether or not to realize oneself. But without teleology it is difficult to make autonomy foundational. If autonomy involves integrity ('identification with one's life'), self-awareness and commitment to a project, it can still be asked why it is valuable. It is insufficient to answer that 'many people desire to be autonomous,' since people's desires are many, conflicting, and mutable. Rather, says Raz, 'the value of autonomy does not depend on choice ... the conditions of autonomy concern a central aspect of a whole system of values of a society ...' The kind of society in question

is a pluralistic society, which both supports and is supported by this central value 'valuing autonomy leads to the endorsement of moral pluralism.'

Other modern liberals, notably Rawls, are cautious about seeking to found political legitimacy on a particular understanding of freedom. They hold that particular individuals and groups can value freedom for different reasons, some welcoming it because it is a purely negative guarantee against interference, others because it embodies their highest aspirations. Against them Raz holds that most people in a liberal state must subscribe to the positive value of freedom as autonomy, otherwise its tolerant pluralism will perish.

4. Civic Republicanism

In the fifth model of political freedom, the 'liberty of the ancients' re-enters the modern debate. In both versions of 'civic republicanism' considered below, political participation plays a key role, a purely instrumental role in the first, and a constitutive role in the second.

4.1 The Instrumental Notion of Civic Republicanism: Skinner

Skinner's model, known variously as 'civic republicanism,' 'classical republicanism,' and 'neoromanism,' is rooted in the Renaissance, in particular in the idea derived from Machiavelli that a state which is 'free from all external slavery' guarantees 'personal liberty, understood in the ordinary sense ... each citizen remains free from any elements of constraint, especially those which arise from personal dependence and servitude, free to pursue his own chosen ends ... To be free ... is simply to be unconstrained from pursuing whatever goals we may happen to set ourselves.' The independence of the state is linked to the freedom of the individual insofar as 'a self-governing republic is the only type of régime under which a community can hope to attain greatness at the same time as guaranteeing its citizens their individual liberty.' Such a state requires two capacities or virtues from its citizens, courage and determination to defend their community against external attack, and practical reason which qualifies them to play an active part in public life. That means that the state can, legitimately, force the citizen to be free. Despite appearances, this 'strenuous view of citizenship' is, according to Skinner (1990), consistent with negative liberty in two respects. First, it does not presuppose any specific teleological theory of human nature. Second, when we are 'forced to be free,' we are being forced not to 'reason about ends,' but to 'recognize the range of actions we have good reason to perform in order to bring about the ends we actually desire.' Skinner illustrates the latter point by a contrast between 'liberals' like Hobbes and Locke on the one hand, and Machiavelli on the other.

The 'liberals' hold only (a) that the law preserves A's freedom by coercing B (the potential law-breaker). Machiavelli holds both (a) and (b) that the law also preserves A's freedom by coercing A 'into acting in a particular way ... forcing us out of our habitual patterns of self-interested behaviour ... into discharging the full range of our civic duties and thereby ensuring that the free state on which our own liberty depends is itself maintained free of servitude.' In short, because 'freedom depends on service and so on our willingness to cultivate the civic virtues ... we may have to be coerced into virtue and thereby constrained into upholding a liberty which, left to ourselves, we would have undermined.' Skinner comments that 'contemporary liberalism ... is in danger of sweeping the public arena bare of any concepts save self-interest and individual rights.' Ideologies suspicious of the whole modern agenda of individual freedom have stepped in to fill that vacuum. Skinner (1998) thinks that 'republican freedom' provides a robust defence against such reactionary positions.

4.2 The Constitutive Notion of Civic Republicanism: Pettit

In the second version of civic republicanism, espoused by Pettit (1997), Berlin's dichotomy of the positive freedom of self-mastery and the negative freedom of noninterference is not exhaustive. Beyond those two lies republican freedom, which means absence of domination, not absence of interference. Domination differs from interference in important respects. Domination is a function of the relationship of unequal power between persons or groups of persons. It relates to the *capacity* of one person or group of persons to interfere in the choices of other persons or groups, rather than to actual *instances* of interference. But that would not be enough to pick out the undesirable character of domination, since any non-anarchist social order will empower certain persons or groups, giving them authority over others in given circumstances. So domination is not just the possession of power, but the possession of *arbitrary* power. The goal of republican freedom, then, is that 'no-one is able to interfere on an arbitrary basis—at their pleasure—in the choices of the free person.' With this formula Pettit seeks to extract the rational kernel of both negative and positive freedom, while showing that neither of them is sufficient in itself. On the one hand, freedom as nondomination 'will facilitate the achievement of autonomy,' as positive freedom requires. On the other hand, it will respect people's empirical choices, as negative freedom requires, but it will go beyond it by showing that dominated people are not really free even if they have the good fortune to suffer no interference when their dominators choose to refrain from exercising their arbitrary power.

Insofar as republican freedom is tied to power, it is essentially egalitarian. It is held to protect each

individual against arbitrary power, and also to be a 'communitarian good,' allowing people to identify with a state that protects their freedom. This version of republican freedom is heavily influenced by Rousseau, purged of totalitarian accretions and updated to the advanced capitalist societies of the late twentieth century. They are now explicitly inclusive, bestowing their benefits on all members of society, and also multicultural, displaying liberal neutrality towards different substantive conceptions of the good. How far such societies can provide a stable balance between the participatory core of republican freedom and the centrifugal drives of modern pluralism remains to be seen.

See also: Contractarianism; Equality and Inequality: Legal Aspects; Equality of Opportunity; Equality: Philosophical Aspects; Liberalism; Liberalism: Historical Aspects; Libertarianism; Republicanism: Philosophical Aspects; Solidarity: History of the Concept

Bibliography

- Berlin I 1958 *Four Essays on Liberty*. Oxford University Press, London
- Cohen G A 1995 *Self-ownership, Freedom and Equality*. Cambridge University Press, Cambridge, UK
- Constant B 1988 The liberty of the ancients compared with that of the moderns. In: *Political Writings*, (trans. and ed.) Fontana B. Cambridge University Press, Cambridge, UK
- Dworkin R M 1977 *Taking Rights Seriously*. Duckworth, London
- Dworkin R 1981 What is equality Parts 1 and 2. *Philosophy and Public Affairs* 10: 185–246, 283–345
- Guest R 1997 *Ronald Dworkin*. Edinburgh University Press, Edinburgh, UK
- Nozick R 1974 *Anarchy, State, and Utopia*. Blackwell, Oxford, UK
- Pettit P 1997 *Republicanism: A Theory of Freedom and Government*. Clarendon Press, Oxford, UK
- Rawls J 1971 *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA
- Rawls J 1993 *Political Liberalism*. Columbia University Press, New York
- Raz J 1986 *The Morality of Freedom*. Clarendon Press, Oxford, UK
- Sandel M J 1982 *Liberalism and the Limits of Justice*. Cambridge University Press, Cambridge, UK
- Sen A 1992 *Inequality Reexamined*. Clarendon Press, Oxford, UK
- Skinner Q 1990 The republican idea of political liberty. In: Bock G, Skinner Q, Viroli M (eds.) *Machiavelli and Republicanism*. Cambridge University Press, Cambridge, UK
- Skinner Q 1998 *Liberty before Liberalism*. Cambridge University Press, Cambridge, UK
- Taylor C 1979 What's wrong with negative liberty? In: Ryan A (ed.) *Essays in Honour of Isaiah Berlin*. Oxford University Press, Oxford, UK
- Van Parijs P 1995 *Real Freedom for All*. Clarendon Press, Oxford, UK

T. O'Hagan

French Revolution, The

The French Revolution was a momentous historical event that set enduring patterns for modern revolutionary movements and for much of modern politics in general. Contemporaries experienced the French Revolution as a set of interlocking changes or stages that seemed driven by some kind of mechanism or impetus. Analysis of this mechanism has preoccupied revolutionaries and scholars ever since. The conflicting political reactions to the French Revolution fed into the development of ideologies from conservatism to communism. Scholarly controversies about the significance and interpretation of the French Revolution have been animated in particular by Karl Marx and Friedrich Engels' use of it as the prime example of a 'bourgeois revolution' that prepared the way for the triumph of capitalism. Although the controversies over the Marxist interpretation continue, recent scholarship has turned in new directions: gender relations; the notion of citizenship and its connection to religion, ethnicity, and race; and the Revolution's impact on the colonies.

1. The French Revolution as Event

The year 1789 marks the clear beginning of the French Revolution, but opinion is much more divided about its ending: did it end in 1794 with the fall of Maximilien Robespierre, in 1799 with the rise of Napoleon Bonaparte, in 1804 when Napoleon established his empire, or in 1815 with the final defeat of Napoleon at the Battle of Waterloo? Debate on the question of endings cannot be resolved definitively because it is so closely entwined with interpretive questions. The choice of ending depends on the determination of the central significance of the French Revolution.

What most struck contemporaries and has fascinated commentators ever since was the unpredictability and yet seeming inexorability of the unfolding of events. Even in the midst of this unfolding, political actors and witnesses spoke of 'the system' of 'the French Revolution,' as if revolutionary events had an internal logic or mechanism pushing in a consistent direction. After spending years studying this logic, and having published his enduringly influential study of the French Revolution, Alexis de Tocqueville still complained of his difficulty in grasping 'this malady of the French Revolution,' this 'virus of a new and unknown kind' (Tocqueville 1977).

1.1 Revolutionary Stages

As a series of political events, the French Revolution is usually divided into five periods or stages: the constitutional monarchy or liberal revolution of 1789–91; the second revolution and the Terror of 1792–4; the

Directorial regime or liberal republic of 1795–9; Napoleon Bonaparte's consulate of 1799–1804; and the Napoleonic empire of 1804–15. This vertiginous succession of regimes established many of the prototypes of modern political authority and even much of modern political language, including the key distinction between Left and Right (determined by the location of deputies vis-à-vis the speaker in the National Assembly). The very possibility of ascertaining periods or stages—already apparent to contemporaries—raised questions about the natural history or morphology of revolution: was it like a sickness or fever from which a polity eventually recovered? Or was it rather a kind of speeding up of historical development, a bootstrap up to the next stage of capitalism or modernity?

1.2 Models of Political Authority

As the French Revolution moved from stage to stage, the revolutionaries generated at least three different models for modern (secular, nonmonarchical) political authority. By abolishing 'feudalism' (eliminating serfdom and seigneurial dues), suppressing noble titles, declaring the 'rights of man,' and founding the first French republic with universal manhood suffrage (1789–92), the French revolutionaries laid the foundations for democratic, representative government. By arresting dissidents, executing opponents, and carrying out a cultural revolution in the form of de-Christianization and republican political re-education (the Terror of 1793–4), the revolutionaries set the mold for 'totalitarian' government with its repressive insistence on political conformity in the name of a future social good. Napoleon Bonaparte developed a third model of paternalist, authoritarian government by a state with a strong military caste. In other words, the French Revolution adumbrated virtually all the possibilities of modern political authority.

2. The Origins of Ideology

The French Revolution gave birth both to the term 'ideology' and to many different ideologies that have continued to organize political discussion right up to the present day. Under the impact of revolutionary events, tradition lost its givenness, hierarchy no longer seemed natural, and monarchy no longer went without justification. The revolutionaries showed that society could be reorganized by an act of self-conscious political will (the definition of revolution itself). The attempt to reshape hearts and minds through festivals, paintings, engravings, operas, plays, and even ordinary items such as snuffboxes and playing cards drew dramatic attention to revolution as a mode of action. This attempt did not produce unanimity,

however; rather it produced social conflict. As a consequence, revolution brought 'the social' as a category into sharper relief. As early as 1792, Antoine Barnave wrote his *Introduction à la Révolution française*, arguing that new forms of wealth produced ineluctably new forms of political authority (Barnave 1988). As the social determinants of politics became more visible, i.e., more of a problem to be investigated, ideologies took shape as formal doctrines about the proper relationship between politics and society.

2.1 From Conservatism to Communism

The first ideology to emerge from the French Revolution was conservatism. Its founding text, Edmund Burke's *Reflections on the Revolution in France* (1790), was written in response to the events of 1789. The National Assembly's efforts to remake the French constitution led Burke to denounce what he termed a new 'political metaphysics' based on natural rights and abstract reason. Unlike many reactionary royalists, however, Burke did not just excoriate the French Revolution as the incarnation of evil and impiety. He understood that the revolutionaries were in the process of erecting an entirely new foundation for politics, and he undertook to counter them point by point. By arguing for a conservatism that accommodated change through practical experience and gradual amelioration, he admitted the necessity of confronting demands for change and the parallel demand for an elaborated doctrine to support the status quo. Even for conservatives, then, the French Revolution changed the rules of the game. The status quo now required explicit defense.

Although liberalism had its origins in the social contract theory of John Locke and the political economy of Adam Smith, after 1789 liberals too had to announce their position vis-à-vis the French Revolution. Most liberals supported the constitutional revolution of 1789 and rejected the Terror of 1793–4. This distinction was not always easy to uphold, for conservatives routinely accused liberals of supporting policies and positions that might open the way to more radical experiments, just as those of the revolutionaries of 1789 had opened the way to the Terror of 1793–4. The unfolding of stages in the revolutionary process thus shaped ideological debate, either explicitly or implicitly, for many generations after the French Revolution.

Socialists, and especially Marxists, traced their lineage directly back to the French Revolution. The early socialists—Claude Henri de Saint-Simon and Charles Fourier, for example—concluded from the French Revolution that society could be reorganized according to a technocratic project (thus 'utopian socialism' according to Marx). Marxists believed that only class struggle could produce a new social order. According to Marx and Engels, the French Revolution marked

the overthrow of the aristocracy by the bourgeoisie. The bourgeoisie would develop capitalism on the ruins of the feudal order it dismantled, thus preparing the way for a subsequent revolution in which the proletariat formed by capitalist development would overturn the bourgeoisie and replace capitalism with communism. The French Revolution was therefore a crucial stage in the class struggle and an essential forerunner to the future communist order.

Marx's emphasis on the need for self-conscious revolutionary organization had been at least partially inspired by Gracchus Babeuf and his 'Conspiracy of Equals' of 1796. Although Babeuf had few followers and his plans were foiled by police infiltration, his dream of a revolutionary seizure of political power by a secretly organized conspiratorial group lived on through the efforts of the Italian revolutionary Filippo (or Philippe) Buonarroti. Buonarroti, a member of Babeuf's circle in 1796, published a history of the Conspiracy of Equals in 1828. It brought the message of revolutionary seizure of power to secret societies throughout Europe. One of Buonarroti's followers, the French revolutionary socialist Auguste Blanqui, combined Buonarroti's appeal for a secret revolutionary brotherhood with the idea of a 'dictatorship of the proletariat,' a term coined by Blanqui to define the political goal for socialism. Marx and Engels incorporated both Babeuf's and Blanqui's notions into their program for proletarian revolution.

2.2 The Foundations of Modern Social Science

Because the French Revolution brought the relationship between society and politics into sharper focus, it also accelerated the development of the social sciences, first in France and then more broadly in Europe. The Directory regime created the first research institute for the social sciences with its department of political and moral sciences in the Institut. Bonaparte, himself a member of the Institut, took more than 150 scientific experts with him on his expedition to Egypt in 1798. They set up an Institut d'Egypt in Cairo and helped inaugurate 'oriental studies' in Europe. They also discovered the Rosetta Stone which enabled experts to decipher Egyptian hieroglyphics.

The French Revolution had an especially dramatic impact on historical studies in France. The first full-length histories of the Revolution were published in the 1820s by François-Auguste Mignet and Adolphe Thiers, journalists and editors in the political opposition. They used the history of the French Revolution to defend the notion of a constitutional monarchy and to criticize the restored Bourbons. In the 1840s, Louis Blanc, Alphonse Lamartine, and Jules Michelet wrote histories of the Revolution as part of their campaign for a republic. Although Marx did not write a history of the French Revolution of 1789, he did read the French historians and made

constant reference in his work to the French Revolution of 1789.

Marx was not alone in considering the French Revolution an important stage in historical development. As the notion of historical evolution spread in the nineteenth century, the French Revolution almost inevitably became an important benchmark. Auguste Comte, one of the founders of sociology, took as his point of departure the last writings of the Marquis de Condorcet, a French nobleman and revolutionary who committed suicide rather than face the guillotine in 1794. Comte proposed three stages of historical development: the theological, the metaphysical, and the positive or scientific. The French Revolution helped bring about the ongoing transition to the third, positive stage by liberating modern social elements from the dead hand of the Old Regime. The task of social science was to complete this transition, which could only be accomplished finally in ideas and opinions, according to Comte, not by revolutionary action. The ultimate aim of his system was the reorganization of society by means of scientific sociology.

Alexis de Tocqueville, who helped lay the foundations for political sociology and political science with his *L'ancien régime et la Révolution* (1856), considered the French Revolution not as a stage so much as an accelerator of the general modern trend toward democracy and equality. Tocqueville had read the previous historians of the French Revolution, though he rarely cited them, and absorbed the view that social changes had inevitable political repercussions. He wove his analysis of social changes into a subtle examination of French political culture and especially the dominance of the central state. In so doing he created an alternative to the Marxist emphasis on class struggle. Tocqueville considered the French Revolution pivotal in the advance of modernity but largely negative in its effects. Tocqueville saw in the French Revolution the tendency of democratic revolution to reproduce the despotism of monarchical authority by leveling all institutions that might have provided a check on state power. Although the French made their revolution for liberty and equality, they gave up liberty in order to be assured of equality and ended up with an equality of servitude to the Napoleonic state. This line of analysis had a long posterity—Hannah Arendt's analysis of revolution has clear affiliations to Tocqueville's—and continues to influence political and social theory right down to the present.

3. Scholarly Controversies

Since analysis of the French Revolution has played such a prominent role in the development of competing political ideologies, it is hardly surprising that it has generated disputes among scholars as well. In the nineteenth century writers used the French Revolution

to argue for and against monarchy, empire, or republicanism as a form of government. After the eruption of the Russian Revolution in 1917, however, debate shifted to the terrain defined by Marxism and communism, and in particular, the French Revolution as the exemplar of bourgeois revolution.

3.1 The Question of Marxism

Within France, Marxist historians of the French Revolution enjoyed the advantage of having captured the chair of the history of the French Revolution at the Sorbonne (University of Paris). For most of the twentieth century, the chair was occupied by a succession of talented socialists and communists. They pioneered the study of social tensions and social classes, of 'history from below.' The most remarkable of them was Georges Lefebvre, who wrote wide-ranging studies of the peasantry during the Revolution. Lefebvre's combination of narrative skills and analytical studies gave him an unrivaled reputation; even those who objected to his Marxist views regarded him as the most accomplished academic historian of the Revolution.

The Right did not have the same initial foothold within the French university system. Hippolyte Taine, for example, wrote his four volumes on the Old Regime and the French Revolution as a general literary and cultural critic, not as a professional historian. Taine's views remained fundamental to the Right-wing criticism of the Revolution. He had nothing but contempt for 'the people' so beloved by Michelet. He frequently likened the lower classes to beasts or insects; at worst, they were a crazed, frenzied rabble. Augustin Cochin toned down this virulence, arguing that the lower classes were manipulated by professional politicians who came to power through the activities of secret societies such as Masonic lodges. The Right took up some of the claims of the Left—that the French Revolution foreshadowed the Russian Revolution of 1917—but gave them a negative rather than a positive spin.

Before the 1970s, the Marxist Left enjoyed the dominant position in French academic debates, in part because of its strength within the university, and in part because the Right was associated with anti-democratic politics. The balance began to shift, however, as critics of the Revolution finally accepted the republican form of government. A more moderate republican Right got support from a growing body of work produced in the Anglophone countries, especially the United States and the United Kingdom, which attacked the central Marxist thesis that the French Revolution was a bourgeois revolution. Critics argued that the capitalist bourgeoisie had not made the Revolution (first aristocrats and then middle-class officials and lawyers had led it, not manufacturers);

that the Revolution had not overthrown feudalism (because feudalism no longer existed); and that the Revolution did not pave the way for capitalism (since industrialization did not take place until the 1840s).

François Furet, a former communist, brought together the new Anglophone work in a frontal assault on the Marxist interpretation. This 'revolutionary catechism,' this 'Lenino-populist vulgate,' he argued, fundamentally distorted the history of the French Revolution by depicting it as a class struggle that anticipated communism. Furet claimed as his ancestors Tocqueville and Cochin; he did not reject democracy, but rather wanted to understand why democracy in the French Revolution developed into Terror or totalitarianism. The social underpinnings of the French Revolution could never explain this skidding out of control into Terror, he maintained; revolution was an autonomous political and ideological movement, not the expression of a social class (Furet 1978). By 1989, the year of the bicentennial of the French Revolution, Furet's version of the French Revolution had become the new orthodoxy. The political balance had shifted, not only in France but almost everywhere else as well. The French Revolution no longer seemed a communist preserve.

3.2 Current and Future Directions

Since 1989, as the polemics over Marxism have subsided, the work of historians from other countries has attracted more attention. American, British, German, and Italian historians now offer alternative views of their own. No one school, whether Left-wing or Right-wing, French or Anglophone, currently dominates the historiography of the French Revolution. The shift of political interest away from Marxism has opened up new areas of research, including gender relations, definitions of citizenship, and the role of the colonies. These new perspectives reflect concerns with human rights and globalization.

Since the French Revolution began by enshrining the notion of 'the rights of man,' it has been a prime site for work on the dilemmas of universalism. Feminists have pinpointed the Enlightenment's masculinist bias as the chief cause of the failure of the revolutionaries to include women in their promise of citizenship; Enlightenment writers, especially Montesquieu and Rousseau, used nature and reason to justify women's exclusion from the public sphere, and the revolutionaries simply drew the logical conclusion from these assumptions. Other feminists, however, have insisted that the Revolution opened up new opportunities for women, legalizing divorce, assuring equal inheritance to girls, and affording women the opportunity to establish political clubs. Unlike women, men who were Protestants, Jews, free blacks, mulattos, or even slaves gained political rights during

the French Revolution. Yet even these enfranchisements revealed telling fault lines in the revolutionary conception of citizenship; the National Assembly, for instance, insisted that the Jews must give up their particular identity and corporate privileges as Jews in order to become French citizens. Since discussions of free black and slave rights inevitably involved the slave colonies, and since France's colony of Saint Domingue (San Domingo) was home to the first successful slave revolt in history beginning in 1791, the interest in citizenship has also stimulated study of France's slave colonies. Because the French Revolution put both revolution and modernity on the agenda, it continually offers rich possibilities for historical and theoretical debate.

See also: Bourgeoisie/Middle Classes, History of; Citizenship, Historical Development of; Democracy; Democratic Theory; Enlightenment; Freedom/Liberty: Impact on the Social Sciences; Ideology: History of the Concept; Nationalism: General; Political Representation; Representation: History of the Problem; Rousseau, Jean-Jacques (1712–78); Social Science, the Idea of; Socialism: Historical Aspects; Tocqueville, Alexis de (1805–59)

Bibliography

- Arendt H 1963 *On Revolution*. Viking Press, New York
- Barnave A 1988 *Introduction à la Révolution française. De la Révolution et de la Constitution*. Presses Universitaires de Grenoble, Grenoble, France
- Buonarotti P 1828 *Conspiration pour l'égalité dite de Babeuf suivie du procès auquel elle donna lieu, et des pièces justificatives*. A la Librairie Romantique, Brussels, 2 Vols.
- Burke E 1790 *Reflections on the Revolution in France, and on the Proceedings in Certain Societies in London Relative to that Event*, 3rd edn. J. Dodsley, London
- Furet F 1978 *Penser la Révolution française*. Gallimard, Paris
- Furet F 1986 *Marx et la Révolution française*. Flammarion, Paris
- Godineau D 1988 *Citoyennes tricoteuses: les femmes du peuple à Paris pendant la Révolution française*. Alinéa, Aix-en-Provence, France
- Hunt L A 1984 *Politics, Culture, and Class in the French Revolution*. University of California Press, Berkeley, CA
- James C L R 1938 *The Black Jacobins: Toussaint L'Ouverture and the San Domingo Revolution*. Secker & Warburg, London
- Landes J B 1988 *Women and the Public Sphere in the Age of the French Revolution*. Cornell University Press, Ithaca, NY
- Taine H 1885–88 *La Révolution*. Hachette, Paris, 3 Vols.
- Tocqueville A de 1856 *L'ancien régime et la Révolution*, 2nd edn. Michel Lévy frères, Paris
- Tocqueville A de 1977 *Oeuvres complètes*. In: Jardin A, Lesourd J-A (eds.) *Correspondance Tocqueville-Louis de Kergolay*. Gallimard, Paris, Vol. 13

L. Hunt

Frequentist Inference

1. Data, Models, and Inference

The starting point of a statistical analysis is a set of data, for example, of counts or measurements. One aim may be simply to study what these data have to tell us. If they consist of a set of real numbers we might want to see, for example, whether they are small or large; tightly concentrated or spread out, whether they are stable or tend to increase with time, etc. If they are points in the plane we can get an idea of the shape of this set of points, for example, whether they cluster about a line. The branch of statistics dealing with this kind of investigation used to be called descriptive statistics, but now goes by the name data analysis or, more precisely, exploratory data analysis (EDA), a term introduced by Tukey (see *Exploratory Data Analysis: Univariate Methods*). Instead we consider how we can quantify the conclusions or decisions drawn from an analysis. Frequentist inference requires that any quantifying measure be interpretable in terms of frequentist probability of events (see *Frequentist Interpretation of Probability*). That is, we assume the data are random quantities produced by some probability distribution and that something is known about this distribution. For example, if we have a set of n measurements of some quantity, we may assume that these measurements are independently and identically distributed. We may stop there or may go further and make some assumptions about this common distribution.

Statistical analyses based on such assumptions were common in the nineteenth century, and isolated instances can be found in the eighteenth century; but the first general framework was proposed by R. A. Fisher (1922) (see *Fisher, Ronald A (1890–1962)*). He states:

The object of statistical methods is the reduction of data. A quantity of data which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole.

This object is accomplished by constructing a hypothetical infinite population of which the actual data are regarded as constituting a random sample. The law of distribution of this hypothetical population is specified by relatively few parameters, which are sufficient to describe it exhaustively in respect of all quantities under discussion.

In modern terminology we might paraphrase this proposal by saying that we construct a mathematical model, according to which the data are produced by a probability distribution assumed to belong to some specified parametric family of distributions.

What are the statistical methods based on these models supposed to achieve? Fisher called the aim 'inductive inference' or 'inductive reasoning,' and described it variously as 'learning by experience' and

as ‘drawing inferences from the particular to the general, from consequences to causes’ or—using statistical language—‘from the sample to the population.’ ‘The purpose of inductive reasoning based on empirical observations,’ he wrote ‘is to improve our understanding of the systems from which these observations are drawn.’

This interpretation was criticized by Neyman (1961) who objected that ‘After a conscientious effort to find the exact meaning of this term [inductive reasoning] I came to the conclusion that, at least in the sense of Fisher, the term is empty.’ Neyman proposed instead that the purpose of statistical analysis is to serve as guide to appropriate action; and, in contrast to Fisher’s inductive reasoning, named it inductive behavior (see also *Neyman, Jerzy (1894–1981)*).

Neyman’s point of view was taken up by Wald who used it to construct a comprehensive framework for statistical decision making (Wald 1950). It involved three principal elements

(a) A family of probability distributions representing the various possible true situations.

(b) A set of actions from which it is the statistician’s task to choose the most appropriate one.

(c) A loss function which measures the loss resulting from any action taken for any of the possible true situations.

Wald’s formulation was castigated by Fisher as completely inappropriate for scientific work. ‘It is important,’ he wrote (Fisher 1973, p. 106) ‘that the scientific worker introduces no cost functions for faulty decisions. To do so would imply that the purposes to which new knowledge was to be put were known and capable of evaluation. As workers in Science we aim at methods of inference which shall be equally convincing to all freely reasoning minds, entirely independently of any intentions that might be furthered by utilizing the knowledge inferred.’

While it is difficult not to sympathize with Fisher’s stress on scientific aims as an alternative to immediate utilitarian use, his statements are distressingly vague. An effort to be more specific, which still seems rather vague, was made by Tukey (1960). For the outcome of a nonaction-oriented statistical analysis he proposed the term ‘conclusion.’ Concerning the difference he writes: ‘Conclusions are established with careful regard to evidence but without regard to consequences of specific actions in specific circumstances. Conclusions are withheld until adequate evidence has been accumulated.’

The distinction between Fisher’s point of view and that of Neyman and Wald, between conclusions and decisions, concerns the interpretation of the results of a statistical analysis. How those results are reached from the given model by means of frequentist inference forms the subject of this article. (For the frequentist meaning of probability see *Frequentist Interpretation of Probability*.) It includes the derivation of appropriate procedures, the comparison of different pro-

cedures, investigating the sensitivity of a procedure to departures from the assumed model, and the checking of the suitability of the model. Another topic of interest, considered in Sect. 3.6, is the selection of an appropriate model.

There is an alternative approach to inference which can embody both the decision theory aspects of the Neyman–Wald approach and Fisher’s views on learning by experience. This is the Bayes approach which requires not only the frequentist’s specification of a family of possible distributions for the observation, but also that the distribution of the observation has itself been drawn from the specified model according to a known probability mechanism. This is coupled with the subjective point of view of probability in which the chance of an event represents a quantitative measure of the observers’ odds on the occurrence or nonoccurrence of the event. This point of view, though attractive, puts a burden on the observer to specify his or her state of mind with possibly unreasonable precision and, of course, leaves open the issue of the generalizability of the conclusions by scientists at large (see also *Bayesian Statistics*).

The problems of inference have been the subject of much discourse among philosophers as well as scientists whose primary interest was not statistics. We refer to Keynes (1921), von Mises (1928) and Jeffreys (1939) among others.

A problem in discussing frequentist inference is that, while distinctions such as those between Fisher, Neyman, and Wald are required in order to formulate the goals of theory, they are artificial in practice. Even in scientific inquiry the cost of initiating extensive further investigation on the basis of weak evidence has to be kept in mind. On the other hand, the most interesting consequences may not be foreseen and the costs of actions difficult to quantify. Is foisting a slightly better-than-average drug with serious side effects on most patients better than doing nothing? What if the drug turns out to have major benefits for a small group of patients not represented in the original clinical trial of the drug? A chemist, E. B. Wilson (1952), after considering the work of the thinkers we mention, pleads eloquently, ‘There is a great need for further work on the subject of scientific inference. To be fruitful it should be carried out by critical original minds who are not only well-versed in philosophy but also familiar with the way scientists actually work (and not just with the way some of them say they work).’

Wilson concludes pessimistically: ‘Unfortunately the practical nonexistence of such people almost suggests that the qualities of mind required by a good philosopher and those needed by a working scientist are incompatible.’

In what follows we

(a) begin, for illustrative purposes, by describing a few simple models that have arisen from a frequentist point of view (Sect. 2), and

(b) discuss by example the principal frequentist inference procedures (hypothesis testing, point estimation, confidence regions, prediction, and model selection) (Sect. 3).

We use the term frequentist inference if all probabilities involved (such as significance level, power, and coefficient) are interpreted as frequentist probabilities in the sense of the article (*Frequentist Interpretation of Probability*). In particular, the assertion that the value of such a probability is p then means that in a large number of independent cases with probability p of some outcome, we expect the outcome to occur with a frequency close to p . In this connection it is important to realize that there is no need for these situations to be at all like each other, as long as the probability p is the same in each (see Neyman 1937).

2. Some Simple Models

Frequentist models are characterized by two principal features.

(a) The variability of the data. If the observations were repeated, different values would be obtained. This variability is represented in the model by postulating that the data are generated by a probability distribution.

(b) The unknown aspect of the situation which the statistical investigation is to elucidate. This unknown feature is represented in the model by the fact that the probability distribution of the observations is assumed to be only partially known.

These two features are seen clearly in the following model, which was mentioned briefly in the preceding section.

Example 1. (Error-measurement model) If $X_1, \dots, X_n$ denote n measurements of an unknown quantity θ , this model can be written as

$$X_i = \theta + \varepsilon_i. \quad (1)$$

Here θ represents the unknown quantity we wish to estimate or test. The errors ε_i are the source of the variability of the observations. This frequency model assumes that, in a replication of the situation, the unknown value of θ would be the same but that the ε 's would take on different values. The most common assumption concerning their variability is that they are independent random variables with a common distribution with mean zero. This distribution may be assumed to be known. More usually it too embodies some unknown features; for example, its form may be known but it may contain an unknown scale parameter.

Simpson and others toward the middle of the eighteenth century considered various possible forms for this distribution; but by the early nineteenth

century it was agreed generally that the most suitable distribution of such measurement errors, in most cases, was the normal. The basis for this belief was the hypothesis of 'elementary errors,' which assumes that observational errors are the sum of a large number of small, elementary errors; and, by the Central Limit Theorem, are therefore approximately normally distributed.

If the ε 's in (1) are assumed to be independently normally distributed with the mean 0 and variance σ^2 , the model (1) is equivalent to assuming that

$$X_1, \dots, X_n \text{ are independent } N(\theta, \sigma^2). \quad (2)$$

Example 2 (Normal one-sample model) Model (2) arises also in contexts quite different from the measurement situation described above. Suppose X is some numerical characteristic of a person such as height, weight, blood pressure, intelligence, etc., or of an animal, plant or manufactured product. However, this time we are not taking several observations X_i on the same subject but one observation each on different subjects drawn at random from a population which, in the model, we shall assume to be infinitely large. In this setting θ is the mean value of X in the population, and (2) is called the normal one-sample model. (The measurement situation can be viewed as a special case if one considers the n measurements actually taken as a sample from an essentially infinite number of measurements that could be taken.)

In these applications the assumption of normality is frequently made even though it is often not suitable and each situation ought to be considered in its own right. Often normality can be brought closer by applying it not to the observations X_i themselves but to some transformation $T(X_i)$. In other cases the normal model (2) is replaced by the more general model

$$X_1, \dots, X_n \text{ are independently distributed} \\ \text{according to } F(x - \theta) \quad (3)$$

where F has mean zero and otherwise is arbitrary, or where it is assumed to be an arbitrary distribution that is symmetric about zero. Some smoothness conditions may also be imposed; for example, that F has a density f .

Example 3. (Linear Model) Large areas of statistical methodology are based on the following extension of (1), due to Gauss (1809) and known as the linear model

$$X_i = \sum_{j=1}^p \beta_j z_{ij} + \varepsilon_i \quad (i = 1, \dots, n) \quad (4)$$

where the z 's are known constants (sometimes called explanatory variables) and the β 's unknown parameters. The ε 's are the observational errors as before and are assumed to have expectation zero.

Examples of (4) abound. Here are a few.

(a) The measurement model (1) is the special case of (4) with $p = 1$, $z_{i2} = 1$ and $\beta_1 = \theta$.

(b) The two-sample problem with $n_1 + n_2$ independent variables $X_1, \dots, X_n: N(\theta_1, \sigma^2)$ and $X_{n_1+1}, \dots, X_{n_1+n_2}: N(\theta_2, \sigma^2)$ corresponds to $p = 2$,

$$z_{11} = \dots = z_{1n_1} = 1, \quad z_{1n_1+1} = \dots = z_{1n_1+n_2} = 0, \\ \beta_1 = \theta_1 \quad (5)$$

and

$$z_{21} = \dots = z_{2n_1} = 0, \quad z_{2n_1+1} = \dots = z_{2n_1+n_2} = 1, \\ \beta_2 = \theta_2 \quad (6)$$

(c) The k -sample model is defined analogously. The two-way layout without interactions assumes that the variables X_{ijk} ($k = 1, \dots, n_{ij}; i = 1, \dots, a; j = 1, \dots, b$) are independent normal with common variance σ^2 and with means

$$E(X_{ijk}) = \mu + \alpha_i + \beta_j (\Sigma \alpha_i = \Sigma \beta_j = 0). \quad (7)$$

The k -sample model and the two-way layout are the simplest cases of analysis of variance models.

Another important class of examples of (4) are models for regression. The simplest case is

(d) Simple linear regression, given by

$$X_i = \alpha + \beta t_i + \varepsilon_i \quad (8)$$

where the t 's are known constants, and α and β unknown parameters.

This is a special case of

(e) Polynomial regression, with

$$X_i = f(t_i) + \varepsilon_i \quad (9)$$

where f is a polynomial (for example, quadratic) of known degree and with unknown coefficients.

Other functions f , of course, are also possible.

A modification of (4) that is sometimes suitable is to have some of the z 's random. For example (7) in some situations is replaced by

$$X_{ijk} = \mu + A_i + \beta_j + \varepsilon_{ijk} (\Sigma \beta_j = 0) \quad (10)$$

where the A 's are unobservable random variables assumed to be independent $N(0, \sigma_A^2)$ and independent of ε . For more, see *Linear Hypothesis*.

When the distribution of the observational errors the ε 's in (1) and (4) is specified, the models considered so far are parametric, that is, they can be smoothly parameterized by Euclidean labels. Models such as (3) with the distributional form of errors unspecified up to

symmetry are semiparametric. Nonparametric models are ones in which as little is assumed as possible.

Example 4. (Nonparametric regression) This model is suitable in situations where, on each member of a sample of size n from some population, we observe not only a response Y but also observe characteristics (covariates) $\mathbf{Z} = (Z_1, \dots, Z_k)$. For example, Y might be the income of a randomly sampled individual, Z_1 might be age, Z_2 educational level, etc. Of principal interest here is the average relation between Y and $\mathbf{Z}$

$$E(Y|Z_1 = z_1, \dots, Z_k = z_k) \equiv m(z_1, \dots, z_k). \quad (11)$$

If we assume nothing about the form of m , then subject to the assumptions that $E|Y| < \infty$ and that $(Y_1, \mathbf{Z}_1), \dots, (Y_n, \mathbf{Z}_n)$ are independent and identically distributed, this model is completely general since we can always write

$$Y_i = m(Z_{i1}, \dots, Z_{ik}) + \varepsilon_i \quad (12)$$

where $E(\varepsilon_i|Z_i) = 0$. Note that if we assume

$$m(z_1, \dots, z_k) = \sum_{j=1}^k \beta_j z_j,$$

then, given $\mathbf{Z}_i = z_i$ ($i = 1, \dots, n$) we are back in the linear model (4).

In order to identify m in (12) we need to make some assumptions, for instance that m is continuous.

An important method for constructing more complex models from simpler ones is that of hierarchical models. As exemplified in (10), such models often contain unobservable random variables which are themselves of interest.

From this point of view Bayes models can be considered as two-stage hierarchical models in which the first stage is a frequentist model involving unobservable parameters for which the Bayes model specifies a distribution as a second stage. If the frequentist model is itself hierarchical, a last stage is added when the model is Bayesianized. To this difference—that the unknown parameters of the frequentist model become unobservable random variables in the corresponding Bayes model of course, is added the difference in interpretation of the probabilities involved as frequentist or subjective.

3. Inference Methods

As indicated in Sect. 1 linked closely with the frequentist paradigm is that of decision theory. The principal types of problems considered in that framework are

- (a) hypothesis testing,
- (b) point estimation,
- (c) confidence regions,

- (d) prediction,
- (e) model selection.

We shall discuss these in the above order.

3.1 Hypothesis Testing

Wald's decision theory grew out of the Neyman–Pearson theory of hypothesis testing (Neyman and Pearson 1933) which the authors formulated as choosing between two decisions, whether a prespecified statement (the hypothesis H) about the distribution P generating the data is or is not correct. If $\theta \in \Theta$ parameterizes the model, this is equivalent to determining whether $H: \theta \in \Theta_0$ or the alternative $K: \theta \in \Theta_1$ is true.

This formulation is satisfactory in many practical contexts such as occur, for example, in medicine, agriculture, or industry. For instance, in Example 3(b) the two samples may correspond to control and treatment groups of patients. In Example 3(b) the assumption of normality is made, and implicitly that of an additive treatment effect measured by the difference $\theta_2 - \theta_1$. The hypothesis to be tested is usually formulated as $H: \theta_2 \leq \theta_1$; i.e., that the treatment has no beneficial effect. The alternative $K: \theta_2 > \theta_1$ claims the existence of a beneficial effect.

The test for this problem is a test function δ which takes on the values $\delta(x) = 0$ or $\delta(x) = 1$ for the sample points x for which we decide to accept or reject H . The performance of such a test is measured by the probabilities of

- (a) Type I errors—rejecting H when H is true, and
- (b) Type II errors—accepting H when H is false, that is, by the probabilities

$$\begin{aligned} P_\theta[\delta(x) = 1] &\text{ when } \theta \in \Theta_0 \quad \text{and} \\ P_\theta[\delta(x) = 0] &\text{ when } \theta \in \Theta_1. \end{aligned} \quad (13)$$

The test δ is required to satisfy the condition

$$P_\theta(\text{Type I error}) \leq \alpha \quad \text{for all } \theta \in \Theta_0, \quad (14)$$

where the level α of the test is preassigned (usually as 0.05 or 0.01) as the maximum probability of false rejection to be tolerated. Subject to (14) we look for tests δ which make P_θ (Type II error) as small as possible.

This formulation is meaningful if it is possible to distinguish sharply between hypothesis and alternatives, between those θ -values for which rejection of H is clearly desirable or undesirable. In Example 3(b), for instance, we can specify that only beneficial effects exceeding $\varepsilon\sigma$ (for some given $\varepsilon > 0$) are of interest. Such a distinction is often possible in the contexts we have mentioned.

The Neyman–Pearson formulation of hypothesis testing in terms of the two types of error works well when we have clear definitions of both the hypothesis and the alternatives. In scientific contexts it is often the

case that the hypothesis is clear but the alternatives are vague. It is then the unlikelihood of observations that would be surprising if the hypothesis were true (but less so if the vaguely entertained alternatives were valid) which takes on particular significance. It is still necessary to define a test statistic T and to suppose the hypothesis is rejected when its values are sufficiently large, since such values would be surprising if H were true but not if H were false. Then the probability of observing values greater than or equal to the observed value of T is the p -value or significance probability of the hypothesis. If it is sufficiently small, it draws attention to anomalies not expected under H . The calculation of such probabilities is central to science. However, whether such attention should be followed by action as assumed in the Neyman–Pearson theory depends on extraneous considerations.

The Bayesian approach to testing assigns prior probabilities to hypothesis and alternative and calculates the ratio of posterior probabilities, the Bayes factor, as a measure of evidence in favor of or against the hypothesis rather than the significant probability. This approach is very attractive in providing exactly what everyone wants: the probability of the hypothesis being true, given the data. However, this probability depends on the prior probability distribution assigned to the parameters; and these priors can overwhelm strong opposite data evidence. Disparate priors can result in conflicting conclusions based on the same data by analysts holding different prior opinions (see also *Hypothesis Testing in Statistics*).

3.2 Point Estimation

The decision-oriented point of view of Neyman and Pearson led to the full-fledged decision theory of Wald, based on assigning losses to the consequences of inappropriate decisions. Perhaps the most successful example of this approach is point estimation (for which the idea of loss functions had already been discussed by Gauss and Laplace).

The prototypical estimation problem is that of estimating an unknown physical constant subject to measurement error discussed in Example 1. The loss function most commonly considered is squared error leading to its expected value, the mean squared error (MSE), as its risk. The standard assumption of independent normal errors with mean zero and common variance implies that the sample mean $\bar{X}$ is the appropriate estimator according to a number of different criteria.

(a) Minimax. This is a worst case analysis in which the maximum risk is minimized.

(b) Uniform minimum variance unbiased. An estimator of θ is said to be unbiased if its expectation is equal to θ . For unbiased estimators the MSE is equal to the variance; and, in the present case, $\bar{X}$ minimizes the variance among all unbiased estimators.

(c) Maximum likelihood (ML). This is a non-decision-theoretic approach based on the likelihood the density of the data viewed as a function of the parameter. Fisher considered the likelihood of an observation as a measure of support for the different θ -values that might have generated it and proposed the θ -value maximizing it as an appropriate estimator of θ . When a uniform prior distribution can be placed on the parameter, the ML estimator is just the mode of the posterior distribution. However, when θ ranges over the infinite line (as is the case in the measurement model and many others), no uniform probability distribution for θ exists; and only an ‘improper Bayes’ interpretation is then possible.

The ML principle is popular for a number of reasons

- (a) general applicability,
- (b) invariance under reparameterization,
- (c) in cases such as Example 2, where meaning can be attached to ‘large sample sizes,’ the method is approximately optimal,
- (d) in many standard cases it leads to explicit solutions.

However, there are examples in which the resulting estimators are completely misleading, (LeCam 1990, Ferguson 1996). In particular, ML is either not applicable or can lead to very poor results in many non- and semiparametric models such as Example 4. For a general discussion of different approaches to estimation see, for example, Bickel and Doksum (2000), Lehmman and Casella (1998).

The assumption of joint normality of the data, for instance in Examples 1–3, is often seriously in doubt. In such cases robustness criteria are invoked and lead to alternatives to the linear procedures appropriate for normally distributed data. For example, the median rather than the mean (or some intermediate trimmed mean) might be used in Example 1, and minimum L_1 (least absolute values) rather than L_2 (least squares) estimators in Example 3. For an extended discussion of robustness see Hampel et al. (1986) and Staudte and Sheather (1990).

3.3 Bayes Estimation

Still another principle is that of Bayes estimation, i.e., minimizing the expected risk under a prior distribution π of the parameter θ . Since this is the same as minimizing the average risk, averaged with respect to the weight-function π , such estimators are of interest both from a Bayesian and a frequentist point of view. For estimating a real-valued parameter $g(\theta)$ with data X and squared error loss, the Bayes estimator is

$$E[g(\theta)|X] \quad (15)$$

the mean of the posterior distribution of $g(\theta)$ given X .

Bayes estimators are essentially never unbiased (Girshick and Savage 1951) since they naturally ‘pull’

the estimator toward one’s prior opinion, the prior mean of $g(\theta)$. For example, in the one-sample model of Example 2. If π is normal (μ, τ^2) , the Bayes estimator is

$$\left(\frac{n\bar{X} + \mu}{\sigma^2 + \tau^2} \right) / \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2} \right). \quad (16)$$

Frequentist and Bayes inference generally merge for large sample sizes since the data then wash out the influence of the prior distribution (provided the prior does not exclude parts of the sample space); see, for example, Blackwell and Dubins (1962).

In estimation even more can be said. If the prior distribution is locally uniform, Bayes estimators are asymptotically optimal in a frequentist sense by the Bernstein–von Mises theorem (see, for example, Lehmann and Casella 1998, Bickel and Doksum 2000). This is, however, a limit result; in practice, the effect of a strong prior often persists even for substantial sample sizes.

Both aspects can be illustrated by the Bayes estimator (16). For fixed μ , the difference of this estimator from $\bar{X}$ is of the order $1/n\tau^2$. For fixed τ^2 , it is accordingly of the order $1/n$, while the difference of both estimators from θ is of the order $1/\sqrt{n}$. On the other hand, the convergence of (16) can become arbitrarily slow if $\tau^2 \rightarrow 0$ as $n \rightarrow \infty$. The basic message is that optimal frequentist estimators are appropriate for Bayesian problems and vice versa.

For the sake of simplicity, we have here focused attention to the estimation of real-valued parameters. Of course, estimation problems arise also for multi-variate parameters (for instance, in Example 3) and for function-valued parameters such as the function m in Example 4 or in the estimation of an unknown density. Squared error loss is then replaced by summed or integrated MSEs. However, the univariate and multi-variate situations present unexpected differences. For example, the typically unique minimax estimator of the univariate case is replaced by an infinite set of minimax estimators; and among these the natural extension of the classical univariate estimator is no longer optimal (Stein 1956).

Point estimates by themselves are hardly ever sufficient. Typically one also requires an idea of the error committed in using the estimator; that is, an estimate of the MSE. This is again the problem of estimating a parameter, and there is little to add. However, interpretation of the estimated measure of error is often difficult. It is clarified by the concept of confidence estimation, which we take up next.

3.4 Confidence Regions

When estimating θ in the measurement error model of Example 1, it is customary to indicate the reliability of

the estimate $\bar{X}$ by attaching to it the error estimate $\pm S/\sqrt{n}$ with $S^2 = \Sigma(X_i - \bar{X})^2/(n-1)$ as a proxy for $\pm \sigma/\sqrt{n}$. In the case of normal errors with mean zero.

$$P_\theta \left[\bar{X} - \frac{\sigma}{\sqrt{n}} \leq \theta \leq \bar{X} + \frac{\sigma}{\sqrt{n}} \right] = 0.67 \quad \text{for all } \theta. \quad (17)$$

If σ is known, (17) provides intervals of fixed length (and with random midpoints) which, in many repetitions of the experiment of taking n measurements, will 'cover' the true θ 67 percent of the time whatever are the true values of θ and σ . The intervals (17) are confidence intervals for θ with confidence coefficient 0.67.

In practice, σ is usually unknown. We can then obtain confidence intervals for θ by replacing σ by its estimator S . If t_α is the $100(1-\alpha/2)$ percent point of the t -distribution with $n-1$ degrees of freedom, the intervals

$$\bar{X} - t_\alpha \frac{S}{\sqrt{n}} \leq \theta \leq \bar{X} + t_\alpha \frac{S}{\sqrt{n}} \quad (18)$$

cover the true θ with probability $1-\alpha$ so that the statement (18) is correct about $100(1-\alpha)$ percent of the time. The length of the intervals is no longer fixed but random, its length tending to increase as the accuracy of the observations (the inverse of which is measured by σ and estimated by S) decreases.

In general, in estimating a parameter θ taking values in an arbitrary space Θ , a $100(1-\alpha)$ percent confidence region for θ is a random subset $C(X)$ of Θ , depending on the data with the property that

$$P[\theta \in C(X)] \geq 1-\alpha \quad (19)$$

for all probability distribution P in the model. Here the region $C(X)$ can, for example, be an interval as above, an ellipse as in the Scheffé regions for the parameters of the linear model in Example 3, or a 'band' about an unknown distribution function.

Confidence regions can be viewed as simultaneously making statements about the acceptability of a family of hypotheses (see Lehmann 1986). In particular, they provide a measure of the acceptability of the alternatives. A 95 percent confidence interval for the treatment effect of Example 3(b), for instance, which includes zero, tells us not only that the null hypothesis is accepted at the 5 percent level but also specifies all alternatives to the null hypothesis which are accepted at that level. An interval which does not contain zero but to which zero is close suggests that although we have seen something that is surprising under the null hypothesis, the data are consistent with alternatives that do not differ materially from zero.

The interpretation of confidence regions is conceptually difficult because the probability statements refer not to any random variation of the parameter, which, although unknown, is considered fixed, but to that of

the data-dependent region. In contrast, a Bayesian analysis views θ as random and constructs credible regions $C^*(X)$ such that the posterior probability of θ falling into $C^*(X)$ is $\geq 1-\alpha$. This is just the kind of conclusion one would like to make; but, as in the case of testing, tends to be strongly influenced by assumptions concerning the prior.

Again, as in point estimation, Bayes regions derived on the basis of optimality criteria of size (length, volume, etc.) for large sample sizes agree approximately with frequentist confidence regions satisfying minimax size properties subject to fixed probability of coverage (see also *Estimation: Point and Interval*).

3.5 Prediction

This is the part of frequentist inference that is closest to the Bayesian approach and that fits least well into the decision-theoretic framework. Typically we are given a sample from a population of the form $(\mathbf{Z}_i, Y_i)$, $i = 1, \dots, n$ where $\mathbf{Z}_i$ is a vector of predictors. Using this sample (which provides information concerning the relationship of $\mathbf{Z}$ and Y) and a new $\mathbf{Z}_{n+1}$ we wish to predict the unobserved Y_{n+1} . The predictor of Y_{n+1} is a function

$$\delta(\mathbf{Z}; \mathbf{Z}_1, Y_1, \dots, \mathbf{Z}_n, Y_n) = \hat{\delta}(\mathbf{Z})$$

such that $\hat{\delta}(\mathbf{Z}_{n+1})$ is used to predict Y_{n+1} . In the continuous case (regression), the classical measure of loss resulting from incorrect prediction is squared error,

$$[Y_{n+1} - \hat{\delta}(\mathbf{Z}_{n+1})]^2.$$

On the other hand, in the case of categorical variables taking on values $\{1, \dots, k\}$ (classification), the standard loss function takes on the value 0 if $Y_{n+1} = \hat{\delta}(\mathbf{Z}_{n+1})$ and 1 otherwise. Prediction has the attractive feature that it allows nonparametric estimates of error since for any procedure we can compute its performance by applying it to the training sample on which it was built, for instance, $(1/n) \sum_{i=1}^n (\delta(\mathbf{Z}_i) - Y_i)^2$. This measure typically underestimates error, but techniques such as cross validation can be used to adjust it.

3.6 Model Selection

An important aspect of a frequentist analysis is the specification of a model. As an illustration, consider Example 3(e) where the regression function is assumed to be well approximated by a polynomial. Suppose the errors are normally distributed with zero mean and common variance. To complete the specification of the model, it is then necessary to decide on the degree d of the polynomial. With classical (least squares) prediction methods, the variance of the prediction error increases and its bias decreases with d . On a frequentist

basis the data can then be used to select d so as to obtain a desirable balance between these two aspects. Criteria for selecting d (and making analogous choices within a family of proposed models in other problems) have been developed by Akaike, Mallows, Schwarz, and others (for more details, see, for example, Linhart and Zucchini 1986):

Bayesian solutions to this problem differ in a number of ways, reflecting whether we assume that the true model belongs to the hypothesized class of models (e.g., is really a polynomial) or can merely be approximated arbitrarily closely by such models. For more on this topic see Shao (1997).

See also: Estimation: Point and Interval; Robustness in Statistics.

Bibliography

- Bickel P, Doksum K 2000 *Mathematical Statistics*, 2nd edn. Prentice Hall, New York, Vol. 1
- Blackwell D, Dubins L 1962 Merging of opinions with increasing information. *Annals of Mathematical Statistics* **33**: 882–6
- Ferguson T S 1996 *A Course in Large Sample Theory*. Chapman & Hall, London
- Fisher R A 1922 On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A* **222**: 309–68
- Fisher R A 1973 *Statistical Methods and Scientific Inference*, 3rd edn. Hafner Press, New York
- Gauss C F 1809 *Theoria Motus Corporum Celestium*. Perthes, Hamburg, Germany
- Girshick M A, Savage L J 1951 *Bayes and Minimax Estimates for Quadratic Loss Functions*. *Proceedings of the 2nd Berkeley Symposium of Mathematics, Statistics, and Probability*. University of California Press, Berkeley, CA
- Hampel F, Ronchetti E, Rousseeuw P, Stahel W 1986 *Robust Statistics*. Wiley, New York
- Jeffreys H 1939 *Theory of Probability*. Clarendon Press, Oxford, UK
- Keynes J M 1921 *A Treatise on Probability*. Macmillan, London
- LeCam L 1990 Maximum likelihood: An introduction. *International Statistical Review* **58**: 153–71
- Lehmann E L 1986 *Testing Statistical Hypotheses*, 2nd edn. Springer, New York
- Lehmann E L, Casella G 1998 *Theory of Point Estimation*, 2nd edn. Springer, New York
- Linhart H, Zucchini W 1986 *Model Selection*. Wiley, New York
- Neyman J 1937 Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society Series A* **236**: 333–80
- Neyman J 1961 Silver Jubilee of my dispute with Fisher. *Journal of the Operations Research Society of Japan* **3**: 145–54
- Neyman J, Pearson E S 1933 On the problem of the most efficient tests of statistical hypothesis. *Philosophical Transactions of the Royal Society Series A* **231**: 289–337
- Shao J 1997 An asymptotic theory for linear model selection (with discussion). *Statistica Sinica* **7**: 221–66
- Staudte R G, Sheather S J 1990 *Robust Estimation and Testing*. Wiley, New York
- Stein C 1956 Inadmissibility of the usual estimator for the mean of a multivariate distribution. *Proceedings of the 3rd Berkeley*

Symposium of Mathematics, Statistics, and Probability. University of California Press, Vol. 1, pp. 187–95

Tukey J W 1960 Conclusions versus decisions. *Technology* **2**: 423–33

von Mises R 1928 *Wahrscheinlichkeit, Statistik und Wahrheit*. Springer, Wien, Austria

Wald A 1950 *Statistical Decision Functions*. Wiley, New York

Wilson E B 1952 *An Introduction to Scientific Research*. McGraw Hill, New York

P. J. Bickel and E. L. Lehmann

Frequentist Interpretation of Probability

If the outcome of an event is observed in a large number of independent repetitions of the event under roughly the same conditions, then it is a fact of the real world that the frequency of the outcome stabilizes as the number of repetitions increases. If another long sequence of such events is observed, the frequency of the outcome will typically be approximately the same as it was in the first sequence.

Unfortunately, the real world is not very tidy. For this reason it was necessary in the above statement to insert several weasel words. The use of ‘roughly the same,’ ‘typically,’ ‘approximately,’ and ‘long sequence’ make it clear that the stability phenomenon being described cannot be stated very precisely. A much clearer statement is possible within a mathematical model of this phenomenon. This discovery is due to Jacob Bernoulli who raised the following question.

It is well known, Bernoulli says in Part IV of his book *Ars Conjectandi* (published posthumously in 1713), that the degree to which the frequency of an observed event varies about the probability of the event decreases as the number of events increases. He goes on to say that an important question that has never been asked before concerns the behavior of this variability as the number n of events increases indefinitely. He envisages two possibilities.

(a) As n gets larger and larger, the variability eventually shrinks to zero, so that for sufficiently large n the frequency will essentially pinpoint the probability p of the outcome.

(b) Alternatively, it is conceivable that there is a positive lower bound below which the variability can never fall so that p will always be surrounded by a cloud of uncertainty, no matter how large a number of events we observe.

Bernoulli then proceeds to prove the law of large numbers, which shows that it is (a) rather than (b) that pertains. More precisely, he proves that for any $a > 0$

$$\Pr \left(\left| \frac{X}{n} - p \right| < a \right) \rightarrow 1 \quad \text{as } n \rightarrow \infty \quad (1)$$

where X/n is the frequency under consideration.

It is easy to be misled into the belief that this theorem proves something about the behavior of frequencies in the real world. It does not. The result is only concerned with properties of the mathematical model. What it does show is that the behavior of frequencies in the model is mirrored in a way that is much neater and more precise, the very imprecise stability phenomenon stated in the first paragraph of this article.

In fact, a result for the model much more precise than (1) was obtained by De Moivre (1733). It gives the normal approximation

$$\Pr \left(\left| \frac{X}{n} - p \right| < c/\sqrt{n} \right) \rightarrow \int_{-c/\sqrt{pq}}^{c/\sqrt{pq}} \phi(x) dx \quad (2)$$

where ϕ denotes the standard normal density. This is again a theorem in the model. The approximate corresponding real-world phenomenon can be seen, for example, by observing a quincunx, a mechanical device using balls falling through 'random' paths to generate a histogram.

De Moivre's result was given a far reaching generalization by Laplace (1810) in the central limit theorem (CLT) concerning the behavior of the average $\bar{X}$ of n identically, independently distributed random variable $X_1, \dots, X_n$ with mean ξ and finite variance σ^2 . It shows that

$$\Pr \left(|\bar{X} - \xi| < c/\sqrt{n} \right) \rightarrow \int_{-c/\sigma}^{c/\sigma} \phi(x) dx. \quad (3)$$

This reduces to (2) when X takes on the values of 1 and 0 with probabilities p and q , respectively. The CLT formed the basis of most frequentist inference throughout the nineteenth century.

The first systematic discussion of the frequentist approach was given by Venn (1866), and an axiomatization based on frequencies in infinite random sequences (*Kollektives*) was attempted by von Mises (1928). Because of technical difficulties his concept of a random sequence was modified by Solomonoff (1964), Martin-Löf (1966), and Kolmogorov (1968), with the introduction of computational complexity. (An entirely different axiomatization based on events and their probabilities rather than random sequences was put forward by Kolmogorov in 1933, and has successfully served as a basis for both frequentist and subjective interpretations of probability). For more details on these different approaches see Barnett (1982).

The frequentist concept of probability described so far has met considerable criticism. One of the main objections is that it is not applicable to many situations to which one might want to apply probability assessments. To see this, consider the following three possibilities.

(a) An actual sequence of repetitions may be available; for example, a sequence of coin tosses or a sequence of independent measurements of the same quantity.

(b) A sequence of repetitions may be available in principle but not likely to be carried out in practice; for example, the polio experiment of 1954 involving a sample of over a million children.

(c) A unique event which by its very nature can never be replicated, such as the outcome of a particular historical event; for example, whether a particular president will survive an impeachment trial. The conditions of this experiment cannot be duplicated.

The frequentist concept of probability can be applied in cases (a) and (b) but not in the third situation. An alternative approach to probability which is applicable in all cases is the notion of probability as degree of belief; i.e., of a state of mind (for a discussion of this approach, see Robert (1994)). The inference methods based on these two interpretations of the meaning of probability are called frequentist and Bayesian, respectively.

Although frequentist probability is considered objective, it has the following subjective feature. Its impact on a particular person will differ from one person to another. One patient facing a surgical procedure with a 1 percent mortality rate will consider this a dire prospect and emphasize the possibility of a fatal outcome. Another will shrug it off as so rare as not to be worth worrying about.

There exists a class of situations in which both approaches will lead to the same probability assessment. Suppose there is complete symmetry between the various outcomes; for example, in random sampling which is performed so that the drawing favors no sample over another. Then we expect the frequencies of the various outcomes to be roughly the same and will also, in our beliefs, assign the same probability to each of them.

Let us now turn to a second criticism of frequentist probability. This concerns the difficulty of specifying what is meant by a repetition in the first sentence of this section. Consider once more the surgical procedure with 1 percent fatalities. This figure may represent the experience of thousands of cases, with the operation performed by different surgeons in different hospitals and—of course—on different patients. The rate of fatalities may vary from one hospital or surgeon to another and may, in particular, vary drastically with the condition, for example, the age and general health, of the patient.

Suppose a young woman requires this operation although her general health is very good. The frequency of a fatal outcome with patients sharing these characteristics may be much lower, and the 1 percent figure in that sense would be quite misleading for her. And yet she might be considered to have been obtained under 'roughly the same conditions,' namely to be drawn at random from the total population of persons

requiring this surgery. To obtain the most useful figures one should identify the most important variables, classify the cases accordingly (for example, young, middle-aged, old; male, female; etc.) and then provide the frequency for each class. They will, of course, be meaningful only for the classes which contain a reasonable number of cases.

1. A Terminology Note

A source of much confusion in the discussion of probability is the fact that 'probability' is used both as a mathematical term, i.e., as a concept in the mathematical model and also in everyday language when talking about real events. When reading about probability, it is important to be aware of these two meanings and to keep them distinct.

See also: Frequentist Inference; Probability: Formal; Probability: Interpretations; Statistical Methods, History of: Post-1900; Statistical Methods, History of: Pre-1900

Bibliography

- Barnett V 1982 *Comparative Statistical Inference*, 2nd edn. Wiley, New York
- De Moivre A 1733 1738, 1756. In: *Doctrine of Chance*, Millar, London, Reprinted by Chelsea Publishers, New York (1967), pp. 235–43 (1738), 243–54 (1754) 1878–1912
- Kolmogorov A 1933 *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Ergebnisse der Mathematik, Vol. 2. Springer, Berlin
- Kolmogorov A 1968 Logical basis for information theory and probability theory. *IEEE Transactions on Information Theory* **14**: 662–4
- Laplace P S *Oeuvres complètes de Laplace*. Gauthier-Villars, Paris
- Martin-Löf P 1966 The definition of random sequences. *Information and Control* **9**: 602–19
- Robert C P 1994 *The Bayesian Choice*. Springer, New York
- Solomonoff R J 1964 The formal theory of inductive inference. *Information and Control* **7**: 224–54
- Venn J 1866 *The Logic of Change*. Reprinted by Chelsea Publishers, New York, (1967)
- von Mises R 1957 *Probability, Statistics and Truth*. Macmillan, New York

P. J. Bickel and E. L. Lehmann

Freud, Sigmund (1856–1939)

1. Family Background and Early Life

Sigmund Freud, neuroscientist turned founder of psychoanalysis, was born on May 6, 1856, in Freiberg, Moravia, now part of the Czech Republic. His father,

Jakob (1815–96), a wool merchant, came from Galicia, and married Amalie (1835–1930, nee Nathansohn), 20 years younger, when he was 40. Jakob had two sons by a previous marriage, Emanuel and Philipp. Jakob was warm and affectionate, with a strong sense of humor, and Freud evidently admired and respected him. Amalie was a lively and attractive woman who was proud of Sigmund, her first-born.

The family multiplied quickly: five daughters and two more sons were born within 10 years. The first of these, Julius, died when 8 months old and Sigmund was aged 19 months. Both parents were Jews, with no religious affiliations, though a devout Catholic Nanny took Sigmund to Church. She left when he was two and a half. A nephew, John, a year older than Freud was very close to him: the two fought and played together. That childhood ambivalence, Freud reflected years later, must have had a profound effect on his character and on his ability to defend himself (Freud 1900).

Financially, times were hard. In 1859 the family moved to Leipzig, and in the following year to Vienna. Freud missed the beauty of the countryside round Freiberg. Although the children were well looked after, poverty was not assuaged by rapid family growth. By 1875 family fortunes improved: Amalie's family were helping, making possible a move to a larger house. Freud had a room of his own that served as a study.

Freud's memories of his childhood as revealed in his writings are without early detail, but one unforgettable recollection was of urinating in his parents' bedroom while they were present. Reprimanding him, his father said bluntly: 'The boy will come to nothing!' Freud's wounded ambitions were reflected recurrently in dreams in which he listed his achievements as if to say: 'You see! I have come to something!' (Freud 1900; see also Shengold 1993).

2. Education and Early Interests

Freud's schooling began with his mother, until his father took over before sending him to a private school. Freud learned rapidly, and was reading Shakespeare (whom he loved all his life) from the age of eight years. When nine years of age, he won a place at the Sperl Gymnasium, where he was a distinguished pupil, passing out, at 17, with the distinction '*summa cum laude*.' He became an accomplished linguist, with a firm foundation in Latin and Greek, a sound knowledge of Hebrew, fluency in English and French, and good Italian. He corresponded with an adolescent friend in Spanish. He became one of the greatest stylists in the German language (Kaufmann 1980). He studied at every opportunity. He loved the Arts, but took an early dislike to music, which he never felt able to appreciate. He was fond of walking, swimming, and skating. He liked travel, and made his first visit to England at the age of 19.

3. Early Career and Scientific Outlook

Freud (1925) said that curiosity and human concerns motivated his interest in medicine, but it was 'hearing Goethe's beautiful essay on Nature read aloud...that decided me to become a medical student.' He was attracted strongly by Darwin—a necessary step towards his own evolutionary approach to the history of the individual's mental development. His neurological work was affected profoundly by the evolutionary approach of John Hughlings Jackson, the founder of modern neurology. Furthermore, Freud's admiration of Goethe as scientist as well as writer was life long. Goethe's belief in the importance of observation and reflection was profound, and his meticulous examination of a single sheep's skull in 1790 led to his deduction that the vertebrate's skull developed from vertebrae. This type of detailed study—applied by Freud to the individual case—is the essence of poetic (or romantic) science that stands in contrast with (though not in opposition to) Newtonian (or classical) science. (Goethe's unfortunate opposition to Newton's theory of color does not vitiate this point.) In psychology, the corresponding distinction is between the ideographic (romantic) approach and the nomothetic (classical). Both methods have drawbacks and advantages; for Freud's psychology, in particular, the nomothetic approach has serious limitations. The distinctions have been discussed extensively by Kaufmann (1980) and by the founder of modern neuropsychology, Luria (1979). It is impossible fully to appreciate Freud's work outside this context.

Freud began his medical training in October 1873 and qualified MD in March 1881. The years were crowded: Freud dealt with the tedium of a year's compulsory military service by translating a book by John Stewart Mill (Jones 1953). He was influenced by many illustrious figures, none more important than Ernst Brücke (1819–92) head of the Institute of Physiology. Freud, as a research scholar, studied the histology of nerve cells, and his work led to a number of significant conclusions. After qualification, Freud continued working at the Institute, where his love for research had full scope. He had already found 'rest and full satisfaction' there (Freud 1925).

4. Neuroscience

But in 1882 Freud was still poor and, reluctantly, accepted Brücke's advice to practice medicine. He left the laboratory to start work in the General Hospital. He spent 6 months under Theodor Meynert (1833–1892), the Professor of Psychiatry, whose work and personality Freud greatly admired. His research at the Institute of Cerebral Anatomy led to important publications. He began to study nervous diseases, and quickly established a reputation as an outstanding diagnostician. He was still urgently short of money: he

was by now engaged to Martha Bernays (1861–1951), and the need to earn enough to keep a wife and family was compelling.

In the spring of 1885 Freud was appointed Lecturer (*Dozent*) in Neuropathology. Brücke's influence resulted in the award of a valuable Traveling Bursary, and Freud fulfilled an old ambition to visit Paris and learn from Charcot at the Salpêtrière. His offer to translate a new volume of Charcot's lectures into German was accepted by the great man, and Freud was soon admitted to Charcot's inner circle. Charcot's lectures and demonstrations were outstanding, but Freud was astonished by his investigations into hysteria. Charcot demonstrated that hysterical symptoms were genuine; that hysteria was common in men as well as women; and that hysterical paralyses produced by hypnosis were identical with spontaneous attacks. Charcot met skepticism with patience and firmness: once, faced with an objection made on theoretical grounds, he replied: 'Theory is all very well, but it doesn't stop facts from existing.' Freud never forgot that remark.

The Paris visit changed the direction of Freud's career. Charcot had broadened his grasp of neurology, but the neuroses had captured his imagination. He left Paris at the end of February 1886, intent on further investigations of hysteria. But before he returned to Vienna he stayed for some weeks in Berlin to study nervous disorders in children, and this important detour led in time to the publication of a number of monographs. Freud's report on his visits, to his sponsors in Vienna, is a model of clarity and concise exposition (Freud 1956).

Indeed, Freud's interest in neuropathology by no means stopped abruptly. Important papers lay ahead: observations on hemianopsia in two very young children (Freud 1888) and his monograph on *Aphasia* (Freud 1891b) occupy a permanent place in the history of aphasiology. Freud had already abandoned the strict localisation theory of brain functioning—to which Meynert strongly adhered—in favor of the hierarchical model, with its concomitant theory of evolution and dissolution of the nervous system, advanced by Hughlings Jackson (e.g., 1884). Freud recognized the importance of these major advances, which were otherwise scarcely appreciated for another quarter of a century. With his young friend Oscar Rie he wrote a work on the unilateral cerebral palsies of children—a book still of repute in neuroscience (Freud 1891a). Freud's overall contributions to the field were massive, and he published around 200 papers.

On his return to Vienna, however, Freud was trying to build up a practice to support a family: he married Martha in September 1886, starting a life long partnership of unusual harmony and mutual support. Both were devoted to their six children: in order, Mathilde, Martin, Oliver, Ernst, Sophie, and Anna—the only one to follow Freud into psychoanalysis. The first 5 years of marriage covered the closing phase of

neurological work, and saw the growth of a friendship with the distinguished physician Joseph Breuer (1842–1925), with profound effects on Freud's career.

5. *The Birth of Psychoanalysis*

Between 1880 and 1882 Breuer had treated the now famous Frau Anna O. for severe hysterical symptoms, including multiple limb paralyses, disturbances of sight and speech, contractures, an intense cough, and anorexia. Accounts of the illness, its treatment, outcome, and significance for the foundations of psychoanalysis, and its effect on the Breuer–Freud friendship vary (e.g., Breuer 1882, Breuer and Freud 1895, Freud 1914, 1925, 1932, Jones 1953, Gay 1988, Ellenberger 1972, Hirschmüller 1989), but, at the risk of distortion by brevity, the present summary seems safe. Anna O. suffered from states of autohypnosis alternating with near-normality. When, under self-hypnosis, she recounted the history of one of her symptoms, recalling the details of its first appearance, Breuer was astonished to find that it disappeared. Under induced hypnosis other symptoms were treated in similar fashion, and the method was christened the 'talking cure.' Breuer told Freud about it in 1882, and during the next two years they discussed it repeatedly. But, in Paris, Charcot showed little interest in the story, and Freud allowed the matter to lapse.

Back in Vienna, Freud treated a number of hysterical patients by hypnotic suggestion. But 'recovery' was too often transient, and Freud's interest in Breuer's treatment of Anna O. revived. When her first symptom appeared, the patient was nursing her dying father, and her recall resulted in 'abreaction' (the release of pent-up emotion attached to the event). Adopting this 'cathartic method' over several years, Freud repeatedly confirmed Breuer's findings. The links with traumata were hitherto unconscious.

A preliminary communication by Breuer and Freud (1893) preceded the *Studies on Hysteria* (Freud 1895d). Breuer contributed a theoretical chapter and an account of Anna O; Freud, four case histories. Differences of opinion became a strain in 1894, and the collaboration ceased in 1896. Freud valued Breuer's achievement, but was thinking on different lines. Breuer believed that pathological processes arose during 'hypnotic states' (e.g., Anna O's autohypnosis). For Freud, this begged the question of how 'hypnotic states' arose. He considered that psychoneurosis involved psychic defences against unacceptable internal states of 'emotional excitation,' habitually sexual in kind. The actual neuroses, however, were states of anxiety arising from unrelieved sexual excitation linked with abstinence or *coitus interruptus*. Breuer could not accept these views; he had failed to find any sexual component in Anna O's disturbance. But Breuer had never recognized Anna O's strong attachment to him ('transference') as sexual; alarmed by its intensity, he brought the treatment to an end. Anna

O. responded to this loss with a phantom pregnancy (Freud 1932). Breuer continued to see mentally ill patients, but never again used the cathartic method.

In time, Freud abandoned the use of hypnosis altogether. The power of transference (from the patient's unrecognized re-experience, in the treatment process, of infantile relationships) meant that excellent clinical progress could still be vitiated by disturbances in the doctor/patient relationship. That recognition profoundly affected the future development of psychoanalysis.

For a time, Freud used encouragement, persuasion, and similar measures as aids to treatment, but increasingly found them unnecessary. They were replaced slowly by the developing technique of free association, which took over completely by 1897–8. The patient lay on a couch (as before) with the analyst sitting behind him, but was now expected to try to say everything that entered his mind, irrespective of apparent irrelevance, distastefulness, or indiscretion. The analyst listened impartially, without giving more weight to any one utterance than he did to another. This major shift in clinical method had surprising results. Conflicting mental forces, undetected by hypnosis and related devices, were now revealed. Memories of striking internal and external mental experiences in early life had been denied access to consciousness because they had been 'alarming or painful or shameful by the standards of the subject's personality' (Freud 1925). Memory was actively opposed by the force of repression, experienced by the clinician as resistance to unwelcome material. Forces striving to break through this barrier stemmed from the drives, founded substantially in infantile sexuality.

6. *Psychoanalytic Career*

Until the spring of 1897 (Jones 1953) Freud believed his female adult patients whose associations pointed to seduction in childhood, usually by the father, as the traumatic root of their illness. But what at first presented as memory was found increasingly to be fantasy. Seduction in childhood undoubtedly occurred, but could no longer be invoked as the root of all neurosis. As it stood, Freud's theory was wrong. This devastating conclusion nonetheless pointed to the importance of fantasies in mental life, and to the realization that fantasy that is dynamically unconscious, that is, inaccessible to introspection, had all the force of reality since it could not be tested in the light of objective experience. A second, fateful, event, did more to foster psychoanalytic thinking than anything else: Freud courageously (and painfully) undertook his self-analysis, beginning in 1897 and continuing systematically through and beyond the writing of *The Interpretation of Dreams* (1900). With interruptions, it was life long.

Freud's letters to his friend Wilhelm Fliess (1858–1928), an ear, nose, and throat specialist living in

Berlin, provide a unique source of information about these events. Never intended for publication, they were rescued by Princess Marie Bonaparte of Greece (the letters from Fliess have not survived). Starting in 1887, they set out in astonishing detail Freud's reflections, doubts, and formulations as they arose, and can be followed until after the turn of the nineteenth century. An edited version (Freud 1954) was followed by the complete letters (Masson 1985). The demise of the seduction theory and Freud's comments on his self-analysis are described in both.

The self-analysis was crucial for the development of basic psychoanalytic concepts. The richness and complexity of infantile life, including its sexuality, was not fully appreciated until well after *The Interpretation of Dreams* saw the light of day. The Oedipus complex, for example, is described in that book, but its fuller significance was underlined only in footnotes added much later. But Freud regarded the dream book as his single most important work. With copious illustrations from his self-analysis and clinical experience with patients, the study goes beyond the meaning of dreams and approaches the baffling problem of how the mind works.

The theory is involved, but a few key points may be mentioned. The function of dreaming is the preservation of sleep. Stimuli arising from outside the organism (noises, the pressure of bedclothes) or from within (pain, excitation), could often be contained when switched into nonmotor channels (dreaming). The formation of the dream is more complex. In normal waking life, accessible mental content is either conscious at any given moment, or is preconscious—capable of reaching awareness with varying degrees of difficulty. But some mental content is dynamically unconscious—barred from reaching the preconscious by repression. During sleep, this barrier is relaxed, and a partial regression to earlier forms of thinking allows normally unacceptable childhood wishes to find expression, though only in disguise. Even in sleep a censorship is partially operative, and the result is a distorted representation of infantile wishes that gives the dream its seemingly bizarre character. Any semblance of coherence is imposed during waking and recollection: the repression barrier is restored and the dream is (unwittingly) clothed in a measure of rationality.

The recollected (manifest) dream content arises from several sources. Stimuli that threaten sleep are the most obvious, but events and experiences of the day before (day residues) are also used as building blocks. These have no intrinsic importance: they simply provide convenient representations for forbidden wishes. But emerging wishes are themselves distorted by primitive thinking before they are disguised further by the censorship. For Freud, the dream was a hallucinatory wish fulfillment, though generally far removed from the wishes of everyday appearance.

Do dreams provoking anxiety or portraying pun-

ishment also conceal wishes? Freud argued that, even in sleep, a wish can be so strong that its forbidden character can still be recognized and occasion distressing anxiety. In punishment dreams, the transgression portrayed brings with it a wish for punishment. This became clearer still when Freud, later, had conceptualized an unconscious sense of guilt.

The model of the mind outlined in the dream book was the foundation for all Freud's later theoretical formulations. Since it was written during his self-analysis, his pride in it also reflected a new self-confidence. Jones (1955) claims that, at the age of 45, Freud had reached full maturity: intellectual development, always precocious, had not hitherto been matched emotionally. For many years there had been 'continual periods of restlessness, uncertainty, instability and even more definitely neurotic disturbance.' A deep self-assurance had been hidden behind 'strange feelings of inferiority, even in the intellectual sphere.' He had over-idealized men who were, certainly, important to him; of these, Breuer had already withdrawn his support, and the relationship with Fliess was loosening. Self-analysis had revealed the neurotic nature of the tie.

Intellectual loneliness did not diminish productivity. *The Psychopathology of Everyday Life* (Freud 1901) dealt with unconscious motivations in forgetting, slips of the tongue and pen, as well as some 'accidents.' The *Fragment* of Freud's analysis of 'Dora' (1905) described her treatment in 1900. *Three Essays on the Theory of Sexuality* (Freud 1905) met with ferocity hard to imagine today. Freud had widened the concept of sexuality: all existing definitions failed to fit the facts and could not, for example, account for the sexual perversions. Sexuality was now understood within a much wider, developmental perspective. It was considered diphasic: development from infancy to roughly the age of five, centered around bodily erotogenic zones, was followed by a period of comparative quiescence (latency) before a second stage of sexual efflorescence started around puberty. *Jokes and their Relation to the Unconscious* (Freud 1905) arose out of the work on the dream book.

7. *The Growth of the Discipline*

By the time these works appeared Freud was no longer intellectually isolated. Already, in 1902, two young Viennese doctors, Kahane and Reitler, and, more famously, Stekel and Adler were intrigued by psychoanalysis. Freud invited them to meet him for discussions; they gathered together on Wednesdays as 'The Psychological Wednesday Society.' This was the nucleus of an expanding group renamed, in 1908, the Vienna Psychoanalytical Society. By then much had happened. In 1904 Eugene Bleuler, Professor of Psychiatry in Zurich, wrote to inform Freud that he and his staff, most notably his Chief Assistant, C. G.

Jung, had been making use of psychoanalytic ideas for the past two years. (In particular, Jung had applied them in his word-association tests.) By 1906 Freud and Jung were corresponding, and in 1907 Jung visited Freud in Vienna. The friendship grew rapidly and when, in 1910, the International Psychoanalytic Association was formed, Freud asked Jung to be President. Jung's bearing, intellect, qualities of leadership, psychiatric background, and dedication seemed suitable qualifications. Freud was attracted by his personality and regarded him as his 'son and heir.' Jung was greatly impressed by Freud's penetrating intellect. The rise and fall of their relationship is chronicled in their correspondence (McGuire 1974).

Otto Rank had joined the group in 1906; and Karl Abraham, a major figure in the psychoanalytic movement, had long been fascinated by Freud's work when he visited him in 1907. Again, the correspondence between the two gives a vivid picture of their relationship; it continues until Abraham's premature death in 1925 (Abraham and Freud E 1965). In 1907, too, Sandor Ferenczi visited from Budapest, establishing himself the following year as a major figure in the discipline. With the visit of Ernest Jones in 1908, and Hans Sachs in 1910, Freud had a close circle of friends and collaborators.

In 1908, Jung organized an international conference in Salzburg. Freud's work was now discussed widely, in Europe, the USA, and Australia. In 1909 Freud lectured at Clark University, Massachusetts, accompanied by Ferenczi, Jones, and Jung. In 1910 a second international congress was held at Nuremberg, organized by Jung. But internal dissent soon surfaced. In 1911 Adler defected: to Freud's dismay, he refused to try to resolve their conceptual differences. Stekel, gifted, charming, but scientifically unreliable (Jones 1955) left about the same time. Although Freud disliked leadership, he was a fatherly figure for his young colleagues, and bickering and rivalry, especially among the Viennese (cf. Jones 1955) threatened stability. Freud's obvious preference for Jung was resented. But by 1912 Jung was showing signs of disaffection. Many Swiss psychiatrists denounced Freud's theory of sexuality, and, during a course of lectures in the United States Jung omitted any reference to it. Personal relations between the two men cooled, but it was not until 1914 that Jung left the movement on the grounds of incompatible views. The story of the rupture is complex: the Freud/Jung letters and the account by Jones (1955) repay study.

Freud's continuing productivity defied these dissensions. He addressed the problem of the sexual enlightenment of children (Freud 1907), discussed children's sexual theories (Freud 1908), and considered the relationship of the character traits of orderliness, parsimony, and obstinacy to anal eroticism—a paper greeted with indignation when it appeared (Freud 1908). He tried to show that psychoanalysis could contribute to the understanding of art in discussions of

Jensen's *Gradiva* (Freud 1907) and of the relationship between creative writing and day dreaming (Freud 1908). He wrote many other papers in these industrious days.

Two extensive clinical studies appeared in 1909. The first (Freud 1909) describes the very first analysis of a child, though one undertaken by proxy. The successful treatment of 'Little Hans,' who had developed a phobia of horses, was carried out by his father, under Freud's supervision. Hitherto, the childhood roots of neurosis in adults had been understood through a process of reconstruction; with Little Hans those sources could be identified at the time they were operative. Freud believed that the analysis could not be conducted by a stranger, and that the affection and authority of the father (in this case, a physician, too) made treatment possible. Child analysis as a speciality was yet unborn. It was many years before the possibility of directly analysing young children, under favorable conditions, was recognized.

The second patient suffered from an obsessional neurosis (Freud 1909). The symptoms included fears that harm would befall his father and a woman he admired; compelling impulses, one of which was to cut his throat with a razor; and self-imposed irrational prohibitions. He was horrified but attracted by an account of a punishment in which rats ate their way up the victim's rectum. The unconscious sadism underlying his 'fascination' with the torture was among the features on which the analysis cast light. A deeper psychoanalytic formulation of obsessional neurosis emerged from the treatment. Later critics of this case report, and that of Dora, have suggested that Freud neglected the role of the mother in the genesis of the disorders. But Dora's treatment was brief, unfinished, and took place when clinical analysis was still in its infancy. In the case of the Rat Man, the criticism, when sound, is largely based on hindsight.

Another case study (Freud 1911) concerned a subject whom Freud had never seen. Dr. Schreber suffered from a paranoid state, and was under care for 6 years. He later wrote his *Memoirs of My Nervous Illness*, though Freud read them only in 1910. The illness fell into two phases: distressing delusions of persecution preceded delusions of grandiosity. Freud made an ingenious and fascinating reconstruction of the course of the illness. His work stimulated much clinical research. Psychoanalytic theories of psychotic illness fall into two broad groups: either it is part of a neurosis–psychosis continuum, or, unlike neurosis, it involves the loss of higher psychic functions and the re-emergence of more primitive ones, hitherto kept in check by the lost capacities. Many claim that Freud supported the first (unitary) theory; others the second (discontinuity theory). The controversy deserves close attention and is summarized elsewhere (Yorke et al. 1989).

In 1910 Freud (1910) published a study of a childhood memory of Leonardo da Vinci, and the first

of three essays on the psychology of love. In 1911, the first of a series on psychoanalytic technique, and from 1910 to 1912 some short papers on theory that paved the way for later and larger works. An excursion into social anthropology, *Totem and Taboo*, was written between 1912 and 1913. Before war broke out in the summer of 1914, Freud's works included a history of the psychoanalytic movement (Freud 1914), giving his version of the disagreements with Adler and Jung; a fascinating paper on the *Moses* of Michelangelo (Freud 1914); a discussion of the scientific status of psychoanalysis (Freud 1913); and a wide ranging paper on narcissism from which so many ideas were later developed (Freud 1914). These were productive years indeed.

During the war Freud was anxious about two of his sons involved in it. Martin fought in Galicia and Russia; Ernst against Italy after that country entered the conflict. Martin was decorated for gallantry. Medical reasons prevented Oliver's acceptance by the army. Optimism about the outcome diminished after 1915. In 1916 the food shortage began to bite and got worse as the war continued. Freud had few analytic patients. Contact between analysts faced obstacles, some severe, even in the central powers.

For all the difficulties, Freud continued to write. He wrote 12 papers on the theoretical basis of psychoanalysis between 1914 and 1917: only five were published, and a sixth manuscript was found in 1983. There is no trace of the others. *Thoughts for the Times on War and Death* (Freud 1915) anticipated the correspondence with Einstein (Freud 1933) *Why War?* At the University, in the winter terms of 1915–6, and 1916–7, Freud gave a total of 28 lectures on psychoanalysis, and wrote them down, from memory, later (Freud 1916–7). Important changes were made in the *New Introductory Lectures* (Freud 1933), but, taken together, the two series offer an excellent entry into Freud's psychological thinking. The most extensive of Freud's case histories (the *Wolf Man*), written in 1914, was published much later (Freud 1918). In 1919 the baffling problem of masochism (sexual pleasure in pain), and a psychoanalytic look at 'the uncanny' (Freud 1919) were among the many subjects Freud addressed.

The return of many old colleagues to Vienna in 1919 ended wartime isolation. For some time life was hard, but psychoanalysis continued to grow worldwide. In 1919 Martha Freud contracted influenza leading to pneumonia and was ill for many months; the following year Sophie Freud died from a similar cause. Freud grieved deeply.

In 1923 Freud's heavy cigar smoking seemed responsible for a growth affecting the jaw and palate, removed surgically. This was the first of many operations for cancer that Freud survived over the next 16 years, bearing discomfort and pain, often severe, with fortitude. But he continued to work. In that same year, he set out major changes in psychoanalytic theory

(Freud 1923). He had already felt obliged to give aggression (more accurately, destructiveness) the same status as libido. For clinical and theoretical reasons the structure of the mental apparatus was redefined. The ego was now the executive mental agency, holding a balance between the conflicting demands of reality, the id, and the superego. The id strove for immediate drive gratification; its mentation followed the primitive primary process; and its content was dynamically unconscious. But it intruded into an unconscious part of the ego where it encountered checks and controls against inappropriate discharge. The superego was thought of as a greatly extended conscience, based on internalized but modified parental aims and proscriptions, but a substantial part of its operations were unconscious. Its functions included that of the earlier censor. With the aggressive drives of the id partly at its disposal, it put pressure on the ego to withstand those inner urges that ran counter to its prohibitions, aims, and ideals. It was responsible for the sense of guilt, conscious, and unconscious. It must be added that higher mental functions such as conscious thinking and reflection (secondary process), reality testing, and judgment were functions of the ego.

Freud continued to elaborate and modify this model (it should not be reified). He made radical changes in his theory of anxiety (Freud 1926); extended his work on masochism (Freud 1924); and in the field of social psychology added to his earlier study of the psychology of groups (Freud 1921) the far reaching *Civilization and its Discontents* (Freud 1930). The work of his female colleagues and his own clinical experience led him to re-examine the many problems of female sexuality (Freud 1931, 1933). This work aroused opposition, particularly among feminists: Eissler (1977) countered some of these criticisms, but there have been many subsequent developments. It is worth remembering that Freud's close friendships included many women (see, e.g., Freud and Andreas-Salome 1972); and many found his work did much to reduce guilt about sexuality. In *Analysis Terminable and Interminable* Freud (1937) emphasized the therapeutic limitations of psychoanalysis, but he was apprehensive about the possible reception of *Moses and Monotheism* (Freud 1939), which extended his inquiries into early social organization (*Totem and Taboo* 1912–3, *Group Psychology* 1921).

8. Exile: England

In 1938 the Nazis entered Vienna. Freud's books had already been burned in Germany (Freud thought this more advanced than burning him). Through the good offices of Princess Marie and Ernest Jones Freud, together with Martha, their daughter Anna, and others, were allowed to leave for London. Freud was quite ill and tired by the journey, but was pleased by the warm reception that awaited him. The British

Psychoanalytical Society gave him a house at 21 Maresfield Gardens in Hampstead. Anna Freud was his devoted nurse. His remarkable *Outline of Psycho-Analysis* (Freud 1940) was written there. It was there, too, that that death ended his considerable pain and suffering on September 23, 1939. His ashes remain at Golder's Green Crematorium.

Of his many honors, he particularly valued the award of the Goethe Prize for Literature (1930) and his Corresponding Membership of the Royal Society (1936).

9. *Freud's Impact on Twentieth-century Thought and Culture*

There is still a widespread view that Freud's impact on twentieth-century thought, on the arts, sciences, social attitudes, and beliefs, has been far-reaching, even immense. Almost every intellectual discipline, it is said, has been influenced by his ideas. The importance of unconscious motivation is acknowledged widely. Furthermore, many of his concepts—such as the 'Freudian slip' of pen or tongue—have passed into the common language of everyday social intercourse; and his views on sexuality have had a liberating impact on social attitudes and broadened sexual mores.

This view is by no means universal. Freud is undeniably a household name, but more than any other major figure, is the subject of widely varying and conflicting claims. These differences may arise in part from the nature of his subject, for it may be said, with a certain truth, that every man is his own psychologist. They may also arise from an investigative method difficult to understand without personal acquaintance. Freud's neuroscience aside, his principle field of inquiry was subjective and internal, and not readily open to methods appropriate to objective and external research. Many academic psychiatrists and psychologists consider that both Freud's clinical method and his theoretical constructs lack empirical support: in this, they generally equate empiricism with mensuration.

This fact stands in the way of dispassionate assessment even for those actively and intelligently involved in debate. The casual bystander, faced with controversy on a substantial scale, may be simply bemused or follow an opinion in line with personal appeal. And if the intelligent and industrious layman turns to Freud's own writings, he will still encounter significant obstacles. Freud's ideas cannot be picked up piecemeal, but comprehensive reading is difficult without informed guidance. Moreover, reliable introductory works are hard to come by. So it seems unlikely that the forces behind the 'Freudian slip,' and the distinctions between the preconscious and the unconscious, are understood widely. (The term 'subconscious' is more likely to be heard—a word Freud never used except by way of criticism.) Finally, while

the liberalization of sexual attitudes may owe something to Freud, social, economic, and political forces are surely of greater importance. These include the availability of reliable contraception.

Indeed, Freud's views on sexuality, when they are not bowdlerized, are misunderstood widely. But the wider point to be made is this. Most people, in all sections of society, who have any interest in Freud, do not know what he really said. They only know what they think he said. And that is just as true for most critics and commentators who seemingly write with authority, including psychologists, 'hard' scientists, sociologists and art critics, whether the views they express are positive or negative. The viewpoint expressed here may at first seem novel. It was presented in extended form at a meeting of the *Psychoanalytic Heritage Foundation* on October 6, 1992, when it became evident that many Freud scholars were thinking on similar lines. At the recent Conference in London (see below), Douglas Watts stated from the platform that psychoanalysts needed to do more to make the general public more familiar with Freud. 'At present all they have is a caricature.'

For reasons that are not dissimilar, though highly complex, even psychoanalysts have widely diverse approaches to both theory and clinical practice and to the weight they attach to basic Freudian principles (Yorke 1995). Psychoanalysis is avowedly pluralistic (Wallerstein 1988, 1990). And although psychotherapy is often said to derive from the 'talking cure,' similar considerations apply.

In a few instances, Freud's influence, directly or indirectly, on other disciplines is evident, though localized. Disciplines not already mentioned include: education (e.g., many members of the Montessori movement); criminology (The Institute for the Study and Treatment of Delinquency in London); family law in relation to children (e.g., liaison between the Child Study Center and the Law School at Yale University). Properly understood, Freud's work and its true derivatives is capable of fostering advances in many intellectual and sociological fields.

It is pleasing to end on a positive note—in relation to Freud's first love, neuroscience. Freud always hoped that the underlying basis for mental activity in terms of the functioning brain would, one day, yield to investigation. For the greater part of a century that hope seemed slim indeed. But the last few years have brought startling developments. Solms (1997), for example, in a fascinating series of experiments, showed convincingly that the neuropsychology of dreaming was very much in line with the argument set out in *The Interpretation of Dreams*. Work on the neuropsychology of affects, for example, attracted interest from an increasing number of psychoanalysts, and in 1999 a new scholarly journal, *Neuro-psychoanalysis*, was founded. In the first issue, Solms and Nersessian (1999) demonstrated, in a wide-ranging survey, the extensive nature of Freud's contributions to affect

theory. On July 21–3, 2000, the first annual Neuropsychology Conference was held in London entitled *Neuropsychology and Psychoanalytic Perspectives on Emotions*, addressed by the world authorities Jaak Panksepp, Antonio Damasio, and Mark Solms, and introduced by Oliver Sacks. The meeting was a heartening success, and the compatibility between the findings of the two disciplines was seen to be remarkable.

All references to Freud, other than those in the bibliography, are to the *Standard Edition of the Complete Psychological Works of Sigmund Freud* (1953–74) Strachey J (ed.) 24 Vols. Hogarth, London. For specific volumes and page numbers see Vol. 24.

See also: Adorno, Theodor W (1903–69); Allport, Gordon W (1897–1967); Bleuler, Eugen (1857–1939); Bowlby, John (1907–90); Defense Mechanisms; Dreaming, Neural Basis of; Ego Psychology and Psychoanalysis; Erikson, Erik Homburger (1902–94); Foucault, Michel (1926–84); Hypnosis, Psychiatry of; Hysteria; Janet, Pierre (1859–1947); Jung, Carl Gustav (1875–1961); Mead, George Herbert (1863–1931); Mental Health and Normality; Nosology in Psychiatry; Pavlov, Ivan Petrovich (1849–1936); Personality Development in Childhood; Personality Theories; Psychiatry, History of; Psychoanalysis: Adolescence (Clinical–Developmental Approach); Psychoanalysis: Current Status; Psychoanalysis, History of; Psychoanalysis in Clinical Psychology; Psychoanalysis in Sociology; Psychoanalysis: Overview; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychology: Historical and Cultural Perspectives; Sexual Perversions (Paraphilias)

Bibliography

- Abraham H C, Freud E L (eds.) 1965 *A Psycho-analytic Dialogue: The Letters of Sigmund Freud and Karl Abraham*. Hogarth, London
- Breuer J 1882 Report to Bellevue Sanatorium, Kreuzlingen, on Bertha Pappenheim. In: Hirschmüller A (ed.) (1989) *The Life and Work of Joseph Breuer*. New York University Press, New York
- Breuer J, Freud S 1893 *On the Psychical Mechanism of Hysterical Phenomena: A Preliminary Communication*. Hogarth, London
- Breuer J, Freud S 1895 *Studies on Hysteria*. Hogarth, London
- Eissler K R 1977 Comments on penis envy and orgasm in women. *Psychoanalytic Study of the Child* 32: 29–83
- Ellenberger H F 1972 The story of 'Anna O': A critical review with new data. *Journal of the History of the Behavioral Sciences* 8: 267–79
- Freud S 1891b *Zur Auffassung der Aphasien* [trans. Stengel E as *On Aphasia* 1953]. Imago, London
- Freud S 1932 Letter to Stefan Zweig, June 2, 1932. In: Freud E L (ed.) *Letters of Sigmund Freud, 1873–1939*. Hogarth, London
- Freud S 1954 The Origins of Psycho-analysis: Letters to Wilhelm Fliess, drafts and notes: 1887–1902. (Bonaparte M, Freud A, Kris E (eds.) Imago, London

- Freud S 1987 *A Phylogenetic Fantasy*. (Grubrich-Simitis I (ed.) Belknap, Cambridge, MA
- Freud S, Andreas-Salome L 1972 *Letters* [trans. Robson-Scott W, Robson-Scott E]. Hogarth, London
- Gay P 1988 *Freud: A Life for Our Time*. Dent, London and Melbourne
- Hirschmüller A 1989 *The Life and Work of Joseph Breuer: Physiology and Psychoanalysis*. New York University Press, New York
- Jackson J H 1884 Evolution and dissolution of the nervous system. In: Taylor J (ed.) *Selected Writings of John Hughlings Jackson*. Basic Books, New York
- Jones E 1953 *Sigmund Freud: Life and Work*. Hogarth, London, Vol. 1
- Jones E 1955 *Sigmund Freud: Life and Work*. Hogarth, London, Vol. 2
- Jones E 1957 *Sigmund Freud: Life and Work*. Hogarth, London, Vol. 3
- Kaufmann W 1980 *Discovering the Mind*, 3 Vols. McGraw-Hill, New York
- Luria A R 1979 *The Making of Mind: A Personal Account of Soviet Psychology*. Harvard University Press, Cambridge, MA
- McGuire W (ed.) 1974 *The Freud/Jung Letters* [trans. Manheim R, Hull R E C]. Hogarth & Routledge, London
- Masson J M (ed.) 1985 *The Complete Letters of Sigmund Freud to Wilhelm Fliess: 1887–1904*. Belknap Press of Harvard University Press, Cambridge, MA
- Shengold L 1993 *'The Boy Will Come to Nothing!': Freud's Ego Ideal and Freud as Ego Ideal*. Yale University Press, New Haven, CT
- Solms M 1997 *The Neuropsychology of Dreaming: A Clinico-Anatomical Study*. Erlbaum, Mahwah, NJ
- Solms M, Nercessian E 1999 Freud's theory of affect: Questions for neuroscience. *Neuro-psychoanalysis* 1: 5–14
- Wallerstein R S 1988 One psychoanalysis or many? *International Journal of Psycho-analysis* 69: 5–21
- Wallerstein R S 1990 Psychoanalysis: The common ground. *International Journal of Psycho-analysis* 71: 3–20
- Yorke C 1992 *Freud's impact on the twentieth century*. Paper read to the Freud Literary Heritage Foundation, New York, October 6
- Yorke C 1995 Freud's psychology: Can it survive? *Psychoanalytic Study of the Child* 50: 3–31
- Yorke C, Wiseberg S, Freeman T 1989 *Development and Psychopathology: Studies in Psychoanalytic Psychiatry*. Yale University Press, New Haven, CT

C. Yorke

Friendship, Anthropology of

Friendship is an informal social relationship. Contrary to kinship, it is based on choice and voluntariness: friends are sought and must be won. Friendship is an acquired not an ascribed status. The relationship is often viewed as affective. However, the 'emotional content' of friendship can vary strongly (Paine 1969, p. 507). Mutuality, concerning interest in the relationship, appears to be one important requirement.

In fact, in some societies there exist asymmetrical friendships, sometimes being difficult to distinguish from patronage. Friendship is based on sharing: besides matters of material value friends share their time, their problems, plans, hopes, and thoughts. Therefore, trust and loyalty are basic requirements for friendship (Kennedy 1986, 128 ff.). But this aspect also involves dangers: if friends share secrets, and the friendship is eventually broken off, it can lead to gossip and enmity. Some authors note that generally no fixed role expectations are attached to friendship. But in fact many societies do have well-defined expectations.

From the perspective of comparative social and cultural anthropology it is obvious that besides common core aspects, notions of friendship are highly variable. To this day there is a lack of data concerning the different meanings friends and friendship can have in different societies, and even within one society. Most languages have many terms for friends (in English among others: *pal*, *chum*, *buddy*, *bosom friend*, *old friend/chap*, *acquaintance*, etc.), which differentiate between kinds and meanings of social relationships called 'friendship.' One also finds different concepts of friendship within one society relating to sex, age, social status, as well as the different places or contexts in which friendship takes place, e.g., neighborhood, school, or place of work. The 'growth' of friendship has also been an underestimated aspect in research. A friendship does not only change in the course of time, it also changes due to the period of its existence: after 20 years a friendship, like all other emotional relationships, is no longer the same as it was in the beginning. As Sarah Uhl (1991, p. 90) states, friendship is a type of 'noninstitutionalized institution' distinguished from blood-brotherhood and fictive kinship—overlooking the fact that with a new stage in a life cycle, friendships may be transformed to other more institutionalized relationships like co-parenthood.

1. Anthropological Interest in Friendship

Uhl (1991) states that research on friendship is difficult due to the fact that despite its being a public phenomenon, friendship is viewed as private. But the same can be said of marriage, and marriage has been one of the most important topics in anthropology since its beginnings. More probably anthropologists have focused too strongly on kinship (see *Kinship in Anthropology*), thus becoming unaware of sometimes even overlapping concepts of friendship.

Until now most empirical research has been conducted in industrialized societies, and basically there are only few studies by cultural anthropologists. If anthropologists have studied friendship at all, they have concentrated on formal relationships with well-defined mutual obligations such as ritual kinship, exchange, trade, or working relationships. These

anthropological studies—most of them conducted in Southern Europe—overemphasized male friendships and neglected friendships between females (Boissevain 1974, Gilmore 1975). For a long time it was supposed that adult women in particular form their most important social relationships inside the family and with relatives. Different forms of friendship between men are on the other hand an important part of public life. These factors have contributed to the lack of research concerning friendship between women, with the exception of some more recent studies (e.g., Kennedy 1986).

One finds very little information on meanings of friendship in nonindustrial societies. One exception is Thomas Kiefer's work, who in 1968 published a major article on institutionalized friendship and warfare among the Tausug of Jolo, in the Philippines. The Tausug perceive strangers and all people not related by kinship as potential enemies. In this context friendship—formally sworn to by oath on the Koran—is very important for forming alliances between kin groups, as well as, on a higher regional level, between local leaders. These can be former or—should the oath be broken—future enemies. This formal friendship in Tausug society derives its meaning from enmity—friends are allies against common enemies (Kiefer 1968). However, Kiefer's study once again focuses on highly formalized male friendship patterns. Institutionalized and formalized friendship patterns, trade and exchange partnerships, like the bond friendship' (*tau soa*) in Tikopia described by Firth (1967), are still today given priority in studies by cultural anthropologists. On the level of intercultural comparison (including more traditional societies), it would be of importance to investigate the assumed declining significance of kinship and the possibly growing importance of friendship.

2. Themes and Methods

Within the realm of anthropological research different dimensions of friendship have been stressed. Eric R. Wolf differentiates between 'emotional' and 'instrumental relations,' which both satisfy a deficit of some kind in each participant of the dyad (1966, 10 ff). The labels emotional/instrumental are somewhat misleading, because both types have a quite instrumental quality. Wolf's approach links friendship with wider societal forms: societies produce different deficits, which different types of friendship respond to. His conception is a useful starting point for further research, embedding friendship in a wider theoretical frame. In that way it differs from Robert Paine's (1969) above-mentioned conception of an emotional, more inward-turning modern middle-class friendship.

Jacobson's approach to research on friendship emphasizes the situational aspect. He is not only concerned with constant friendship and its character-

istics but 'with the labelling process itself, that is, with the situations in which a person gives and takes away the label of "friend".' (Jacobson 1975, p. 225). Both friendship and affiliation with an ethnic group can change depending on the situation. In combination with an examination of group identities, research on friendship could lead to important results in the investigation of interethnic relationships. Friends can be made across ethnic boundaries; besides binding themselves by marriage, people can form alliances between different ethnic groups. An example of the discussion of interethnic relationships in folklore about friendship is given in a story about a Kipsigis and a Masai, who exchange honey and arrows, cheating each other in the process (Dundes 1971, pp. 176–7). Dundes explains that the topic of making and breaking friendship in African societies is a common motif in narratives. In a story about a crippled and a blind person, for example, who join forces during an attack by enemies, the latter takes the first on his back and runs, while the crippled person directs him. No matter whether the failing of a friendship or its advantages are the central theme, exchange is always the constituent core of the described relationship. The different motifs in African narratives show the conflict between social norms and actual behavior. This discrepancy between a value system and behavior is another aspect of friendship that has not yet been thoroughly studied.

Reina (1959) discusses two concepts of friendship, existing parallel to one another among Indians and Ladinos in Guatemala. The Ladino term *cuello* means taking up many relationships and alliances, which can change under different circumstances. The Indian concept of friendship implies having only very few friends, who are made during adolescence. This example shows that divergent notions of friendship can prevent close relationships between members of different ethnic groups. But without more detailed descriptions of intraethnic concepts of friendship, studies about cultural differences and interethnic friendship are impossible.

In early network analysis the issue of friendship played an important role (Bott 1957). Network analyses seemed to be appropriate for research about weak institutionalized social relationships in complex societies. Friendship is one possible content of such social relations, but until now research has been restricted mostly to Western industrialized societies (e.g., Allan 1989, Fischer 1982). Analysis of networks of friends in more 'traditional' societies must be preceded by studies which describe who is referred to as a friend, and what kind of different categories of friends exist within a given cultural context. As shown above, it cannot be taken for granted that concepts of anthropologists and informants are congruent. With a better knowledge of the meanings and terms attached to the notion of friendship, network analysis and other quantitative methods might also be applied in inter-

ethnic situations. From the interplay between participation in everyday life, an evaluation of one's own experiences, and a systematic investigation, a more differentiated picture of the ambiguous term 'friendship' could emerge.

3. Fieldwork and Friendship

Friendship is not only an important research topic, it is also an important relationship during fieldwork (see *Fieldwork in Social and Cultural Anthropology*). The situation regarding friendship between anthropologists and people with whom they live, often for long periods of time, is hardly documented, even though it is quite probable that these are very relevant personal experiences during fieldwork and in some cases they are even the precondition of fieldwork. Exceptions being the articles by Friedman Hansen (1976), Hendry (1992), and Reina (1959). Friendship formed in the course of fieldwork has neither been treated very honestly, nor has it been dealt with systematically in anthropology to this day. Most anthropologists note having made many friends during their fieldwork. In publications it is obligatory to thank all 'friends,' without whom research would have been impossible. In most cases the nature of this friendship is not described, nor is the significance it had for the fieldwork. Therefore, the question of how anthropologists cope with unfamiliar concepts of friendship is left open. Most studies do not make explicit what is meant by friendship, and few authors relate whether friendship with informants is called so from the perspective of the fieldworker or from the emic view. Other intercultural forms of friendship—those between anthropologists and informants are only one special case—have gone unstudied.

In the literature on fieldwork, friendship is sometimes depicted from a very one-sided angle: 'Informants become friends. This creates another dilemma, for it leads to a very instrumental, and often dishonest, approach to friendship. The interest of science demands that you milk your informants. You do this by making them your friends, by exchanging confidences, giving presents, talking for hours about subjects that bore you. Are these friends or scientific objects? The subjects themselves often realize that they are being used. For some fieldworkers the scientific ends justify the means. They have no moral problem. For many others, however, the dilemma remains.' (Boissevain 1985, p. 273)

Other authors warn against making friends with informants on the grounds of possible role conflicts. Spradley, for example, advises anthropologists, repeating the same questions over and over again on their endless quest for explanations, to refrain from bothering their friends with this boring task. Apart from this, however, he has no objection to making friends during fieldwork (Spradley 1979, p. 28).

An exceptionally good account of friendship in the course of fieldwork has been published by Joy Hendry (1992). In detail she describes the changes within the relationship to her Japanese friend that spanned nearly two decades. She also compares, at least marginally, to what extent concepts of who is related to as a friend, as well as expectations regarding how friends should relate to one another, correspond in Great Britain and Japan. Hendry mentions that in Japan it is not possible to bring inequality in line with friendship. As long as she did research in rural Japan, similarities and equality dominated the relationship to her friend Sachiko, who also did not come from a rural area. Later, influenced by their respective cultures, their lives developed differently from one another. Apart from individual changes due to the respective transformations in their lives, i.e., from being students to married mothers of school children, the actual research project nearly caused a breaking off of their friendship. Conflicts arose when Hendry began to study manners of communicating in Sachiko's *nakama* (reference group). She integrated her into the project, which caused Sachiko to have role conflicts within her own *nakama*. However, the author gathered more insight out of this conflict than out of any other contacts with Japanese women. She evaluated her fieldwork experience as one aspect of participant observation—it having been, on the one hand, scientifically successful but personally nearly having led to a loss. The situation became more relaxed when Hendry left Japan, after enough time had elapsed, and she had analyzed and discussed the experience with Sachiko in a number of letters.

The example concerning Guatemaltecan Indians equally shows that friendship in the context of fieldwork may cause problems due to an initial unawareness of cultural differences between notions of friendship. Guatemaltecan Indians have clearly defined expectations regarding friendship, exclusiveness being one criteria for friendship. Ruben E. Reina states that not only his informants, but also those of his wife, terminated relationships, because both anthropologists had also formed contacts to other informants, thus not meeting the expectations of exclusiveness. Only through detailed explanations of why anthropologists carrying out fieldwork must speak to more than one person was the problem solved (Reina 1959, p. 48).

Role conflicts are one of the major topics discussed in writings by anthropologists on friendships formed in the context of fieldwork. In an article by Judith Friedman Hansen (1976), role conflicts are mentioned as the central problem. This is hardly astonishing, since not only during fieldwork but also in other work-related situations friendship can cause conflicts regarding loyalty, possessiveness, and role expectations. At home anthropologists have to play different roles within their families, partnerships, and within academia as well—thus, cross-cutting and intersecting

friendship patterns are not typical of fieldwork alone. The problem is only more obvious and perceived more consciously during fieldwork. Coping with these conflicts requires time, much talking, patience, and a deeper understanding of the rules of the other culture.

See also: Friendship: Development in Childhood and Adolescence; Interpersonal Attraction, Psychology of; Interpersonal Trust across the Lifespan; Patron–Client Relationships, Anthropology of

Bibliography

- Allan G A 1989 *Friendship: Developing a Sociological Perspective*. Westview Press, Boulder, CO
- Boissevain J 1974 *Friends of Friends: Networks, Manipulators and Coalitions*. Basil Blackwell, Oxford, UK
- Boissevain J 1985 Ethnographic fieldwork. In: Kuper A, Kuper J (eds.) *The Social Science Encyclopedia*. Routledge & Kegan Paul, London, pp. 272–4
- Bott E 1957 *Family and Social Network*. Tavistock, London
- Dundes A 1971 The making and breaking of friendship as a standard frame in African folktales. In: Maranda P, Köngäs Maranda E (eds.) *Structural Analysis of Oral Tradition*. University of Pennsylvania Press, Philadelphia, pp. 171–85
- Firth R 1967 [1936] Bond friendship. In: Firth R (ed.) *Tikopia Ritual and Belief*. Allen & Unwin, London, pp. 108–15
- Fischer C S 1982 *To Dwell Among Friends: Personal Networks in Town and City*. University of Chicago Press, Chicago
- Friedman Hansen J 1976 The anthropologist in the field: Scientist, friend, and voyeur. In: Rynkiewicz M A, Spradley J P (eds.) *Ethics and Anthropology. Dilemmas in Fieldwork*. Wiley, New York, pp. 123–34
- Gilmore D C 1975 Friendship in Fuenmayor: Patterns of integration in an atomistic society. *Ethnology* 14: 311–24
- Hendry J 1992 The paradox of friendship in the field: Analysis of a long-term Anglo–Japanese relationship. In: Okely J, Callaway H (eds.) *Anthropology and Autobiography*. Routledge, London
- Jacobson D 1975 Fair-weather friend: Label and context in middle-class friendships. *Journal of Anthropological Research* 31: 225–34
- Kennedy R 1986 Women's friendships on Crete: A psychological perspective. In: Dubisch J (ed.) *Gender and Power in Rural Greece*. Princeton University Press, Princeton, NJ
- Kiefer T 1968 Institutionalized friendship and warfare among the Tausug of Jolo. *Ethnology* 7: 225–44
- Leyton E (ed.) 1974 *The Compact. Selected Dimensions of Friendship*. Newfoundland Social and Economic Papers, No. 3. Memorial University of Newfoundland, Newfoundland
- Paine R 1969 In search of friendship. *Man* 4: 505–24
- Reina R E 1959 Two patterns of friendship in a Guatemalan community. *American Anthropologist* 61: 44–50
- Spradley J P 1979 *The Ethnographic Interview*. Holt, Rinehart and Winston, New York
- Uhl S 1991 Forbidden friends: Cultural veils of female friendship in Andalusia. *American Ethnologist* 18: 90–105
- Wolf E R 1966 Kinship, friendship, and patron–client relations in complex societies. In: Banton M (ed.) *The Social Anthropology of Complex Societies*. Tavistock, London

B. Beer

Friendship: Development in Childhood and Adolescence

Most children and adolescents have friends. Friendships are characterized by mutual liking and expectations that friends spend more time with one another than acquaintances, that cost–benefit balances in the social exchanges between friends are favorable, and, among older children and adolescents, that friends are loyal, trusting, and intimate (Bukowski et al. 1996). This article deals with friendships during the first two decades of life and their developmental significance.

1. Incidence

Infants and toddlers sometimes show preferences for one another although it is not always easy to compare these relationships to the friendships of older children. The word ‘friend’ begins to be used by children during the third and fourth years and, by that time, about three-fourths of children are involved in reciprocated friendships. Proportions rise through middle childhood (when about 85 percent of children have a best friend and usually several good friends) and decline slightly during adolescence. Only about 10 percent of older children and adolescents have no friends (Hartup and Stevens 1997).

Friendship networks are small among young children, averaging about 1.7 friends for boys and 0.9 friends for girls during the preschool years. These numbers increase to 3.0 to 5.0 best friends during middle childhood and adolescence, depending on whether one includes unreciprocated choices as well as reciprocated ones. Time spent with friends changes with age too, being greater in middle childhood and adolescence than at any other time (29 percent of time awake among teenagers).

2. Friendship Expectations

Friendships are understood by children and adolescents (especially) to provide the individual with companionship and intimacy more frequently than parent–child relationships, but less frequently with experiences of compliance and control. Most dramatically, friendships are understood to differ from parent–child relationships in power sharing (Furman and Buhrmester 1985). Overt affection is provided less frequently to children by friends than by parents but nurturance in general is understood to be provided by both.

Differences between children and adolescents in the ways that they describe their friends occur mainly in the number, complexity, and organization of information and ideas about these relationships (Rubin et al. 1997). Among young children, for example, friend-

ship expectations are most likely to emphasize common activities and concrete reciprocities as well as the ‘here and now’ (‘We play’). Among older children, friends are described as sharing values and rules about loyalty and trust; friends also expect to spend time together and engage in constructive conflict resolution. Among adolescents, friends are expected to share interests, be understanding, and empathetic; similarity between oneself and one’s friends is especially important among teenagers. These developmental changes have sometimes been interpreted as reflecting changes in the child’s understanding of reciprocity or mutuality (Youniss 1980). Indeed, the emphasis on mutuality in these expectations at all ages suggests that the meaning structure underlying these relationships is ‘symmetrical reciprocity.’

3. Behavior with Friends

Children spend more time with their friends than with nonfriends. Some researchers, in fact, have used time-sharing as a means of identifying friendships (Hinde et al. 1985), although most investigators prefer validation by the children themselves or an outside observer (a parent or a teacher).

The social exchanges of friends and nonfriends are different in four major respects (Newcomb and Bagwell 1995): *positive engagement* (friends talk, smile, and laugh together more frequently than nonfriends); *task activity* (friends are more task-oriented and spend more time on-task than nonfriends); *relationship mutuality* (friends are more affirmative with one another, more mutually oriented in their interaction, and more likely to emphasize equality in their exchanges than nonfriends); and *conflict management* (friends use disengagement and negotiation as opposed to power assertion proportionally more often than nonfriends). Mutuality and symmetry thus mark the social exchanges of children and their friends as well as adolescents; these results are another reason for suggesting that reciprocity and symmetry are the behavioral hallmarks of friendship during the first two decades of life.

4. Similarity between Friends

Children and adolescents are more similar to their friends than to nonfriends. Greatest similarity is evident between friends in gender, race, and age—extending from early childhood through adolescence. School-aged children are more similar to friends than to nonfriends in prosocial behavior, antisocial behavior, shyness dependency, depression, and social acceptance/rejection. Friends also share biases in social perception, although it should be recognized that these similarities do not mean that friends are carbon copies of one another (Hartup 1996). By

adolescence, friendship similarities vary according to attribute, with greatest similarities occurring in school-related attitudes, aspirations, and achievement along with life-style behaviors that include smoking, drinking, drug use, and sexual activity. Generally, friends are not very similar to one another in personality traits measured with paper and pencil tests—either in childhood or adolescence. Similarities between friends are multidetermined. Owing to the manner in which schools and neighborhoods are organized in most industrialized countries, children are more likely to come together with children who are like themselves than who are unlike them. Once in the same school or neighborhood, they are more likely to be exposed to the same socialization agents (e.g., teachers). Beyond these sociodemographic conditions, children are also more attracted to children who are like themselves rather than different; children and adolescents also become more like one another as an outcome of mutual socialization. The relative extent to which mutual selection and socialization contribute to the similarity between friends depends on many conditions, including characteristics of the children themselves, the nature of their interaction, and the behavioral attributes being measured (Urberg 1999).

5. Gender Differences

One of the most notable characteristics of friendships in childhood and adolescence is their gender concordance. Opposite-sex friendships are relatively rare, even among preschoolers. Boys and girls do not differ in the proportion who have friends, although friendship networks are likely to be somewhat smaller among girls than among boys (Rubin et al. 1997).

Everyone knows that the activities of boys and their friends differ from the activities of girls and their friends. Certain interpersonal provisions also differ. Girls, for example, expect to be provided with greater affection, intimacy, and instrumental help by their friends than do boys. Intimacy concerns are more common in girls' talk about their friendships than in boys' talk; self-ratings of girls' friendships are more intimate; and self-disclosure between girls and their friends is more common than between boys and their friends. These gender differences are recognized and accepted by both boys and girls. Not well studied, however, are gender differences in intimacy when the construct is defined in terms of camaraderie and shared mastery rather than in terms of empathy and self-disclosure.

6. Children Who Have Friends vs. Those Who Do Not

Children who have friends are more likely to evince better adjustment over time than those who do not. Research supports this conclusion in that: (a) children

and adolescents who have friends are better adjusted than those who do not; (b) children referred for clinical treatment are more likely *not* to have friends than children who have never been referred for treatment; and (c) youngsters who have friends feel better about themselves and are less lonely than youngsters who do not (Hartup 1996). These results have been widely reported, but they are difficult to interpret. First, 'having friends' usually means having 'good friends' in the sense that the relationships in question are harmonious and mutually supportive. This confounding is important since research with adults shows that *not* having problematic friendships is more closely related to good adjustment than actually having supportive and intimate ones (Hartup and Stevens 1997). Second, correlational results cannot be used to support causal interpretations: having friends may contribute to individual well-being but the reverse may also be true. Third, existing data may not be adequately refined: having friends may be associated with certain behavioral adaptations (e.g., self-esteem) but not others (conduct disorder).

Longitudinal studies assist in sorting out these issues when adequate assessment of early status is included in the research design and multiple measures are used to assess outcome. Accordingly, certain studies show that having a friend increases self-esteem and decreases psychosocial difficulties among preadolescents during changes from lower to middle to high school. Also, having friends in middle childhood predicts young adult feelings of self-worth, family attitudes, and the absence of depression, but not sociability, school performance, educational aspirations, and job performance (Bagwell et al. 1998). These results indicate that merely having a friend contributes selectively rather than generally to developmental outcome.

One other consequence of having friends during childhood should be mentioned—success in early romantic relationships. Having same-gender friends during middle childhood forecasts having romantic relationships during adolescence. Having friends in middle childhood also enhances success in early romantic relationships, even when the quality of early parent-child relationships is factored out (Sroufe et al. 1999). These results conform to theoretical predictions that childhood opportunities for intimacy in an asexual context are necessary precursors to heterosexual adaptation.

7. Characteristics of One's Friends

Considerable weight in determining developmental outcome is carried by characteristics of the child's best friends. In many instances, more can be learned by knowing who children's friends are than by knowing merely that they have a best friend. For example, following divorce or remarriage (both stressful events), having socially skilled friends who have few behavior problems promotes resilience whereas having imma-

ture friends who have many behavior problems does not (Hetherington 1999). Conversely, considerable evidence shows that friendships between antisocial children increase the risk of antisocial behavior among both children (Coie and Dodge 1997).

What socialization experiences between friends contribute to these outcomes? Although observational learning (modeling) may make a contribution, it is also the case that the interaction between friends varies according to the characteristics of the children involved. Overtly aggressive boys, for example, are more contentious and conflict-instigating with their friends than nonaggressive boys are, and engage in more deviant talk (Dishion et al. 1995). 'Relationally aggressive' children, who hurt others by exclusion or doing things that damage the victim's relationships with other children, have friendships that are intimate and exclusive but, at the same time, marked by frequent displays of relational aggression (Grotper and Crick 1996). One can easily guess how this admixture of intimacy and relationship-oriented aggression incites conflict and uncertainty, as well as provides instruction in how to draw others into dysfunctional relationships.

8. Friendship Stability

Children who have friends at one time are likely to have friends at other ages. Changes in partners are common, although friendships are more stable than is sometimes appreciated (Hartup and Stevens 1997). Preschool-aged children often maintain these relationships for many months. Friendship stability, however, increases with age until about 70 percent of adolescents report that their friendships last for a year or more; the majority of 13- and 14-year olds who have friends report that these relationships have lasted between one and five years.

Stability varies according to child characteristics: friendships between aggressive, antisocial children are less stable than friendships among nonaggressive children (Dishion et al. 1995). Disabilities and psychosocial difficulties are also associated with reduced friendship stability. In these instances, stability is reduced as a consequence of limitations on the child's ability to regulate emotion and other deficits in social skills.

Friendship stability has considerable developmental significance. Kindergarten children, for example, who maintain friendships across the school year show improvement in attitudes toward school and other school-related behaviors, whereas children who do not maintain their friendships show poorer adaptations. In other instances, friendship stability is correlated with outcome according to characteristics of the child's friends: stable friendships with children who have conduct problems, for example, increase children's own behavior problems, whereas stable rela-

tionships with children who are socially withdrawn do not increase the children's social withdrawal. Friendship stability thus affects children differently according to the behavior in question and other aspects of their relationships (Berndt et al. 1999).

9. Friendship Quality

Friendships among children and adolescents vary in their social and emotional qualities. Some relationships are supportive and intimate, some conflict-ridden. One can differentiate among these relationships in terms of reciprocity, self-disclosure, overt hostility, and covert hostility. Friendship structures have been most commonly described, however, simply as 'positive' or 'negative' (Bukowski et al 1996).

Supportiveness between children and their friends is positively correlated with school involvement and achievement while being negatively correlated with school-based problems. Supportiveness between school-aged friends is positively correlated with self-esteem in early adulthood, as well as negatively correlated with later identity problems (Bagwell et al. 1998). Friendship quality is associated with most developmental outcomes, however, in combination with child characteristics and friendship stability. For example, boys with very low-quality friendships and who are delinquent at age 13 to 14 escalate more over the next year in delinquent activity than either delinquent or nondelinquent boys in better relationships (Poulin et al. 1999). Other results show that sociability increases most and social withdrawal increases least across the transition from elementary school to junior high school when friendships are *both* high quality and stable (Berndt et al. 1999). Although causal direction has not been established in these data, the weight of the evidence suggests that good developmental outcomes are most likely when a child has friends, those friends are socially skilled, and these friendships are supportive and stable. Friendships between children who are not well socialized are mixed blessings.

10. Conclusion

Friendships are developmentally significant from early childhood through adolescence and beyond. Friendships are cognitive and affective resources, fostering social problem solving, self-esteem, and well-being. Friends socialize one another (sometimes for better, sometimes for worse). Supportive and intimate friendships generally are developmental advantages, whereas contentious relationships between individuals at risk generally are disadvantages.

Even so, friendships are not all alike. Children differ in the company they keep and in the social and affective qualities of their relationships. Developmental assessment is more complete when one knows who

a child's friends are and what these relationships are like. Developmental outcomes need to be better understood, however, especially in terms of children's involvement in relationships that can place them at risk. Particularly needed are studies of the manner in which friendship experience combines with temperament and early experience, social skills, family relationships, and the social context (especially the child's encounters with stress) to affect developmental outcome. One must also consider that childhood friendships may be more important in accounting for developmental outcome in some individuals than others.

See also: Divorce and Children's Social Development; Friendship, Anthropology of; Infancy and Childhood: Emotional Development; Interpersonal Trust across the Lifespan; Trust: Philosophical Aspects; Trust, Sociology of

Bibliography

- Bagwell C L, Newcomb A F, Bukowski W M 1998 Preadolescent friendship and peer rejection as predictors of adult adjustment. *Child Development* **69**: 140–53
- Berndt T J, Hawkins J A, Jiao Z Y 1999 Influences of friends and friendships on adjustment to junior high school. *Merrill-Palmer Quarterly-Journal of Developmental Psychology* **45**: 13–41
- Bukowski W M, Newcomb A F, Hartup W W (eds.) 1996 *The Company They Keep: Friendship in Childhood and Adolescence*. Cambridge University Press, Cambridge, UK
- Coie J D, Dodge K A 1997 Aggression and antisocial behavior. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology*. Wiley, New York, Vol. 3, pp. 779–862
- Dishion T J, Andrews D W, Crosby L 1995 Antisocial boys and their friends in early adolescence: Relationship characteristics, quality, and interactional process. *Child Development* **66**: 139–51
- Furman W, Buhrmester D 1985 Children's perceptions of the personal relationships in their social networks. *Developmental Psychology* **21**: 1016–24
- Grotjeter J K, Crick N R 1996 Relational aggression, overt aggression, and friendship. *Child Development* **67**: 2328–38
- Hartup W W 1996 The company they keep: Friendships and their developmental significance. *Child Development* **67**: 1–13
- Hartup W W, Stevens N 1997 Friendships and adaptation in the life course. *Psychological Bulletin* **121**: 355–70
- Hetherington E M 1999 Social capital and the development of youth from nondivorced, divorced, and remarried families. In: Collins W A, Laursen B (eds.) *Relationships as Developmental Contexts, Minnesota Symposia on Child Psychology, Vol. 30*. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 177–209
- Hinde R A, Titmus G, Easton D, Tamplin A 1985 Incidence of 'friendship' and behavior with strong associates versus non-associates in preschoolers. *Child Development* **56**: 234–45
- Newcomb A F, Bagwell C L 1995 Children's friendship relations: A meta-analytic review. *Psychological Bulletin* **117**: 306–47
- Poulin F, Dishion T J, Haas E 1999 The peer influence paradox: Friendship quality and deviancy training within male ado-

- lescent friendships. *Merrill-Palmer Quarterly-Journal of Developmental Psychology* **45**: 42–61
- Rubin K H, Bukowski W M, Parker J G 1997 Peer interactions, relationships, and groups. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology*. Wiley, New York, Vol. 3, pp. 619–700
- Sroufe L A, Egeland B, Carlson E 1999 One social world: The integrated development of parent-child and peer relationships. In: Collins W A, Laursen B (eds.) *Relationships as Developmental Contexts*. L. Erlbaum Associates, Mahwah, NJ, Vol. 30, pp. 241–61
- Urberg K A 1999 Introduction: Some thoughts about studying the influence of peers on children and adolescents. *Merrill-Palmer Quarterly-Journal of Developmental Psychology* **45**: 1–12
- Youniss J 1980 *Parents and Peers in Social Development: A Sullivan-Piaget Perspective*. University of Chicago Press, Chicago

W. W. Hartup

Frontiers in History

Frontiers are as ancient as the state and the history of frontiers begins as far back as Herodotus, but the frontier as a myth is a recent invention. In Europe, the word 'frontière' first appears in French in the fourteenth century to indicate a facade in architecture or a military order of battle and only gradually came to mean the limits of state control. It passed into English and was widely used in the sixteenth century to mean a barrier against attack. In the thirteenth century the Slavic word 'greniz' or 'granitsa' was applied, particularly by the Teutonic Order, to the area of contact between Germans and Slavs, and replaced the Gothic term 'mark'. In the scholarly literature, definitions of frontiers have shifted from line and place to process, symbol and mythology as the disciplines involved—anthropology, geography and history—interacted creatively with one another under the general influence of the contextual approach. This has not been a linear evolution; nor has it been uniformly accepted, least of all by advocates of geopolitics, who still perceive frontiers as constituting a power grid, even when they have been determined by cultural or ideological factors.

1. Frontiers and the Historical Sciences

It is just over 100 years since Frederick Jackson Turner published his famous essay that still serves as the main frame in a triptych of historical approaches to frontiers in history (Turner [1893] 1921). Given the large number of American historians and their methodological restlessness, the comparative historiography of fron-

tiers has been dominated by their revisions of Turner. Flanking Turner's frontier on one side is a different spatial concept tied to the rise and consolidation of the centralized state that developed out of the French experience, as interpreted by the Annales school. On the other side a third panel represents the symbolic geographies, that is, the construction of imaginary borders between civilizations, as exemplified by the classical dualism of Europe and Asia, and the trialism of the Cold War: the West, the Communist Bloc and the Third World.

Anthropologists came later than historians to frontier studies, but then arrived in considerable force. In the 1950s they began to engage in extensive fieldwork on the margins of sedentary and nomadic societies (Khazanov 1994). Their insights were rapidly absorbed by historians, who had already been seriously exploring frontiers for over half a century. Finally, geographers based in local history and influenced by cultural studies began to redefine the ideas of space and frontiers in terms of linguistic and social contexts (Godlewka and Smith 1994). As a result the current state of research on frontiers is multidisciplinary and comparative.

2. Types of Frontiers

Frontiers may be separated into three basic types: consolidated state frontiers; dynamic frontiers of advancing settlements; and symbolic frontiers. On occasion, features of two or even three of these types coincide with one another. Each has a variety of subtypes. All are characterized by the existence of zones of influence and exchange on both sides of the real or imaginary line that presumes to divide them.

2.1 State Frontiers in Antiquity

The earliest historical evidence of frontiers expresses the needs of the state to define the limits of public authority, as in the empires of ancient Mesopotamia and Egypt, or later of private property in the Greek city states. In both cases, boundary markers were endowed with a sacred character and their removal signified the victory of a conqueror (Foucher 1991). The Athenian and Roman Empires and, independently, the Chinese Empire introduced two additional features: the construction of walls or defensive lines, and, second, the definition of peoples beyond the frontiers as culturally inferior, that is, barbarian. The attempts of the Roman and Chinese emperors to establish fixed frontiers repeatedly broke down under the pressure of migration, raiding and the development of extensive cultural and economic exchange. In the case of the Chinese-steppe frontier, the process of acculturation, transculturation and assimilation, sometimes simplified to 'trade or raid,' worked in

both directions long before the Great Wall was constructed and continued into the twentieth century. Relations between the Romans and peoples on the periphery of their control varied greatly according to location over centuries. Even the walls (limes), especially in the UK and Tunisia, were designed less to exclude, and more to control the movement of people and goods. The collapse in the West is increasingly regarded as a process of mutual accommodation of Romans and barbarian tribes through the emergence of a new military elite of the warlord type (Whittaker 1994).

2.2 Islamic State Frontiers

The Ottoman Empire represents a subtype of state frontier, intermediate between the ancient empires and the modern European state system. Embodying the precepts of Islam, it did not recognize until the late eighteenth century the concept of state frontiers in the European sense, but rather a division of the world into two cultural spheres: the Dar ul-Islam, the abode of Islam, and the Dar ul-Harb, the abode of war. But even this rigid division could not be strictly maintained. For example a contested space occupied by nomads opened up all along the Islamic-Byzantine frontiers during the relative stability of the tenth century. By the thirteenth century, there had arisen a culture of holy warriors (gazhis) with a corresponding type on the Byzantine side (Akritai), at first composed of Greek, and then of Turkomen tribesmen. This intermediate zone where war and trade often alternated facilitated the penetration and conquest of the Byzantine Empire by the Ottomans (Witteck 1938). At the western end of Ottoman power, an even more complex military frontier emerged where three empires, the Ottoman, Habsburg and Venetian (triplex confinium) confronted and co-existed with one another for three centuries (ca. 1500–1800). Small wars, population transfers, natural disasters and deeply rooted uncertainties shaped a border society that invented a variety of survival strategies, from cross-border trade and blood brotherhood to religious conversions, and gave rise to a variation of the ghazi warrior culture in the form of the Christian Uskoks (Kaser 1997, Roksandic 1998). Along the northern Ottoman frontier, similar conditions existed at the meeting ground of both the Poles, Russians and Ottomans in the Ukraine, and of the Russians, Ottomans and Iranians in the Transcaucasus, which also produced a warrior culture of Cossacks. Curiously, Ottoman specialists have yet to make the frontier experience the central theme in the history of the empire.

By the nineteenth century, the Ottoman frontier societies became more sharply divided by fundamental cultural conflicts. As Ottoman power receded in southwestern Europe and along the Black Sea, littoral

new state boundaries were drawn by international negotiation (first with Russia in the Treaty of Kuchuk-Kainardji, 1774) and great power intervention. Later, both the idea of rigid linear frontiers along national lines and the power to enforce them came from western Europe.

2.3 West European State Frontiers

A very different kind of frontier emerged in Western Europe during the transition from the feudal fragmentation of political power to the centralized absolutist monarchy. France, which set the pace frontiers as the limit of state power, evolved from a military to an ideological, financial and linguistic line accompanied by a more scientific cartography; in 1815, the French government produced the first map in the world drawn on a geodesic triangulation. Propagandists of the ancient regime and men of the revolution both claimed for France its 'natural frontiers', first in the form of the four rivers (Rhône, Saône, Meuse, and Scheldt), and then invoking Strabo and Caesar, the quadrilateral limits of ancient Gaul, the Rhine, Alps and the Ocean. To these myths the revolutionaries added another, the exact correspondence between 'natural' and 'national' frontiers before it was entirely clear who or what constituted the nation. It took another century before the administrative, customs and juridical instruments were sufficiently firmly established to enforce the claim (Bergeron and Revel 1989). In the process, the national territorial boundary was shaped as much by the local, frontier societies as by the central state, and as a result, the inhabitants retained their sense of place, even as they accepted their identity with the nation (Sahlins 1989).

The French (or Jacobin) model that the frontiers of a state ought to correspond to the natural and the national with one language and one culture was internationalized, at least for France, in the treaties of Vienna of 1815. Subsequently, it was widely diffused throughout Europe, Latin America and—with the participation of the British—Africa and the Middle East with mixed results.

One subtype of the French model emerged in the late nineteenth and early twentieth centuries in the small successor states of central, southeastern and eastern Europe. But there the prolonged existence of multicultural societies (Habsburg, Ottoman, and Russian) shaped by dynastic interests delayed into the twentieth century the shift from historic to linguistic (national) frontiers with dire consequences. By this time, internal migration, colonization, forced resettlement, and expulsions had created a kaleidoscope of peoples which could not be neatly compartmentalized by drawing linear frontiers.

Redrawing the map of southeastern Europe as the Ottoman power gradually receded was a cartographic nightmare, the result of a series of unhappy compro-

mises between the conflicting interests of the great powers and the aspirations of the subjugated peoples striving not only for independence but also for frontiers based on historicist claims—Greater Serbia, Greater Croatia, Greater Romania, Greater Hungary, Greater Bulgaria, Greater Greece, etc.—claims that were mutually irreconcilable.

German frontiers represent another subtype, sharing elements of the consolidated state and the dynamic moving frontier. In the north, west and south, the German frontiers evolved along the lines of the French model. But in the vast eastern space of the German-Slavic encounter, the discrepancy between sovereignty and settlement remained unresolved for 1000 years. For German statesmen, publicists and intellectuals from the eighteenth to the mid twentieth century, the task of consolidating a fixed state frontier with the Slavic east became a profoundly disturbing sociopsychological as well as political problem producing its share of historical myths (Demandt 1990).

3. The Dynamic Type of Frontiers

The second type of frontier may be characterized as an advancing line of settlements engaged primarily in agricultural or mixed economy, confronting a seminomadic and/or technologically less developed culture. In the modern period, three subtypes may be distinguished: the United States, the former British dominions and the Imperial Russian and Chinese Empires. The United States was the only former British colony where, once it had achieved independence, the frontier experience became both a national myth and a dominant historiographic theme. The other former British dominions constitute another subtype, where geographical and political factors placed the idea of the frontier in a different context. The continental expansion of Russia and China constitute a third subtype where the advance of the two societies brought them into direct contact and competition with one another.

3.1 The American Frontier as a Subtype

Neither the revisionists in American historiography nor the comparativists on an international scale have been able to free themselves entirely from the Turnerian concept of the frontier, even as they demolished its tenets one by one. What Turner did was to establish the frontier as a historical problem and to ask what it meant to be an American. The complexity of his approach guaranteed its longevity even as a foil. He combined elements of economic geography, social structure and ideology even though he was not always clear or consistent in drawing the relationship between them. His brilliant essays added up to something less than a full-blown theory and left the way open for

multiple interpretation, extensive revision and critical refutation. Like many great theses in historiography, its importance lies more in the literature it generated than in its original propositions. The debate over his legacy took place in two arenas, the national (American) and the comparative (global).

In American historiography, the revisions and reformulations of the Turner thesis radiate around two conceptual poles: new methodologies, particularly anthropology and environmental studies, and fresh interpretive perspectives reflecting the shifting concerns of a restless society in search of itself. At present, as in so many other areas, the task of charting the currents of the debate has itself become a minor cottage industry (Nash 1991). Turner's frontier as an advancing line of settlement has given way, in the work of the regionalists, to a multiplicity of locations. But Turner's tendency to conflate the frontier as a line and 'the Great West' as a region continues to be a source of confusion. There has also been a trend away from spatial to cultural definitions of the frontier, that is, from place to process. Turner's concept of the frontier as dividing 'savagery from civilization' or, alternatively, as a forward movement into 'empty lands' has, under the influence of anthropologists, yielded to the idea of the frontier as a zone of interaction involving two or more previously distinct societies that engendered a variety of cultural and commercial 'frontier exchanges' between Native Americans and colonists. Turner's concept of the West as an environment that bred rugged individualism and privileged self-reliance was challenged by historians and sociologists who gave equal billing to social collectivities, and uncovered a darker side of the frontiersman in the persistence of a gun culture and the cult of violence. Environmentalist historians exploded Turner's arcadian picture of natural harmony on the frontier by exposing the depredations, manipulations and 'species shifting' impact of the western advance (Cronon et al. 1992). Turner's insistence on the defining role of the frontier experience in building American democratic institutions was counterbalanced by the contributions of urban life and the antecedent values the pioneers brought with them. At one end of the historiographical spectrum, Turner's thesis lies in ruins and even the term 'frontier' is rejected or excoriated. At the other end, attempts have been made to synthesize the newer ideas on race, class, gender and region, in order to rescue once again the importance of the frontier in American history. The persistence of interest in Turner, marked more often by invoking his 'ghost' or 'spirit' rather than the body of his work, lies in large part with the power of his explanatory theme to explain American national identity. Pluralistic interpretations have never conveyed the same emotional force. Turner's thesis has passed into the realm of historical myth and nowhere is this more evident than in the realm of mass culture.

The commercialization of American history found

the perfect vehicle in the frontier. The transformation of the frontier into a popular myth and its subsequent permutations was a function of changes in both the medium and the message. New social and environmental concerns left their mark on Turnerian stereotypes. In film and television, the heroic cowboy in the white hat, the intrepid farmer fighting off the Indians, the cavalry arriving at the last minute became more ambiguous figures. Black cowboys and troopers, environmentally sensitive Native Americans, even women gun-slingers challenged the dominant white male image. The Marlboro Man made his last stand on American billboards.

3.2 *The British Empire as a Subtype*

An analysis of subtypes of the dynamic frontier cannot be treated without reference to the comparative approach largely inspired by, if not always accepting, the Turner thesis and its revisions. The influence of the debate over the Turner thesis has been felt most keenly in the historiographies of the former dominions of the British Empire, and the Russian and Chinese Empires. Historians of North America and South Africa have produced a particularly well-designed comparative analysis of their respective frontiers (Lamar and Thompson 1981). It sought to achieve theoretical clarity and consistency by establishing a universal definition rather than adopting the USA experience or Turner thesis as a model, even though the original inspiration remained clear. Three elements in defining frontier societies are considered essential: first, the geographical-territorial, conceived as a zone rather than a line with different carrying capacities, attractiveness and resources; second, the cultural, in terms of the interpenetration between two previously distinct societies, one indigenous and the other intrusive; and third, the operational, the process of opening and closing, the former dated by the arrival of representatives of the intrusive society and the latter, more complex stage because it was variable and potentially reversible, as it indeed proved to be in South Africa, by the establishment of a single political hegemony through extermination (Tasmania), expulsion (Trans-Mississippi), subjugation (South Africa) or incorporation (Brazil). Specific comparisons of the USA and South Africa reveal more differences than similarities, although the white communities in both mythologized the frontier experience as formative. Before the recent reversal in South Africa, the critical difference appeared to be the different relationship of the indigenous people in the USA and South Africa to capitalism.

Comparative studies, with Turnerian overtones, of the frontier experience in Canada, Australia and New Zealand suggest that despite superficial similarities with the USA, geographical factors and the prolonged tutorship of the metropolitan center in the UK

prevented the rise of a mythology of frontier exceptionalism (Nugent 1994). In Canada there were close parallels between the farming frontier in Ontario and Saskatchewan, and in North Dakota and Montana. But the expansion west was delayed by the physical barrier of the Canadian shield. When it came, it was promoted by Royal Governors established national police (the Royal Mounties) before mass immigration and negotiated settlements with the Blackfoot tribes thus heading off the local violence and large-scale white-native American wars that characterized the American west. To an even greater extent, the geography of Australia, dominated by a vast arid region in the center, 'a hinterland but no heartland,' relegated the bulk of the population to the cities on the east and south coast while royal land policy did not favor the small farmer in competition with the large sheep ranges. Yet even here there are regional similarities between the semi-arid fringes of California and Victoria. And Australia had its frontier folk hero in the form of the bush-ranger, although he was connected to the convict life rather than the freebooter born in an open society. The Northwest Frontier of India was the only zone of encounter between the British colonial expansion and a seminomadic population where the odds were rated even. The character of this military frontier resembled more that of the great older empires, the Roman, Chinese and Ottoman, confronting a persistent threat to security along one or several of their peripheries. The British inherited the problem from their predecessors the Moghuls, who, like the founders of several Chinese dynasties, were originally 'eastern barbarians' before they came to settle in the fertile river valleys of the south. Although the Northwest Frontier could hardly be said to have dominated British colonial politics, there was a constant tension between the proponents of a forward policy to anchor the frontier on the Hindu Kush and the advocates of 'masterly inactivity' (Fraser-Tytler 1950). And there was concern and debate over the Russian advance behind the tribes. The British crossed the frontier in force three times to fight the Afghans without settling much of anything. It was a classical military frontier zone exhibiting sporadic warfare, trade, smuggling, movement of peoples and a mythology of heroism immortalized by Kipling. Today, the Northwest frontier between Afghanistan and Pakistan has lost nothing of this except its glamour.

3.3 *The Imperial Russian and Chinese Empires as a Subtype*

The long, porous and shifting Eurasian frontiers of Russia and China have also attracted comparativists who employ Turner's thesis and its revisions as a point of departure in their analysis. However, even before Turner, nineteenth-century Russian historians, like S. M. Solov'ev and V. O. Kliuchevskii, stressed the

central role of colonization and frontier in shaping the history of their own country, and drew from many of the same formative influences as their American counterpart (Bassin 1993). By contrast, though, their interpretations emphasized the negative effects: the draining of resources from the center, the unfavorable conditions of climate and soil, the threat of nomadic incursions. Historians in the early Soviet period added to this list the exploitation of the conquered people. A strong reversal set in only after World War II when the new state school of Soviet historians extolled the Russian peasant colonizer as a heroic figure and employed the specialized term 'merging' (sblizhenie), a form of assimilation, to signify the civilizing mission of the Russian people. A third turn has been taken only recently by the emerging regional (Siberian) school of historians who, in readjusting their focus to the specific features and values of the periphery at odds with those of the center have, nonetheless, acknowledged their debt to Turner.

Meanwhile, Western scholars, mainly Americans and British, assumed the mantle of the pre-revolutionary Russian historians and invoked Turner's thesis, in order to draw parallels and contrasts between the Russian conquest and colonization of Siberia and the American western expansion. The earliest attempts to apply the Turner thesis to the Russian experience lagged behind the extensive revisionist literature in the USA that had already forced a re-evaluation of its central tenets. However, since the 1980s a new generation of American and British students of Russian colonization have incorporated new ideas borrowed from the historiography of the USA, rather than from direct comparison with the American experience. Multiple locus and process have replaced conquest and resistance (Barrett 1995).

On the Chinese side, after more than six decades the dominant figure in frontier studies remains Owen Lattimore. His earliest work combined the keen first-hand observations of the linguistically competent, experienced traveler with historical and anthropological training (Lattimore [1940] 1951). Studded with comparative insights (Roman Empire, American West, Northwest Frontier Province), its central thesis was to establish that frontiers, as distinct from boundaries, were the outer limit of zones on the margins of a socioeconomic systems that represented their 'optimal limit of growth.' While admitting that Turner was 'an acute observer' he added that 'what he saw so clearly he saw standing on his head. In large measure, when he thought he saw what the frontier had done to society, he was really seeing what society did to the frontier' (Lattimore 1962). By comparing the social and economic character of Chinese and Russian empires, he was able to distinguish between their different 'margins of expansion.' The study of Chinese frontiers has increasingly focused on the long-term sinefication of the frontier provinces: Manchuria, Inner Mongolia, Sinkiang, and Tibet.

4. Symbolic Frontiers

The frontiers of symbolic geographies focus on the meeting ground of distinctive civilizations. The attempt to define and defend these frontiers was a feature of both the Greco-Roman and Chinese empires, the first to identify the boundary between civilized and barbarian. It was continued by the Islamic and Christian (European) traditions which were contemporaneous and contiguous hence more conflictual. But religious fissures (Sunni-Sh'ia on the one hand and Eastern Orthodox and Latin on the other) created intercivilizational frontiers that proved just as difficult to cross. Within Europe, a significant shift in symbolic geography occurred in the eighteenth century when the philosophers realigned the main cultural divide between the enlightened and the backward from a south–north to an west–east axis (Wolff 1994). But the exact demarcation line was much disputed, especially when confused with the symbolic frontier between Europe and Asia as Metternich's famous quip, 'Asia begins at the Landstrasse,' well illustrates.

In the twentieth century, the Soviet rhetoric of frontiers infused the ancient 'sacred' concept with an exclusionist ideological meaning enabling another rhetorician, Churchill, to coin the phrase 'Iron Curtain,' which retained its symbolic power long after events demonstrated its imprecision and inadequacy. The collapse of the Soviet Union and the Communist bloc overturned the theory of stable state frontiers, which had dominated American political science from 1950–90 (Anderson 1997). It also spurred a recurrent debate about the cultural line dividing western and eastern Europe and the West from other civilizations. Both during and after the Cold War there were divergent attempts to locate the historic origins of the dividing lines. Some held that Latin and Orthodox Christianity constituted a single civilization for 1,000 years until the communists rang down the curtain and arbitrarily split it along the Oder–Neisse line (Halecki 1950). Others, like Arnold Toynbee, drew the frontiers of civilizations along religious lines and argued that totalitarian Russia had its roots in Eastern Orthodoxy. After 1989, several émigré and former dissident intellectuals in the region resurrected the idea of a different more westerly and long-established cultural frontier between 'central Europe,' which was equated with liberal democracy and the spirit of capitalism, from 'eastern Europe' associated with the authoritarian state and cultural backwardness. Only in this context is it possible to measure the enormous psychological impact on the people of the region of being accepted or rejected for membership in NATO or the European Union, real institutions which, nevertheless, constitute the new symbolic frontiers of civilization.

See also: Area and International Studies in the United States; Intellectual Trends; Borders, Anthropology of;

Civilization, Concept and History of; Colonization and Colonialism, History of; Historiography and Historical Thought: Classical Period (Especially Greece and Rome); Historiography and Historical Thought: Current Trends; Historiography and Historical Thought: East Asia; Historiography and Historical Thought: Islamic Tradition; Historiography and Historical Thought: Modern History (Since the Eighteenth Century); Nations and Nation-states in History; State Formation; State, History of

Bibliography

- Anderson M 1997 *Frontiers: Territory and State Formation in the Modern World*. Polity Press, Malden, MA
- Barrett T 1995 Lines of uncertainty. The frontiers of the north Caucasus. *Slavic Review* 54(3): 578–601
- Bassin M 1993 Turner, Solov'ev and the 'Frontier Hypothesis': The nationalist significance of open spaces. *Journal of Modern History* 3: 473–511
- Bergeron L, Revel J (eds.) 1989 *L'Espace Français*. Seuil, Paris
- Cronon W, Miles G, Gitlin J 1992 *Under an Open Sky: Rethinking America's Western Past*. W. W. Norton, New York
- Demandt A (ed.) 1990 *Die Grenzen in der Geschichte Deutschlands*. Verlag C H Beck, Munich, Germany
- Foucher M 1991 *Fronts et Frontières. Un Tour du Monde Géopolitique*. New edn. Fayard, Paris
- Fraser-Tytler WK 1950 *Afghanistan. A Study of Political Developments in Central Asia*. Oxford University Press, London
- Godlewka A, Smith N 1994 *Geography and Empire*. Blackwell, Oxford
- Halecki O 1950 *The Limits and Divisions of European History*. Sheed and Ward, London
- Kaser K 1997 *Freier Bauer und Soldat. Die Militarisierung der agrarischen Gesellschaft an der kroatisch-slawonischen Militärgrenze (1535–1881)*. Verlag, Vienna
- Khazanov A M 1994 *Nomads and the Outside World*, 2nd ed. University of Wisconsin Press, Madison, WI
- Lamar H, Thompson L (eds.) 1981 *The Frontier in History: North America and Southern Africa Compared*. Yale University Press, New Haven, CT
- Lattimore O [1940] 1951 *Inner Asian Frontiers of China*. American Geographical Society, New York
- Lattimore O 1962 *Studies in Frontier History. Collected Papers, 1928–1958*. Oxford University Press, London
- Nash G D 1991 *Creating the West: Historical Interpretations, 1880–1990*, 1st edn. University of New Mexico Press, Albuquerque, NM
- Nugent W 1994 Comparing wests and frontiers. In: Milner C A II et al. (eds.) *Oxford History of the American West*. Oxford University Press, New York pp. 803–33
- Roksandic D (ed.) 1998 *Microhistory of the Triplex Confinium*. Institute on Southeastern Europe, Budapest, Hungary
- Sahlins P 1989 *Boundaries. The Making of France and Spain in the Pyrenees*. University of California Press, Berkeley, CA
- Turner F J [1893] 1921 *The Frontier in American History*. Henry Holt, New York
- Whittaker C R 1994 *Frontiers of the Roman Empire. A Social and Economic Study*. Johns Hopkins University Press, Baltimore, MD

Wittek P 1938 *The Rise of the Ottoman Empire*. Royal Asiatic Society, London

Wolff L 1994 *Inventing Eastern Europe. The Map of Civilization on the Mind of the Enlightenment*. Stanford University Press, Stanford, CA

A. J. Rieber

Functional Brain Imaging

Since the 1990s the field of cognitive neuroscience has emerged as a very important growth area in neuroscience. Cognitive neuroscience combines the experimental strategies of cognitive psychology with various techniques to actually examine how brain function supports mental activities. Leading this research in normal humans are the new techniques of functional brain imaging: positron emission tomography (PET) and magnetic resonance imaging (MRI) along with event related potentials (ERP's) obtained from electroencephalography (EEG) or magnetoencephalography (MEG). This article focuses on PET and MRI.

1. Antecedents of Functional Brain Imaging

1.1 Neuropsychology

Early discussions of the 'mind-brain problem' treated the brain largely as a 'black box.' Then, in 1861, French surgeon and anthropologist Pierre Paul Broca described a clear relationship between a patient's difficulty speaking and an injury to a specific part of the patient's brain due to a stroke. Since this seminal observation, a vast body of scientific literature has accumulated implicating various parts of the human brain in specific aspects of human behavior, including language. This approach to the study of human brain function has achieved a remarkable level of sophistication and represents an important tool in neuroscience.

The features of brain organization arising from the study of patients with brain injury nevertheless raise some questions of interpretation. The size and location of brain injury varies greatly from patient to patient, making a precise correlation between damage to a particular area of the brain and the function normally served by that area sometimes difficult to determine. Furthermore, each patient may be assumed to have some features of brain organization that are unique to him or her. Finally, it remains uncertain whether one can simply attribute a lost or disrupted function to a particular area of injury. Because of the interconnected nature of areas of the brain, injury in one area is likely to have effects on other areas that cannot necessarily be predicted from the location and size of the injury. Thus, as valuable as our insights concerning the

organization of the human brain have been, from the study of patients with stroke and other types of brain injury, it has remained an open question exactly how this information relates to the normal organization of the human brain.

Only recently have scientists interested in this question had the opportunity to explore it analytically—to peer inside the black box as it functions normally. This ability stems from the developments in imaging technology since the 1980s, most notably PET and MRI. These techniques can now capture localized changes in blood flow in the normal human brain associated with behaviorally induced changes in brain cellular activity (Raichle 1998).

1.2 Blood Flow and Brain Function

The quest for an understanding of the functional organization of the normal human brain, using techniques to assess changes in brain circulation, has occupied mankind for more than a century (for a historical review see Raichle (2000)). An enduring theme throughout this work has been that changes in the cellular activity of the normal brain of humans and laboratory animals is invariably accompanied by changes in the blood flow to the region of the brain affected.

1.2.1 Early quantitative blood flow measurements in humans. At the close of World War II, Seymour Kety and co-worker opened the first chapter in quantitative studies of brain circulation and metabolism in humans and laboratory animals. Working with Lou Sokoloff and others, Kety developed the first quantitative methods for measuring *whole* brain blood flow and metabolism in humans. The introduction of an *in vivo* tissue autoradiographic measurement of regional blood flow in laboratory animals by Kety's group (Kety 1960, Landau et al. 1955) provided the first glimpse of quantitative changes in blood flow in the brain related directly to brain function.

Soon after Kety introduced quantitative methods for measuring whole brain blood flow and metabolism in humans, David Ingvar, Neils Lassen, and their Scandinavian colleagues introduced methods applicable to man that permitted regional blood flow measurements to be made using scintillation detectors arrayed like a helmet over the head (Lassen et al. 1963). They demonstrated directly in normal human subjects that blood flow changed regionally during changes in brain functional activity. The first study of functionally-induced regional changes in blood flow using these techniques in normal humans was actually reported by Ingvar and Risberg at an early meeting on brain blood and metabolism (Ingvar and Risberg 1965) and was greeted with cautious enthusiasm and a clear sense of its potential importance for studies of

human brain function by Kety (Kety 1965). However, despite many studies of functionally induced changes in regional cerebral blood, which followed, this approach was not embraced by most neuroscientists or cognitive scientists at the time. It is interesting to note that this indifference was to disappear almost completely in the 1980s, a subject to which the article will return.

2. Introduction of Modern Imaging Techniques

The modern era of medical imaging began in the early 1970s when the world was introduced to a remarkable technique called X-ray computed tomography, now known as X-ray CT or just CT (for review see Kevles 1997). South African physicist Allan M. Cormack and British engineer Sir Godfrey Hounsfield independently developed its principles but it was Hounsfield who constructed the first scanner successfully.

2.1 X-ray Computed Tomography (CT)

CT takes advantage of the fact that different tissues absorb varying amounts of X-ray energy. The denser the tissue, the more it absorbs. A highly focused beam of X-rays traversing the body will exit at a reduced energy level depending on the tissues and organs through which it passes. A beam of X-rays passed through the body at many different angles through a plane provides sufficient information to reconstruct a picture of that body section. It was crucial to the development of CT that clever computing and mathematical techniques emerged for processing the vast amount of information necessary to create the images themselves. Without the availability of sophisticated computers, the task would have been impossible to accomplish.

CT had two consequences. First, it changed the practice of medicine forever, because it was much superior to standard X-ray techniques. For the first time, physicians could safely and effectively view living human tissue such as the brain with no discomfort or risk to the patient. Second, CT immediately stimulated scientists and engineers to consider alternative ways of creating images of the body's interior using similar mathematical and computer strategies for image construction. These efforts went beyond the picture of human anatomy to focus on function.

2.2 Positron Emission Tomography (PET)

One of the first groups to be intrigued by the possibilities opened by CT consisted of experts in tissue autoradiography, a method used for many years in animal studies to investigate organ metabolism, biochemistry and blood flow. In tissue autoradio-

graphy, a radioactively labeled compound is injected into a vein. After the compound has accumulated in the organ under investigation, the animal is sacrificed and the organ (e.g., the brain) removed for study. The organ then is sectioned carefully, and the individual slices are laid on a piece of film sensitive to radioactivity. Much as the film in a camera records a scene as it was viewed originally, this X-ray film records the distribution of radioactively labeled compound in each slice of tissue.

When the X-ray film is developed, scientists have a picture of the distribution of radioactivity within the organ and hence can deduce the organ's specific functions. The radioactive compound injected determines the type of information. A radioactively labeled form of glucose, for example, measures brain metabolism, because glucose is the primary source of energy for the cells of the brain. Central to functional brain imaging with PET is the measurement of brain blood flow, which is accomplished by the injection of radioactively labeled water.

Investigators adept at tissue autoradiography were fascinated when CT was introduced. They suddenly realized that if they could reconstruct the anatomy of an organ by passing an X-ray beam through it, as CT did, they could also reconstruct safely the distribution of a previously administered radioisotope. One had simply to measure the emission of radioactivity from the body section. This realization was the birth of autoradiography of living human subjects.

One crucial element in the evolution of human autoradiography was the choice of radioisotope. Workers in the field selected a class of radioisotopes that emit positrons, which resemble electrons except that they carry a positive charge. A positron produced with the tissue almost immediately combines with a nearby electron. The positron and the electron annihilate one another in this interaction, emitting two high-energy gamma rays in the process. Since the gamma rays travel in nearly opposite directions, radiation detection devices arrayed in a circle around the organ of interest can detect the pairs of gamma rays and, with the aid of computers, locate their origin with remarkable precision. The crucial role of positrons in human autoradiography gave rise to the name positron emission tomography or PET (Phelps et al. 1975, Ter-Pogossian et al. 1975).

2.3 Magnetic Resonance Imaging (MRI)

Finally, another technology emerged contemporaneously with PET and CT. This was MRI. MRI is based upon yet another set of physical principles that have to do with the behavior of hydrogen atoms or protons in a magnetic field. These principles were discovered independently by Felix Block and Edward Purcell and colleagues in 1946 and expanded to imaging by Paul Lauterbur in 1973 (for review see Kevles 1997).

Initially MRI provided superb anatomical information but inherent in the data also was important metabolic and physiological information. An opening for MRI in the area of functional brain imaging emerged when it was discovered that during changes in neuronal activity there are local changes in the amount of oxygen in the tissue (Raichle 1998). By combining this observation with a much earlier observation by Pauling and Coryell (1936) that changing the amount of oxygen carried by hemoglobin changes the degree to which hemoglobin disturbs a magnetic field, Ogawa et al. (1990) were able to demonstrate that *in vivo* changes in blood oxygenation could be detected with MRI. The MRI signal (technically known as T2* or 'tee-two-star') arising from this unique combination of brain physiology (Raichle 1998) and nuclear magnetic resonance physics (Pauling and Coryell 1936, Thulborn et al. 1982) became known as the blood oxygen level dependent or BOLD signal (Ogawa et al. 1990). There followed quickly several demonstrations of BOLD signal changes in normal humans during functional brain activation (Kwong et al. 1992, Ogawa et al. 1992) giving birth to the rapidly developing field of functional MRI or fMRI.

It is important to keep in mind that when a BOLD signal is detected blood flow to a region of brain has changed out of proportion to the change in oxygen consumption. When blood flow changes more than oxygen consumption, in either direction, there is a reciprocal change in the amount of deoxyhemoglobin present locally in the tissue changing the local magnetic field properties. As you will see both increases and decreases occur in the BOLD signal in the normal human brain.

3. Functional Brain Imaging

3.1 The Strategy of Functional Brain Imaging

As a result of collaboration among neuroscientists, imaging scientists, and cognitive psychologists, a distinct behavioral strategy for the functional mapping of neuronal activity emerged. This combination of disciplines and the strategies resulting from it probably account for the remarkable success and the resulting acceptance of functional brain imaging in cognitive neuroscience. The general strategy of functional brain imaging is based on a concept introduced by the Dutch physiologist Franciscus C. Donders in 1868 (reprinted in Donders (1969)). Donders proposed a general method to measure thought processes based on a simple logic. He subtracted the time needed to respond to a light (say, by pressing a key) from the time needed to respond to a particular color of light. He found that discriminating color required about 50msec. In this way, Donders isolated and measured a mental process for the first time by subtracting a *control state* (i.e.,

responding to a light) from a *task state* (i.e., discriminating the color of the light).

One criticism of this approach has been that the time necessary to press a key after a decision to do so has been made is affected by the nature of the decision process itself. By implication, the nature of the processes underlying key press, in this example, may have been altered. Although this issue (known in cognitive science jargon as the *assumption of pure insertion*) has been the subject of continuing discussion in cognitive psychology, it finds its resolution in functional brain imaging, where changes in any process are signaled directly by changes in observable brain states. Events occurring in the brain are not hidden from the investigator as in the purely cognitive experiments. Careful analysis of the changes in the functional images reveals whether processes (e.g., specific cognitive decisions) can be added or removed without affecting ongoing processes (e.g., motor processes). Clearly, some areas of the brain active at one stage in a hierarchically designed paradigm can become inactive as task complexity is increased. Even more surprising, areas of the brain active in the resting state can also become inactive during certain cognitive states (Raichle et al. 2001). While changes of this sort are hidden from the view of the cognitive scientist they become obvious when brain imaging is employed.

A final caveat with regard to certain cognitive paradigms is that the brain systems involved do not necessarily remain constant through many repetitions of the task. While simple habituation might be suspected when a task is tedious, this is not the issue referred to here. Rather, when a task is *novel* and, more importantly, conflicts with a more habitual response to the presented stimulus major changes can occur in the systems allocated to the task. Such changes have both practical and theoretical implications when it comes to the design and interpretation of cognitive activation experiments. Functional brain imaging obviously provides a unique perspective that is unavailable in the purely cognitive experiment.

3.2 The Assumptions of Functional Brain Imaging

It is important to point out that the underlying assumptions of current brain-mapping studies with PET and fMRI do not embrace a modern version of phrenology. The phrenologists of the nineteenth century posited that single areas of the brain, often identified by bumps on the skull, uniquely represented specific thought processes and emotions. In contrast, modern thinking posits that single areas of the brain each contribute quite simple mental operations that form the elementary components of the observable behaviors. Observable behaviors and thought processes emerge through the cooperative interactions of many such areas. So, just as diverse instruments of a large orchestra are played in a coordinated fashion to

produce a symphony, a group of diverse brain areas, each performing unique elementary and unique mental operations, work together in a cooperative fashion to produce human behavior. The prerequisite for such analyses is the conviction that complex behaviors can be broken down into sets of constituent mental operations. This is not a new view. In 1909, the great neuroanatomist Korbinian Brodmann commented that ‘With regard to the infinite variety and richness of the psyche one should rather envisage the situation that in each particular case supposed “elementary functional loci” are active in differing numbers, in differing degrees and in differing combinations. One must further accept that specific complex processes occur mainly in one locality and others mainly in another. One must therefore also assume a certain regional preference for higher activities, sometimes more in occipital and temporal areas, sometimes more in frontal. *Such activities are, however, always the result (and not merely the sum) of the function of a large number of suborgans distributed more or less widely over the cortical surface; they can never be the product of a morphologically or physiologically independent “center”*’ (Brodmann 1909, 1994).

3.3 The Future of Functional Brain Imaging

Functional brain imaging took a major step forward with the introduction of fMRI. This resulted not only from the improved spatial resolution of the technique and the wide availability of MRI scanners, many now appearing outside of hospitals in departments of psychology, but also the ability to obtain information on the response of the brain to individual stimuli (Buckner et al. 1996). Earlier studies with PET required that stimulus types be ‘blocked’ in order to accommodate the time necessary to make a single measurement of blood flow (usually about 1 m). With fMRI stimulus types can be randomized in ways that are identical to those used traditionally by cognitive psychologists. Thus, following the completion of the study, image data can be sorted according to stimulus type or response characteristics (Brewer et al. 1998, Wagner et al. 1998), providing a much more sophisticated analysis of brain function in relation to behavior or stimulus type.

Even with the introduction of fMRI we still do not have an adequate description of the temporal dimension of responses occurring within the brain. We know from direct measurements of the responses of neurons in the brains of laboratory animals that information is processed among areas of the brain with time differences in the range of tens of milliseconds. This is far too fast to be captured with fMRI. Electrical techniques such as the EEG and MEG do capture this temporal dimension but do so with very poor spatial resolution. The hope, therefore, is to combine eventually the spatial dimension of responses

captured with fMRI with the temporal dimension provided by EEG or MEG. This combination of techniques remains a technically challenging problem but progress has been made (Liu et al. 1998, Rosen et al. 1998).

Those readers interested in a general survey of the current state of functional brain imaging work might wish to consult a variety of new books that are appearing regularly (Moonen and Bandettini 1999; Toga and Mazziotta 1996, 2000).

See also: Cortical Activity; Differential Optical Imaging; Magnetoencephalography; MRI (Magnetic Resonance Imaging) in Psychiatry; Neuroanatomical Methods.

Bibliography

- Brewer J B, Zhao Z, Desmond J E, Glover G H, Gabrieli 1998 Making memories: Brain activity that predicts how well visual experience will be remembered. *Science* **281**: 1185–7
- Brodmann K 1909 *Vergleichende Lokalisationlehre der Grosshirnrinde*. Barth, Leipzig, Germany
- Brodmann K 1994 *Localisation in the Cerebral Cortex*. Smith-Gordon, London
- Buckner R L, Bandettini P A, O’Craven K M, Savoy R L, Petersen S E, Raichle M E, Rosen B R 1996 Detection of cortical activation during averaged single trials of a cognitive task using functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America* **93**: 14878–83
- Donders F C 1969 On the speed of mental processes. *Acta Psychologica* **30**: 412–31
- Ingvar G H, Risberg J 1965 Influence of mental activity upon regional cerebral blood flow in man. *Acta Neurological Scandinavica* **Suppl 14**: 183–6
- Kety S 1960 Measurement of local blood flow by the exchange on an inert diffusible substance. *Methods Medical Research* **8**: 228–36
- Kety S 1965 Closing comments. *Acta Neurological Scandinavica* **Suppl 14**: 197
- Kevles B H 1997 *Naked To The Bone. Medical Imaging in the Twentieth Century*. Rutgers University Press, New Brunswick, NJ
- Kwong K K, Belliveau J W, Chesler D A, Goldberg I E, Weiskoff R M, Poncelet B P, Kennedy D N, Hoppel B E, Cohen M S, Turner R, Cheng H M, Brady T J, Rosen B R 1992 Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences of the United States of America* **89**: 5675–9
- Landau W M, Freygang W H J, Roland L P, Sokoloff L, Kety S S 1955 The local circulation of the living brain: values in the unanesthetized and anesthetized cat. *Transactions of the American Neurology Association* **80**: 125–9
- Lassen N A, Hoedt-Rasmussen K, Sorensen S C, Skinhoj E, Cronquist B, Bodforss E, Ingvar D H 1963 Regional cerebral blood flow in man determined by Krypton-85. *Neurology* **13**: 719–27

- Liu A K, Belliveau J W, Dale A M 1998 Spatiotemporal imaging of human brain activity using functional MRI constrained magnetoencephalography data: Monte Carlo simulations. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 8945–50
- Moonen C T, Bandettini P A 1999 *Functional MRI*. Springer, New York
- Ogawa S, Lee T M, Kay A R, Tank D W 1990 Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America* **87**: 9868–72
- Ogawa S, Tank D W, Menon R, Ellermann J M, Kim S-G, Merkle H, Ugurbil K 1992 Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America* **89**: 5951–5
- Pauling L, Coryell C D 1936 The magnetic properties and structure of hemoglobin, oxyhemoglobin and carboxyhemoglobin. *Proceedings of the National Academy of Sciences of the United States of America* **22**: 210–6
- Petersen S E, Fox P T, Posner M I, Mintum M A, Raichle M E 1988 Positron emission tomographic studies of the cortical anatomy of single-word processing. *Nature* **331**: 585–9
- Petersen S E, Fox P T, Posner M I, Mintun M A, Raichle M E 1989 Positron emission tomographic studies of the processing of single words. *Journal of Cognitive Neuroscience* **1**: 153–70
- Phelps M E, Hoffman E J, Mullani N A, Ter-Pogossian M M 1975 Application of annihilation coincidence detection to transaxial reconstruction tomography. *Journal of Nuclear Medicine* **16**: 210–24
- Raichle M E 1998 Behind the scenes of functional brain imaging: A historical and physiological perspective. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 765–72
- Raichle M E 2000 A brief history of human functional brain mapping. In: Toga A W, Mazziotta J C (eds.) *Brain Mapping. The Systems*. Academic Press, San Diego, pp. 33–75
- Raichle M E, Fiez J A, Videen T O, MacLeod A M, Pardo J V, Fox P T, Petersen S E 1994 Practice-related changes in human brain functional anatomy during nonmotor learning. *Cerebral Cortex* **4**: 8–26
- Raichle M E, MacLeod A M, Snyder A Z, Powers W J, Gusnard D A, Shulman G L 2001 A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 676–82
- Rosen B R, Buckner R L, Dale A M 1998 Event-related function MRI: Past, present and future. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 773–80
- Talairach J, Tournoux P 1988 *Coplanar Stereotaxic Atlas of the Human Brain*. Thieme Medical Publishers, New York
- Ter-Pogossian M M, Phelps M E, Hoffman E J, Mullani N A 1975 A positron-emission tomograph for nuclear imaging (PETT). *Radiology* **114**: 89–98
- Thulborn K R, Waterton J C, Matthews P M, Radda G K 1982 Oxygenation dependence of the transverse relaxation time of water protons in whole blood at high field. *Biochimica et Biophysica Acta* **714**: 265–70
- Toga A W, Mazziotta J C 1996 *Brain Mapping. The Methods*. Academic Press, New York
- Toga A W, Mazziotta J C 2000 *Brain Mapping. The Systems*. Academic Press, New York

Wagner A D, Schacter D L, Rotte M, Koutstaal W, Maril A, Dale A M, Rosen B R, Buckner R L 1998 Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity. *Science* **281**: 1188–91

M. E. Raichle

Functional Data Analysis

Most statistical analyses involve one or more observations taken on each of a number of individuals in a sample, with the aim of making inferences about the general population from which the sample is drawn. In an increasing number of fields, these observations are curves or images. Curves and images are examples of functions, since an observed intensity is available at each point on a line segment, a portion of a plane, or a volume. For this reason, we call observed curves and images ‘functional data,’ and statistical methods for analyzing such data are described by the term ‘functional data analysis’ (FDA), coined by Ramsay and Dalzell (1991). Though the individual methods and techniques are important, functional data analysis is also a general way of thinking, where the basic unit of information is the entire observed function rather than a string of numbers. A few illustrations will be offered in this article, but Ramsay and Silverman (1997) may be consulted for many more examples and methods, for consideration of the underlying philosophy, and for further references to the literature.

The goals of functional data analysis are essentially the same as for other branches of statistics, and include the following:

- (a) to represent and transform the data in ways that aid further analysis,
- (b) to display the data so as to highlight various characteristics,
- (c) to study important sources of pattern and variation among the data, and
- (d) to explain variation in an outcome or dependent variable by using input or independent variable information.

We shall illustrate the nature of functional data, these goals, and FDA tools available or under development, through a series of examples.

1. Human Growth Data: Looking at Velocity and Acceleration

Figure 1 shows the growth patterns of 10 Swiss boys involved in the Zurich Longitudinal Growth Study (Falkner 1960). Each boy’s height is measured at 29 separate ages, but it is more instructive to think of this as a sample of 10 growth curves or functions than as a set of 290 numerical observations. Of particular interest in the study of growth is the way that growth speeds up and slows down at various stages of

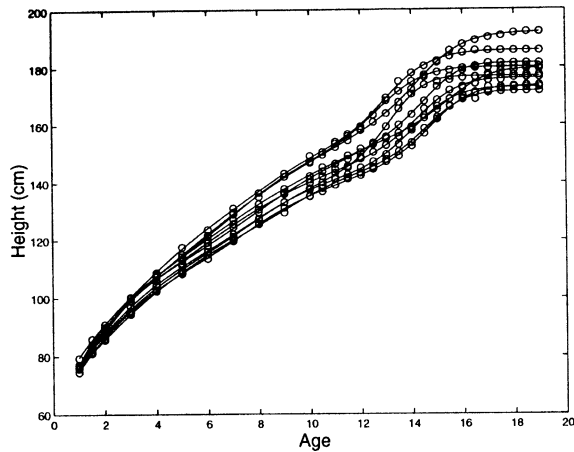


Figure 1

The heights of 10 Swiss boys measured at 29 ages
Notes: Points indicate unequally spaced ages of measurement; solid line is a function fitted to these data by monotone smoothing.

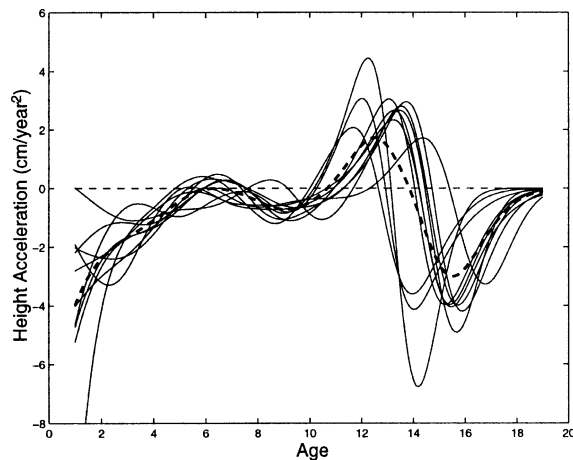


Figure 2

The estimated accelerations of height for 10 Swiss boys
Notes: Heavy dashed line is cross-sectional mean. This is a rather poor summary of the curves because of the presence of the timing of the pubertal growth spurt, called phase variation.

development, and this is much better illustrated by studying the velocity and acceleration of height. Figure 2 gives estimates of the acceleration or second derivative of these 10 height functions, and shows that around puberty there is a strong pulse of acceleration in growth, followed by sharp deceleration. This is the pubertal growth spurt (PGS), and there is also often a smaller spurt, hardly visible in Fig. 1 but obvious in Fig. 2, at around age 6.

2. The Mean Function and the Registration Problem

How would one estimate an overall acceleration curve to quantify the size of these spurts? The dashed curve in Fig. 2 is the average acceleration over the sample at each time point, but the PGS for this average is smaller in size and longer in duration than that of any of the individual spurts. Averages are supposed to look like typical observations; so what went wrong?

The problem is that the curves vary in two ways, termed amplitude variation and phase variation. Alternatively, these are called vertical and horizontal variation, respectively, since amplitude concerns the height of a function at a given point, and phase the location in time of a feature such as an acceleration peak. In short, the actual acceleration peaks in the PGS vary in both size and timing. An artificial example demonstrating these two sorts of variation is presented in Fig. 3. The phase shift displayed in this figure is constant across the timescale, but many data sets require more complicated and nonlinear transformations of the time axis in order to align curve features like the peaks and valleys in Fig. 3. The presence of phase variation arises in FDA because many processes involve a system time, or in this case biological time, that does not unfold at the same rate as clock time. In this case, the biological clock of each boy marks off time at its own unique rate.

A key part of the FDA methodology is time-warping or registration of the data curves to align specific features or to minimize variability, since estimating the mean and most other analyses are thrown off by the presence of phase variation. The time-warping functions may be simple constants, linear functions, or more complicated, nonlinear

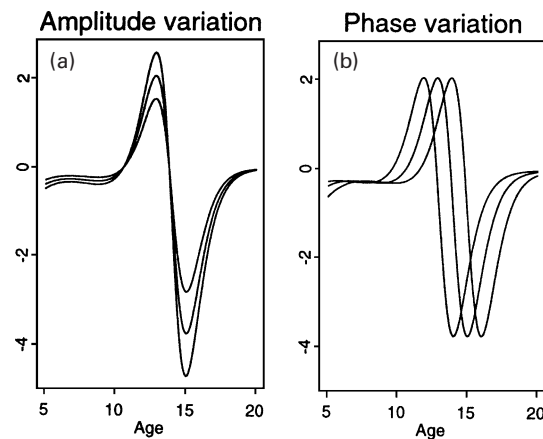


Figure 3

Panel (a): three height acceleration curves varying only in amplitude; Panel (b): three curves varying only in phase

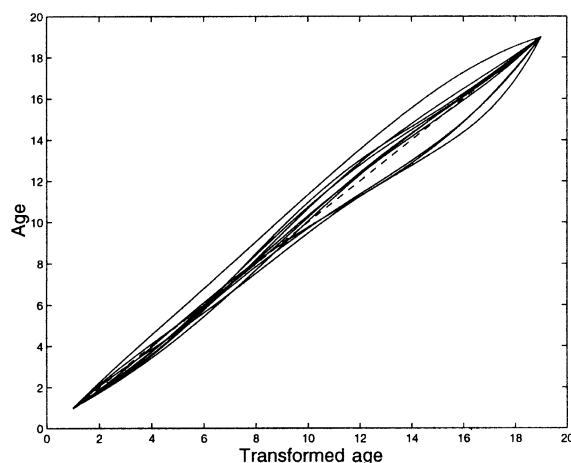


Figure 4

Time-warping functions taking chronological age on the horizontal axis into biological age on the vertical axis

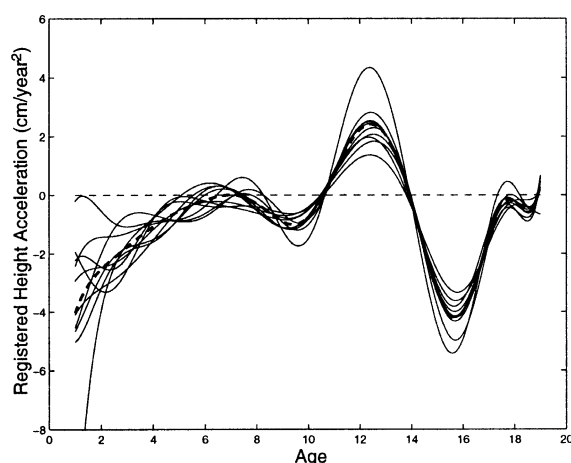


Figure 5

The height acceleration curves in Fig. 2 aligned or registered by measuring height with respect to the biological time shown in Fig. 4

Notes: Heavy dashed line is the mean acceleration. This provides a much better summary of the actual curves than Fig. 2

functions. The amount or shape of warp may be of great interest. See Ramsay and Silverman (1997, Chaps. 5 and 8). Figure 4 shows time-warping functions and registered height acceleration curves for the growth curves of Fig. 1 estimated using a method developed by Ramsay and Li (1998). The time-warping curve for each individual shows his maturation pattern. A warping function consistently above

the diagonal dashed line implies that clock time lags behind development, and therefore development is late in clock terms. If height is measured relative to biological age, the curves in Fig. 2 are aligned as shown in Fig. 5. Differences in height acceleration after registration correspond to pure amplitude variation in growth patterns. The mean of the registered curves in Fig. 5 is a much better summary than the mean of the unregistered data in Fig. 1.

3. The Nondurable Goods Index and More Derivatives

Valuable information may be gained by plotting one derivative of a functional observation against another, a technique called phase/plane plotting. The US index of nondurable goods production is an important economic indicator, and considerable insight may be gained by plotting the acceleration of this index against its rate of increase. If the index had a smooth trend plus a sinusoidal seasonal variation, this plot would be approximately circular with suitable choice of scales. The typical year 1970 in Fig. 6 has three main cycles, of which the spring cycle is rather smaller than the summer and fall/winter cycles. Comparisons of plots for different years is instructive. During the stable period 1964–6, there is little year-to-year change, but the period 1973–5 in Fig. 6 shows considerable instability, due primarily to the end of the Vietnam War. A typical plot for a more recent year, 1994, shows that the whole pattern has changed, indicating structural changes in the methods and economics of production. For further details, see Ramsay and Ramsey (2001).

4. Functional Principal Components Analysis

Many FDA methods are adaptations of classical multivariate methods such as principal components analysis (PCA), linear modeling, and analysis of variance (see *Linear Hypothesis* and *Multivariate Analysis: Overview*). Functional PCA demonstrates the way in which a set of functional data varies from its mean, and, in terms of these modes of variability, quantifies the discrepancy from the mean of each individual functional datum. Figure 7 (fig. 6.2 in Ramsay and Silverman (1997)) shows the first four principal component weight curves for temperature records for 35 Canadian weather stations. The first mode of variability is overall temperature, but with larger effects in the winter; the second corresponds to variation in annual range of temperature; the third partly to a time-shift of the overall cycle, and the fourth to short, hot summers and long, mild winters, or vice versa. Individual weather stations can then be

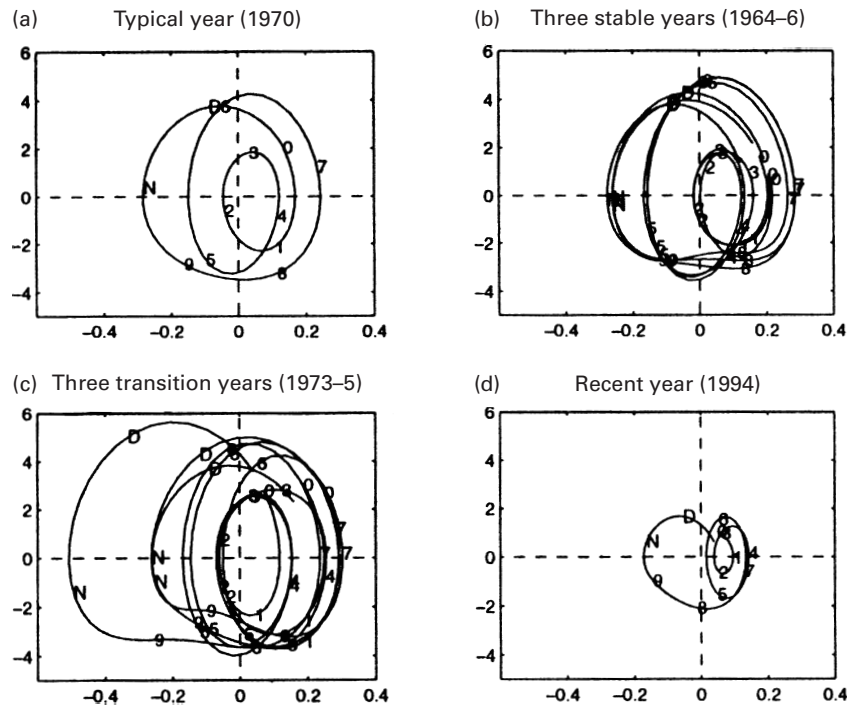


Figure 6

A phase/plane plot of *US index of nondurable goods production*. Each panel plots second derivative on vertical axis against first derivative on horizontal axis. Plots show structure of seasonal variation in index, and how structure varies with time

characterized according to their scores on these four criteria. The interpretability of the components can often be improved by a suitable rotation of the original principal components; see Ramsay and Silverman (1997, Chap. 6).

5. A Functional Linear Model and Regularization

An example of functional linear modeling is provided by the study of the dependence of the acceleration of the lower lip (*lipacc*) in speech on neural activity, as measured by electromyographical (EMG) recording. It is reasonable to suppose that the neural activity at time s continues to affect lip movement for only a short period of time, which we denote by δ and this leads to a model of the form

$$\text{lipacc}(t) = \alpha(t) + \int_{t-\delta}^t \text{EMG}(s) \beta(s, t) dt + \varepsilon(t),$$

developed by Malfait et al. (2000). The bivariate function $\beta(s, t)$ plays the same role as the regression coefficient in standard linear regression, and quantifies the relationship between the independent functional

variable EMG, and the dependent functional variable lip. One particular issue is the appropriate choice of the time lag δ the length of time within the model that neural activity can continue to affect movement. Models for increasing values of δ correspond to nested linear models, and in this case a study of the amount of variation explained demonstrates that there is little effect beyond a lag of about half a second.

An important theoretical and practical issue is the necessity or otherwise of incorporating some form of smoothing or regularization into a functional analog of a classical multivariate procedure. In principal components analysis, the principal components, such as those displayed in Fig. 7, are smoothed or regularized and in some procedures, such as canonical correlation analysis (see *Multivariate Analysis: Overview*) or certain types of linear modeling for the lip/EMG data, regularization is necessary to give any meaningful results at all; see Leurgans et al. (1993), and Ramsay and Silverman (1997, Chaps. 11, 12).

6. Modeling with Derivatives: A Central Theme

It is the smoothness of the processes generating functional data that differentiates this type of data

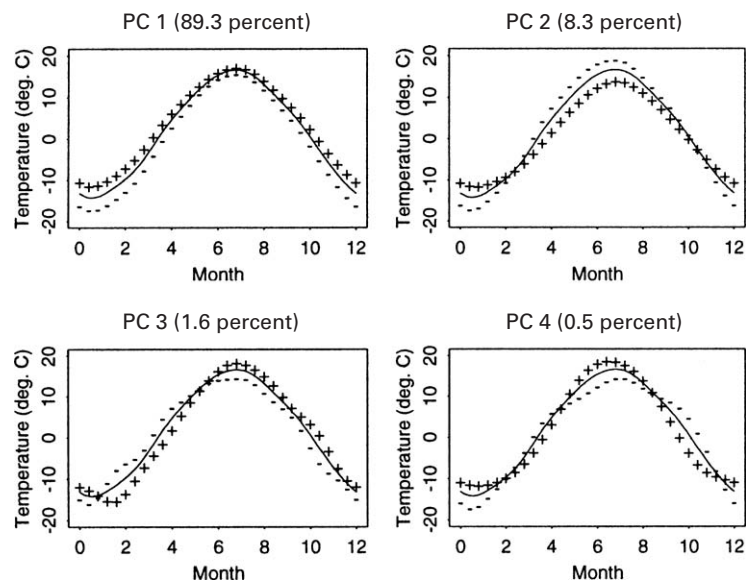


Figure 7

A principal components analysis of mean temperature curves for 35 Canadian weather stations

Notes: Solid line is the mean curve taken across all stations. The effects of adding (+) and subtracting (−) a suitable multiple of each component are also shown.

from more classical multivariate observations. This smoothness means that we can work with the information in the derivatives of functions or images. We looked at acceleration in the growth data, found something interesting in the joint variation of the velocity and acceleration of nondurable goods pro-

duction, and used derivatives to define roughness penalties; see Ramsay and Silverman (1997, Chap. 4) for the weather data.

We may also use derivatives to construct models for data. Such models are called differential equations, and are part of the standard model-building toolkit in

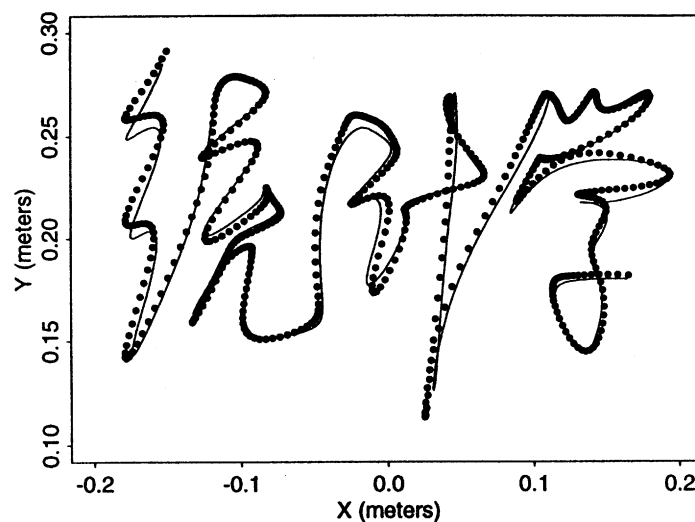


Figure 8

Observations of the tip of the pen while writing the word 'statistics' in Mandarin

Notes: Dots indicate 2400 observations; solid line is an approximation based on a second-order linear differential equation.

the natural sciences and engineering. Differential equation models are often called dynamic models, since they are especially effective at representing processes unfolding in time. Ramsay (2000) developed a linear differential equation for handwriting involving velocity, acceleration, and the third derivative of the tip of the pen. Figure 8 shows a sequence of pen positions during the writing of the word ‘statistics’ in Mandarin, and also shows as a solid line its approximation on the basis of the dynamic equation estimated from 50 such samples of Chinese script. The equation not only captured the essential features of this particular sample, but was also able to model a fair amount of the sample-to-sample variation as well. And, finally, it also provided an excellent account of the velocity, acceleration, and jerk functions at the same time. Ramsay and Silverman (1997, Chaps. 13, 14) take up this topic of dynamic modeling with functional data, including an account of a technique called ‘principal differential analysis,’ developed by Ramsay (1995).

7. Relationships with Other Branches of Statistics

Of course, statisticians have been exploring curves for a very long time, and there are many older statistical methodologies that can continue to inform and enrich an FDA. First of all, FDA usually begins with some kind of nonparametric regression (e.g., see *Exploratory Data Analysis: Multivariate Approaches (Nonparametric Regression)*) in order to replace discrete and possibly noisy curve values by smooth, continuous functions. Density estimation also involves data smoothing. Typically, time series analysis (Brillinger 1981, Box and Jenkins 1994) considers long sequences of curve values corresponding to equally spaced time values that are more or less stationary; that is, the data show little overall trend, and the covariance among neighboring values does not change much from one time region to another. However, methods for non-stationary processes are also under active development, and there are strong structural links between times series methods such as state-space modeling and FDA methods such as smoothing. Longitudinal data analysis or multilevel modeling (Diggle et al. 1994, Searle et al. 1992) usually looks at shorter sequences of curve values sampled over many cases, and studies the structure of between- and within-curve variation. It seems likely that this field and FDA will merge in a number of useful ways. Multivariate statistics has many methods developed specifically for functional data that have preceded FDA, but which have not explicitly exploited the smoothness of functional data or made other uses of derivatives (see *Probability Density Estimation; Longitudinal Data; Multivariate Analysis: Overview*).

8. Conclusion

By the smoothness of a functional observation we have meant the possibility that one or more derivatives, such as velocity or acceleration, may be examined or used in the modeling process. Otherwise, FDA makes only minimal assumptions about the structure of variation within and between curves. We have seen, however, that teasing apart amplitude and phase variation is a central issue, and one that has largely been ignored until recently. Certain types of models considered to be fairly routine and well understood in multivariate statistics, such as linear regression, have many more variants in the functional domain, and are being researched actively. Software for FDA is also being developed, and it is likely that FDA will soon be considered as a routine item in a data analyst’s black bag of instruments and nostrums. Further information on FDA and related issues may be obtained in Pinheiro and Bates (2000), Ramsay and Silverman (1997), and Simonoff (1996).

Bibliography

- Box G E P, Jenkins G M 1994 *Times Series Analysis: Forecasting and Control*, 3rd edn. Prentice-Hall, Englewood Cliffs, NJ
- Brillinger D R 1981 *Time Series: Data Analysis and Theory*. Holden Day, San Francisco
- Diggle P J, Liang K-Y, Zeger S L 1994 *Analysis of Longitudinal Data*. Clarendon Press, Oxford, UK
- Falkner F (ed.) 1960 *Child Development: An International Method of Study*. Karger, Basle, Switzerland
- Leurgans S E, Moyeed R A, Silverman B W 1993 Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society, Series B* **55**: 725–40
- Malfait N, Ramsay J O, Froda S 2000 The historical functional linear model. Unpublished manuscript. McGill University, Montreal, PQ
- Pinheiro J C, Bates D M 2000 *Mixed-effects Models in S and S-PLUS*. Springer, New York
- Ramsay J O 1995 Principal differential analysis: Data reduction by differential operators. *Journal of the Royal Statistical Society* **58**: 495–508
- Ramsay J O 1998 Estimating smooth monotone functions. *Journal of the Royal Statistical Society, Series B* **60**: 365–75
- Ramsay J O 2000 Functional components of variation in handwriting. *Journal of the American Statistical Association* **95**: 9–15
- Ramsay J O, Dalzell C 1991 Some tools for functional data analysis (with discussion). *Journal of the Royal Statistical Society* **53**: 539–72
- Ramsay J O, Li X 1998 Curve registration. *Journal of the Royal Statistical Society, Series B* **60**: 351–63
- Ramsay J O, Ramsey J B 2001 Functional data analysis of the dynamics of the monthly index of non-durable goods production. *Journal of Econometrics* (in press)
- Ramsay J O, Silverman B W 1997 *Functional Data Analysis*. Springer, New York

Searle S R, Casella G, McCulloch C E 1992 *Variance Components*. Wiley, New York
 Simonoff J S 1996 *Smoothing Methods in Statistics*. Springer, New York

J. O. Ramsay and B. W. Silverman

Functional Equations in Behavioral and Social Sciences

Theories in the behavioral, social, and natural sciences are often formalized by equations involving unknown functions, i.e., by functional equations. For instance, a theorist may be reluctant to make specific assumptions regarding the form of functions involved in a mathematical model, but the qualitative formulation of the model itself may impose constraints on the initially unknown functions. Such constraints often reduce the possibilities and occasionally are so severe that they restrict the possible forms to a few. A recent general reference to functional equations is Aczél and Dhombres (1989) where references to earlier and still useful surveys are given.

In the simplest cases there is one unknown, real-valued, function ϕ . An example, named after the famous French mathematician A. L. Cauchy, is the Cauchy equation $\phi(x+y) = \phi(x) + \phi(y)$, where x, y are in the set of real numbers $\mathbb{R}$. More restrictive domains are sometimes studied. With no further restrictions, the solutions to the Cauchy equation can be wild. But assuming that ϕ is monotonic, or continuous, or bounded over a finite interval, the solution reduces to $\phi(x) = cx$ for some constant c . By taking $\psi = \exp \phi$, we get $\psi(x+y) = \psi(x)\psi(y)$, which is the so-called 'lack of memory property' and it is easy to see from the previous case that the strictly monotonic solutions are $\psi(x) = e^{cx}$, $c \neq 0$.

A celebrated psychophysical example is the connection between empirical just noticeable differences and G. T. Fechner's hypothesis that sensation corresponds to subjective differences being equal. By one interpretation of what he meant, one is led to the family of functional equations, named Abel equations after the famous Norwegian mathematician N. H. Abel, $\eta[x+g(k,x)] - \eta(x) = h(k)$, where the variable x and the parameter k are in the set of nonnegative real numbers $\mathbb{R}_+$. For the special case of Weber's law where $g(k, x) = (k-1)x$, this equation reduces to $\eta(kx) - \eta(x) = h(k) = \eta(k) - \eta(1)$, i.e., with $\phi(z) = \eta(\exp z) - \eta(1)$, $k' = \log k$, $x' = \log x$, we have $\phi(k' + x') = \phi(k') + \phi(x')$, which if k' is treated as a variable is Cauchy's equation. The restriction of strict monotonicity yields $\eta(x) = a \ln x + b$, $a \neq 0$, b constants, as the solution. For references and a general discussion of closely related issue see Falmagne (1985, Chap. 4).

Sometimes a functional equation has multiple, qualitatively different solutions. Yet the scientist arriving at the functional equation has the strong intuition that only one of these solutions is really appropriate for the scientific problem in question. Whenever this happens, the challenge is to discover additional behavioral properties that seem to be empirically correct and that serve to isolate the desired solution. A reason for fully determining the several solutions is that the same functional equation may arise in an entirely different empirical context and, for that context, one of the previously unacceptable solutions may be appropriate. So the complete characterization is clearly of interest.

As will be seen, functional equations arise in the social sciences in at least three main ways. One occurs when one knows how to measure numerically the same attribute in two different ways, which is the case more often than not. Then the two measures are related by an unknown strictly increasing function. An empirical law linking the two underlying measurement structures manifests itself as an equation restricting that unknown function.

Functional equations also arise when some invariance condition holds. The Fechner hypothesis above is an example. Others are given later.

Economic aggregation problems are a third source of functional equations. Consistent aggregation both rules out some *ad hoc* aggregation functions and leads to families of functions that are indeed consistent. Despite the intuitive appeal of additive aggregation, that assumption is inconsistent with the most common production functions used in economics. Examples are given of production functions and of permissible aggregation rules.

1. Independent Measures of an Attribute

1.1 A Physical Example: Mass Measures

Consider mass measurement where one decides which of two objects x and y is more massive by observing which arm of an equal-arm beam balance drops (in a vacuum). This provides a mass ordering. One can construct a numerical measure m of mass, unique up to multiplication by a positive constant (ratio scale), by placing pairs of objects, denoted by $x \circ y$, on the pans. Under reasonable, empirically-testable assumptions m is additive over the operation of combining, i.e., $m(x \circ y) = m(x) + m(y)$. One can also construct a measure unique up to power transformations (log-interval scale) by varying the velocity of objects and determining which has greater momentum. Under empirically reasonable assumptions, the measure of momentum is multiplicative in powers of a mass and a velocity measure, i.e., $\alpha m'(x)^\beta v'(x)^\beta$, where $\alpha, \beta > 0$ but are otherwise unspecified. It can be shown that the two

measures m and m' preserve the same mass order, and so they must be related by a strictly increasing function ξ . A natural question is: What is the form of ξ ?

The following qualitative distribution property is empirically justified. Suppose (x, v) denotes the object x moving at velocity v and that $(x, v) \sim (x', v')$ means the two objects have the same momentum. Suppose also that $(y, v) \sim (y', v')$. Then, distribution asserts $(x \circ y, v) \sim (x' \circ y', v')$. This condition forces ξ to be multiplicative in the sense of satisfying $\xi(rs) = \xi(r)\xi(s)$, where the variables are real numbers with $r > 0, s > 0$. By setting $\xi = \exp \phi(\ln)$ we see from the strictly increasing solutions of the Cauchy equation that $\xi(r) = r^\beta$, $\beta > 0$. Thus, one mass measure is a power function of the other one. See Luce et al. (1990, p. 125).

1.2 Preferences Among Gambles

Something quite analogous occurs in studying people's preferences among uncertain alternatives, often called gambles. A binary gamble $(x, C; y, \bar{C})$ means that x is the consequence if event C occurs and y otherwise when a chance 'experiment' with possible outcomes $E = C \cup \bar{C}$ is conducted. The preference order over gambles, which is assumed to be connected and transitive, is denoted by $\succsim$. We assume there is a status quo e , and a consequence x is said to be a gain if $x \succsim e$ and a loss if $e \succsim x$. A well developed theory of preferences over gambles, which generalizes subjective expected utility (Savage 1954), was worked out during the 1980s (see Quiggin 1993). It establishes that under certain conditions a utility function U over gambles and a weighting function W over chance events exists such that for gains x, y , and $g = (x, C; y, \bar{C})$

$$U(g) = \begin{cases} U(x)W(C) + U(y)[1 - W(C)], & \text{if } x \succsim y \\ U(x)[1 - W(\bar{C})] + U(y)W(\bar{C}), & \text{if } x \precsim y \end{cases} \quad (1)$$

$U(g) = U[(x, C; y, \bar{C})]$ is usually written just as $U(x, C; y, \bar{C})$, which is a minor abuse of notation. Because we assume $x \sim (x, C; x, \bar{C})$, U on gambles is also on consequences as well. This representation is called rank-dependent utility because the weight used in the average depends upon the ordering of the consequences. The model can be extended to cover losses and mixed gains and losses, but we do not go into that here (see Luce 2000).

One can also consider having or receiving pairs of valued consequences and gambles. Suppose g and h are gambles (including, as special cases, pure consequences) and let $g \oplus h$ denote having or receiving both. This binary operation $\oplus$ is called joint receipt. Assume that the preference order $\succsim$ extends to the domain of joint receipts. It is plausible that, for gains, $\oplus$ and $\succsim$ satisfy the same formal properties as does mass where no change from the status quo, e , plays the

role of an identity. Thus, there is a real-valued 'value measure' V that preserves the order $\succsim$ and is additive over $\oplus$, i.e., $V(x \oplus y) = V(x) + V(y)$.

A natural question is: How do U and V relate? Luce and Fishburn (1991) considered the following linking hypothesis (or putative 'law') called segregation: For all gains x, y and events C

$$(x, C; e, \bar{C}) \oplus y \sim (x \oplus y, C; y, \bar{C}) \quad (2)$$

Note that both sides yield the same thing under the same conditions, so this is a highly rational property. If we introduce the notation $G(u, v) = U[U^{-1}(u) \oplus U^{-1}(v)]$ one can show that segregation implies that for $u = U(x)$, $v = U(y)$, $w = W(C)$ and $u \geq v$

$$G(uw, v) = G(u, v)w + v(1 - w) \quad (3)$$

For commutative the solution to this functional equation (Luce and Fishburn 1991) is that for some real δ

$$G(u, v) = u + v - \delta uv \quad (4)$$

From this equation one can show that U and V are related as follows. If $\delta = 0$, then $U = \alpha V$. If $\delta \neq 0$, then either $U_+ = 1 - e^{-\kappa V}$ or $U_+ = e^{\kappa V} - 1$, where $U_+(x) = |\delta|U(x)$, $x \succsim e$. Thus, if one knows how V depends on the consequences, then one knows up to a constant how U_+ does also. Conditions are given that determine the form of V over money and W over probability. [In like manner, a theory for losses is possible with constant δ' and $U_-(x) = |\delta'|U(x)$, $x \precsim e$.]

From these expressions and the additivity of V over $\oplus$ one can deduce formulas for the utility U over $\oplus$ and for gambles when the consequences are a gain and a loss. Some interesting non-bilinear expressions arise (Luce 2000).

Recently, Luce and A. A. J. Marley (summarized in Luce 2000) have suggested a generalized representation of binary gambles over gains, namely, for $x \succsim y \succsim e$,

$$\phi_0[(x, C; y, \bar{C})] = \phi_1[u_1(x)w_1(C)] + \phi_2[u_2(y)w_2(\bar{C})] \quad (5)$$

where $\phi_1(0) = \phi_2(0) = 0$. The ranked-additive decomposition in this form has been axiomatized by Wakker (1991, 1993) and the separability of u_1w_1 can be defended axiomatically for the first component (Luce 2000). Assume, as seems reasonable,

$$x \sim (x, E; y, \emptyset) \quad (6)$$

$$y \sim (x, \emptyset; y, E) \quad (7)$$

$$x \sim (x, C; x, \bar{C}) \quad (8)$$

where $\emptyset$ denotes the null event and E the universal one underlying the gamble. Then noting $w_1(E) = 1$, $w_2(\emptyset) = 0$, the first indifference yields $\phi_0(x) = \phi_1[u_1(x)]$. If we set $y = x$, $v = u_1(x)$, $w = w_1(C)$, $w_2(\bar{C})$

$= Q[w_1(C)] = Q(w)$, then the second of the above conditions implies $x \sim (x, \emptyset; x, E)$, and so $\varphi_1(v) = \varphi_0(x) = \varphi_2[u_2(x)]$ or rewriting $u_2(x) = \varphi_2^{-1}[\varphi_1(v)] = \psi(v)$. Thus, setting $\varphi = \varphi_1$ the third yields the functional equation

$$\varphi(v) = \varphi(vw) + \varphi(\psi^{-1}[\psi(v)Q(w)]), \quad v \in [0, k], w \in [0, 1] \quad (9)$$

where the unknown functions φ, ψ are strictly increasing, and Q is strictly decreasing. Aczél et al. (2001) have determined all solutions to this equation. From these results, Luce and Marley (see Luce 2000) showed the following utility representation: For $x \succcurlyeq y \succcurlyeq e$, there exist constants $c > 0$ and $\mu > -1/k^c$ such that

$$U(x, C; y, \bar{C}) = \frac{U(x)W(C) + U(y)[1 - W(C)] + \mu U(x)U(y)W(C)}{1 + \mu U(y)W(C)} \quad (10)$$

where $U = (u_1)^c$ and $W = (w_1)^c$. If, however, one supposes that segregation holds, then one can show that $\mu = 0$, which of course is just the rank-dependent model.

2. Invariance Principles

2.1 Value Function for Money

Section 1.2 left open the question of how V depends upon money. The most obvious assumption is that if x, y are sums of money, then $x \oplus y = x + y$. Thus, $V(x) + V(y) = V(x \oplus y) = V(x + y)$, which is Cauchy's equation. Because V is strictly increasing, V is thus proportional to money. Some empirical work suggests that, in general, this hypothesis may be too strong (Thaler 1985). An alternative assumption is that $\oplus$ is invariant under proportional changes in money, i.e., for any $\lambda > 0$,

$$x \oplus y \sim z \Rightarrow \lambda x \oplus \lambda y \sim \lambda z \quad (11)$$

Using the additivity of V ,

$$V^{-1}[V(\lambda x) + V(\lambda y)] = \lambda z = \lambda V^{-1}[V(x) + V(y)] \quad (12)$$

For $x, y \geq 0$, functional equation arguments lead, through multiplicative functions, to $V(x) = \alpha x^\beta$ and $x \oplus y = (x^\beta + y^\beta)^{1/\beta}$, for some $\alpha > 0, \beta > 0$. This result, combined with those of Sect. 1.2, gives U to within one additional free parameter β (note that κ of U_+ and α are not separately identifiable).

2.2 Prelec's Weighting Function

Section 1.2 also left open the question of the form of the weighting function W . When the gambles are

characterized in terms of probabilities $p = \Pr(C)$, much empirical work shows that in general for some p_0 in $]0, 1[$

$$W(p) \begin{cases} > p, & \text{if } 0 < p < p_0 \\ = p, & \text{if } p = p_0 \\ < p, & \text{if } p_0 < p < 1 \end{cases} \quad (13)$$

with W strictly increasing initially concave, and then convex. Such weighting functions are described as inverse S-shaped. Until Prelec (1998), suggestions of the mathematical form have not been derived from behavioral principles. He provided a somewhat complex behavioral invariance condition leading to $W(p) = \exp[-\gamma(-\ln p)^\alpha]$. Gonzalez and Wu (1999) showed that the Prelec form fits empirical data remarkably well. Not only is this function inverse S-shaped for $\alpha < 1$, but for $\alpha = 1$ it includes the 'rational' weighting function that arises from the conditions that gambles of the form $(x, p; 0)$ have a separable representation $U(x, p; 0) = U(x)W(p)$ and that $((x, p; 0), q; 0) \sim (x, pq; 0)$ holds.

Luce (cited in Luce 2000) gave a simpler condition, called reduction invariance, equivalent to Prelec's function which generalizes a property of the rational condition to the non-rational case: For all gains x and p, q, r in $[0, 1]$,

$$((x, p; 0), q; 0) \sim (x, r; 0) \text{ implies, for } N = 2, 3, \text{ that } ((x, p^N; 0), q^N; 0) \sim (x, r^N; 0) \quad (14)$$

From this, one first shows that the right side holds for N replaced by any real exponent $\lambda > 0$. Then, using separability, one arrives at the functional equation

$$(W^{-1}[W(p) + W(q)])^\lambda = r^\lambda = W^{-1}[W(p^\lambda) + W(q^\lambda)] \quad (15)$$

Note that with $V = \ln W(\exp)$, $x = \ln p, y = \ln q$ this reduces to the same functional equation as that for the value of money.

2.3 Additive Conjoint Weber Laws

Consider stimulus pairs (a, x) such as a tone of intensity a in a background noise of intensity x . Denote by $P(a, x; b, y)$ the probability that a subject perceives a in x more easily than b in y . Theoretical psychophysicists wish to understand the mathematical structure of such families of probabilities.

To do so, they explore various regularity or invariance conditions. For example, a homogeneity condition called a 'conjoint Weber law' assumes that for all real $\lambda > 0$

$$P(\lambda a, \lambda x; \lambda b, \lambda y) = P(a, x; b, y) \quad (16)$$

That is, the probability is invariant under multiplication of all four intensities by any positive number λ . A second important hypothesis is that of 'additive probability' in the following sense: There are con-

tinuous functions f strictly increasing, r strictly decreasing, and H strictly increasing in the first and strictly decreasing in the second variable such that

$$P(a, x; b, y) = H[f(a) + r(x), f(b) + r(y)] \quad (17)$$

In place of the last equation, Falmagne and G. Iverson (see Falmagne 1985, pp. 289–91) assumed the weaker condition

$$P(a, x; b, y) = H[q(a, x), g(b, y)] \quad (18)$$

of simple scalability where H is as above and g is strictly increasing in the first and strictly decreasing in the second variable, and the variable pairs (c, z) lie in subsets closed under multiplication by any $\lambda > 0$. They showed that then P must be of one of two forms:

$$P(a, x; b, y) = M \left[\frac{T(a/x)x}{T(b/y)y} \right] \quad (19)$$

or

$$P(a, x; b, y) = Q\left(\frac{a}{x}, \frac{b}{y}\right) \quad (20)$$

where $T > 0$, M , and Q are continuous, T and M are strictly increasing, and Q is strictly increasing in the first and strictly decreasing in the second variable.

In particular, they determined the form of all representations that are both additive and satisfy a conjoint Weber law as follows. The last equation when rewritten as

$$P(a, x; b, y) = Q(e^{\ln a - \ln x}, e^{\ln b - \ln y}) \quad (21)$$

is additive with

$$H(u, v) = Q(e^u, e^v), f(a) = \ln a, r(x) = -\ln x \quad (22)$$

The first homogeneous, simple scalable form is additive if and only if

$$M \left[\frac{T(a/x)x}{T(b/y)y} \right] = H[f(a) + r(x), f(b) + r(y)] \quad (23)$$

Holding $b = b_0$, $y = y_0$ constant and letting

$$K(t) = T \left[\frac{b_0}{y_0} \right] y_0 M^{-1}(H[t, f(b_0) + r(y_0)]) \quad (24)$$

we get

$$K[f(a) + r(x)] = T \left[\frac{a}{x} \right] x \quad (25)$$

Thus, for all $\lambda > 0$,

$$K[f(\lambda a) + r(\lambda x)] = \lambda K[f(a) + r(x)] \quad (26)$$

This functional equation and its method of solution is similar to but somewhat more complicated than those in Sects. 2.1 and 2.2. Moreover, it has two distinct sets of solutions:

$$f(a) = Aa^\beta + B, \quad r(x) = Cx^\beta + D \quad (27)$$

and

$$f(a) = A \ln a + B, \quad r(x) = C \ln x + D \quad (28)$$

Accordingly, for all homogeneous, additive, conjoint Weber laws, the probabilities are of one of the following three forms:

$$P(a, x; b, y) = Q \left(\frac{a}{x}, \frac{b}{y} \right) \quad (29)$$

$$P(a, x; b, y) = G \left(\frac{a^\beta + \gamma x^\beta}{b^\beta + \gamma y^\beta} \right) \quad (30)$$

$$P(a, x; b, y) = G \left(\frac{ax^\beta}{by^\beta} \right) \quad (31)$$

where Q is as above, G is continuous and strictly increasing, and $\beta > 0$, $\gamma \neq 0$ are constants.

Results of this character are useful both in understanding the implications of general invariance principles and in suggesting to experimentalists how to plot and analyse their data.

3. Aggregation

Consider an industry with m producers using n inputs. Denote by x_{jk} the amount of inputs k used by producer j . Let $y_j = F_j(x_{j1}, \dots, x_{jn})$ denote the output of producer j and $z_k = G_k(x_{1k}, \dots, x_{mk})$ denote the total consumption of input k . Consistent aggregation holds if there is a macro production function F and macro input function G such that

$$\begin{aligned} F(z_1, \dots, z_n) &= F[G_1(x_{11}, \dots, x_{m1}), \dots, G_n(x_{1n}, \dots, x_{mn})] \\ &= G[F_1(x_{11}, \dots, x_{1n}), \dots, F_m(x_{m1}, \dots, x_{mn})] \\ &= G(y_1, \dots, y_m) \end{aligned} \quad (32)$$

This equation is called $m \times n$ generalized bisymmetry. Although a full treatment deals with general domains (see, e.g., Aczél and Maksa 1996, Aczél 1997), here it is assumed that all of the variables and functions lie in the set of non-negative real numbers $\mathbb{R}_+$. Assuming

that all functions are strictly monotonic and continuous, Maksa (1999) has proved that there are continuous functions α_k , β_{jk} , γ_j , and φ , $j = 1, \dots, m$, $k = 1, \dots, n$ such that

$$F(z_1, \dots, z_n) = \varphi^{-1}[\alpha_1(z_1) + \dots + \alpha_n(z_n)] \quad (33)$$

$$F_j(x_{j1}, \dots, x_{jn}) = \gamma_j^{-1}[\beta_{j1}(x_{j1}) + \dots + \beta_{jn}(x_{jn})] \quad (34)$$

$$G(y_1, \dots, y_m) = \varphi^{-1}[\gamma_1(y_1) + \dots + \gamma_m(y_m)] \quad (35)$$

$$G_k(x_{1k}, \dots, x_{mk}) = \alpha_k^{-1}[\beta_{1k}(x_{1k}) + \dots + \beta_{mk}(x_{mk})] \quad (36)$$

Aggregation functions of this quasi-additive type had been obtained earlier, but under stronger regularity conditions (for a survey see Aczél 1997). Maksa's achievement was to get the same results under appreciably weaker conditions.

One way to interpret this result is that if the variables and functions are 'properly' measured, i.e., as $\varphi(F)$, $\varphi(G)$, $\gamma_j(F_j)$, $\alpha_k(G_k)$, $\alpha_k(z_k)$, $\beta_{jk}(x_{jk})$ and $\gamma_j(y_j)$, then aggregation is additive. This is, however, very different from assuming that the 'raw' inputs and outputs are related additively. Consider the following examples. Two commonly used production functions are the Cobb-Douglas (CD) function

$$F(z_1, \dots, z_n) = az_1^{c_1} z_2^{c_2} \dots z_n^{c_n}, \quad a, c_k > 0 \quad (37)$$

and the Constant Elasticity of Substitution (CES) Function

$$F(z_1, \dots, z_n) = (c_1 z_1^b + \dots + c_n z_n^b)^{1/b}, \quad c_k > 0, b \neq 0 \quad (38)$$

Aczél and Maksa (1996) have shown that if any one of the aggregation functions is simple addition at the level of the 'raw' inputs or outputs, then it is impossible for any production function to be either CD or CES with $b \neq 1$. This does not say that inputs (or outputs) cannot actually be additively aggregated. Rather, it says that such additive aggregation is not consistent with common and realistic production functions. (One can prove that it is consistent only with linear production functions.)

Another question addressed is: If the F_j functions are in some sense of the same form, does that dictate that the macro F is also of that form? This is called the representativeness problem. The answer depends upon exactly what one means by 'the same form.' In a trivial sense, the answer is 'Yes' since the F and F_j 's in the above quasi-additive solution have roughly the same structure. A somewhat more sophisticated answer is that it need not be so. For example, if all of the F_j are CD functions, then

$$\gamma_j(y) = \ln \left(\frac{y}{a_j} \right), \quad \beta_{jk}(x) = c_k \ln x \quad (39)$$

Or if they are all CES functions, then

$$\gamma_j(y) = y^b, \quad \beta_{jk}(x) = c_k x^b \quad (40)$$

Neither conclusion restricts φ and α_k in any way. Thus one may choose F to be CD or CES or neither.

So the general conclusion from solving the aggregation problem is that consistent and even representative aggregation is possible only for appropriately chosen functions. In general neither is possible if the aggregating functions are preselected in the 'wrong way.' The implications for the possibility of macro production models are considerable.

Related ideas and results concerning aggregation of probabilistic models of choice, with results for Luce's choice model as a special case, appear in Aczél et al. (1997).

4. Conclusions

For many static questions in the behavioral and social sciences, functional equations serve as a powerful tool in understanding theoretical constraints. Three types of constraint have been illustrated: (a) the linking of measurement structures that have a common qualitative attribute; (b) the imposition of invariance principles of various types of which three were illustrated; and (c) the question of whether one can aggregate over inputs and separately over industries and arrive at the same macro conclusion.

See also: Algorithms; Axiomatic Theories; Mathematical Psychology; Measurement, Representational Theory of; Measurement Theory: Conjoint; Ordered Relational Structures; Psychophysical Theory and Laws, History of; Utility and Subjective Probability; Contemporary Theories

Bibliography

- Aczél J 1997 Bisymmetry and consistent aggregation: Historical review and recent results. In: Marley A A J (ed.) *Choice, Decision, and Measurement: Essays in Honor of R. Duncan Luce*. Erlbaum, Mahwah, NJ
- Aczél J, Dhombres J 1989 *Functional Equations in Several Variables with Applications to Mathematics, Information Theory and to the Natural and Social Sciences*. Cambridge University Press, Cambridge, New York
- Aczél J, Maksa G 1996 Solution of the rectangular $m \times n$ generalized bisymmetry equation and of the problem of consistent aggregation. *Journal of Mathematical Analysis and Applications* **203**: 104–26
- Aczél J, Maksa G, Marley A A J, Moszner Z 1997 Consistent aggregation of scale families of selection probabilities. *Mathematical Social Sciences* **33**: 227–50
- Aczél J, Maksa G, Ng C T, Páles Z 2001 A functional equation arising from ranked additive and separable utility. *Proceedings of the American Mathematical Society* **129**: 989–98

- Falmagne J-C 1985 *Elements of Psychophysical Theory*. Oxford University Press, New York
- Gonzalez R, Wu G 1999 On the form of the probability weighting function. *Cognitive Psychology* **38**: 129–66
- Luce R D 2000 *Utility of Gains and Losses: Measurement-Theoretical and Experimental Approaches*. Erlbaum, Mahwah, NJ
- Luce R D, Fishburn P C 1991 Rank- and sign-dependent linear utility models for finite first-order gambles. *Journal of Risk and Uncertainty* **4**: 25–59
- Luce R D, Krantz D H, Suppes P, Tversky A 1990 *Foundations of Measurement*. Academic Press, New York Vol. III
- Maksa G 1999 Solution of generalized bisymmetry type equations without surjectivity assumptions. *Aequationes Mathematicae* **57**: 1–25
- Prelec D 1998 The probability weighting function. *Econometrica* **66**: 497–527
- Quiggin J 1993 *Generalized Expected Utility Theory: The Rank-dependent Model*. Kluwer Academic Publishers, Boston
- Savage L J 1954 *The Foundations of Statistics*. Wiley, New York
- Thaler R H 1985 Mental accounting and consumer choice. *Marketing Science* **36**: 199–214
- Wakker P P 1991 Additive representations on rank-ordered sets. I. *Journal of Mathematical Psychology* **35**: 501–31
- Wakker P P 1993 Additive representations on rank-ordered sets. II. The topological approach. *Journal of Mathematical Economics* **22**: 1–26

J. Aczél and R. D. Luce

Functional Explanation: Philosophical Aspects

Functional explanation has received a great deal of philosophical attention in recent years, partly as a consequence of a surge of interest in the philosophy of biology, and partly as a consequence of attempts to make functions and functional explanation central to issues in the philosophy of psychology. In both of these areas there has been some controversy over what the correct account of functional explanation is, with the contending accounts splitting into two main camps. Both camps agree that functional explanation is dependent on an account of functions, but disagree on what the correct account is. Within evolutionary biology the favored view is that functions emerge over time as a consequence of evolutionary processes, so a certain feature of an organism will only have a function if it has a suitable history. The opposed view is that the operation of a system can be analyzed into the way in which parts of the system function to make the system behave in that manner. This version has many names, but is perhaps best known as a ‘systems’ or ‘Cummins’ account of functions and functional explanation, but it has also been called ‘causal-role’ functionalism. In what follows both of these views will be examined; in addition the controversy over what it is that functional explanations explain will be outlined.

1. Functions and Systems

The ‘systems’ account of functions takes as its starting point the fact of (biological) complexity and proposes to understand the complex functioning of a biological entity (or system) in terms of the working of its parts. The fact of complexity suggests that the systems are goal directed, and if this is taken for granted one can recover what is distinctive of many explanations in biology, that they invariably cite the effects of causes as explaining why the cause is there. It does this by specifying how a part of the system operates in the production of the aimed-at goal.

The best known account of how such an explanation operates is given by Cummins (1975, 1983). He distinguishes explanation of changes, which is done by ‘transition’ theories, from explanation of properties. Transition theories are suitable for explaining ‘What caused S to acquire P?’ whereas a property theory explains what it is for S to instantiate P. This is best done by constructing an analysis of S, such an analysis adverting to the properties of S’s components and their organization. So the kinetic theory of gases explains what it is for a gas to have temperature via an account of the properties of the molecules contained within the gas. Cummins calls the analysis of such a system compositional analysis: the analysis of a dispositional property being called functional analysis. The general claim is that in science the ascription of a function to something is to specify a capacity, such a capacity being identified by its role in enabling a system to operate in a certain manner. Thus, in biology the capacities of an organism are explained by analyzing them into a number of systems (circulatory, digestive, etc.), such systems having a particular task to perform. How these tasks get performed is in turn explained by functional analysis, such an analysis citing the capacities of parts of these systems. Such capacities are the functions of the parts, so we get the Cummins-style definition of function:

X functions as an *F* in *s* (or: the function of *X* in *s* is to *F*) relative to an analytical account *A* of *s*’s capacity to *G* just in case *X* is capable of *F*-ing in *s* and *A* appropriately and adequately accounts for *s*’s capacity to *G* by, in part, appealing to the capacity of *X* to *F* in *s* (Cummins 1975, p. 762).

This account clearly does not restrict the attribution of functions to biological organisms; one can recognize the attribution of this type of function to parts of societies within sociological and anthropological functionalism. Here the system was seen as a society or culture, and the parts were the subsocietal structures or the recurrent patterns of behavior, the function of which was to maintain the structural continuity of the whole. The fact that this account of functionality is applicable to any system has given rise to two criticisms of this approach, one seeing a problem in how to specify nonarbitrarily which behavior or capacity of the containing system is relevant for the

functional analysis to analyze, the second seeing a difficulty in producing a sufficiently determinate notion of a 'system' for which a functional analysis is appropriate. The problem is that the choice of both a capacity and a system may seem to depend upon our decisions in a way which could make the functional analysis of any capacity or system seem justifiable. The pertinent question here concerns the explanatory gain made by the appeal to functional analysis. There seems to be no explanatory pay-off in some cases where functional analysis may be applied. Take the 'system' which is my study, which includes as parts various items of furniture, books, and me. The property of the room that we want explained is its temperature. Say my presence in the study raises the temperature of the room by $1/2$ a degree. One does not gain any insight into the working of this system by attributing to me the function of enabling the room (the system) attain its present temperature.

Cummins was aware of this trivialization of functional analysis, and suggested constraints on the choice of a system and capacity that could be analyzed illuminatingly. One problem with the above analysis is that the capacity to be explained via functional analysis, the temperature of the room, is similar in kind and complexity to the property doing the explaining, my body temperature. One gets an explanatory pay-off when the analyzing property is both less complicated than, and different in kind from, the analysandum. This suggests the third constraint imposed by Cummins: one is not going to get a less sophisticated property doing all the explanatory work in showing how a complex system operates. The analysis will have to appeal to the operation of component parts and their properties, their combined action producing the complex behavior of the system. It is through the organization of interdependent parts that simplicity generates complexity, and functional analysis succeeds when it shows how this can be achieved.

The choice of system and capacity to be analyzed, then, is constrained: functional analysis will be applied fruitfully only to a complex system whose behavior is produced by the action of simpler but well organized parts. It is unsurprising, then, that the systems most often favored as exemplars of systems suitable for functional analysis are those that have been designed. The interdependence of the parts of such systems is often attained by careful planning. The second theory of functions that will now be examined focuses on this 'design' aspect of the systems to whose parts functions can be aptly attributed.

2. Functions and Selection

It is the improbability of complex systems existing by chance that suggests that such systems are generated by design, and so have parts which function so as to

produce the intended behavior of the system. With artifacts it is easy to see how functionality can be ascribed: the complex system is designed intentionally to have the capacity to produce certain effects, and the parts are chosen for their ability to contribute towards that goal. They are selected for that task. This type of functionality is derived from intentionality. The achievement of Darwin and Wallace was to provide a nonintentional explanation of how such complexity could originate, and they did this via the theory of natural selection. The major strength of the historical account of functions is that it ties attribution of functions to 'design' by natural (nonintentional) selection. The essential features of this account are dependent upon the idea that to attribute a function to an item is to say that the item has that feature or behavior because it has a certain history: in the past that feature produced certain consequences, those consequences themselves having the effect of reproducing items with that feature. It is this latter aspect which makes the account teleological. One can say that the item has the relevant property in order to produce those consequences. It is because of (some of) its effects that the item with this capacity is selected for. What the theory of natural selection does is make respectable the apparent purposefulness of the 'in order to.' It does this by showing how environmental pressures on variable items can produce varied reproduction of those items; those items with certain features will become more prevalent because those features have typical effects on the reproductive capacity of the items which bear them.

The long neck of the giraffe, to take a standard example, is an adaptation and has the function of enabling giraffes to eat the leaves at the top of tall trees. On the historical account, this capacity is the function of the long neck because previous giraffes whose genes produced longer necks than their contemporaries were better producers of offspring, and were so just because of the advantage conferred on those giraffes in the food wars. This results, in the long run, in more long-necked giraffes, until the giraffe is a long-necked species. Put as briefly and generally as possible, the etiological account says that functionality arises because some variants within a group (a mutant gene, for example) have novel features with capacities which are favourable to their possessors' ability to reproduce (i.e., more favourable than the nonpossession of those features by the other members of the group). Such features are transmitted to their descendants, proliferating within the group in the process. The feature of the organism will then, and only then, have as its function the exercise of the favorable capacity.

As with Cummins-style functions, this account of functionality is very general. It applies not just to biological entities, but to all items that have the required characteristics of:

- (a) variability in the items to be selected,

- (b) selection of some of those items, and
- (c) variable reproduction of the items selected for.

Selection cannot take place if there is no initial variation; the same selection forces operating on a homogeneous population will have no effect. For a cause to be there in order to produce certain effects it must also be the case that instances of properties with those causal powers have been selected in the past because they produced those effects. The selection process requires that the causal property interact with salient features of the environment in such a way that other items without that property are seen to 'suffer' some loss as a consequence of their non-possession of the property. Finally, it is because the favored properties are transmitted to descendants of the items having those features that the present items have these properties. Then it can be said that the property is, or was functional, its function being to produce those types of effects which lead to its ancestors reproducing items with the property in question.

As the giraffe example indicates, it is in its forging a connection between function and adaptation that the strength of this account resides. Adaptations are produced historically via the operation of natural selection, so they are functional. A creature may have a feature which renders it 'adapted' to its environment because it so happens that that feature will help it flourish in that environment, but such a feature, lacking the relevant selectionist history, will be adaptive for the creature, but not an adaptation. The historical (etiological) theory of functions claims that it is because adaptations are selected, not accidental, features of organisms that we are entitled to the use of functional explanation.

The historical and systems account of functions are not necessarily competitors. It has been seen that the systems for which functional analysis is appropriate are complex, with the analysis proceeding by showing how the operation of simpler, interdependent parts can produce the relevant capacity of the system. Such systems, it can be claimed, come into being either by intention or by selection, so functional analysis will be appropriately applied in just those cases where some design-mechanism is operating. On this view, functional analysis is an accessory to the historical theory of functions. There are, however, two aspects of the historical conception that render it significantly different from the systems version. The first is a consequence of what has been suggested above: it is probable that complex systems are generated by some design process, but it is not impossible for such a system to emerge by chance. For such a system, a functional analysis would still be appropriate, but the historical version of functions would disallow the attribution of any function to the system or its parts, these not having the appropriate history. Second, the historical theory has the ability to include malfunctions within its purview. On the systems account, if a part 'functions' so as to produce a system with a

different capacity from that expected, that part cannot be said to be malfunctioning. A functional analysis will proceed as usual, with the new capacity being analyzed, and the analysis showing just how it is produced given the nature of the contained parts, and their causal powers. On the historical conception, given a suitable account of reproduction, one can allow for mistakes to be made in the reproducing process, so that the items which are considered to be descendants of some common ancestor need not be perfect copies of that ancestor, and in particular may not have the causal powers for which items of that type have been selected previously. The item may then not do what it is supposed to do—it may malfunction. To appreciate this better, and to make manifest the power of the explanations which the historical theory makes available, that theory needs to be fleshed out in more detail.

3. *Elaborations: Malfunction and Novelty*

An essential aspect of the historical account is that a function results from selection operating over a period of time on items that are descended from each other. What is it to be descended in the relevant sense? Initially one may think in terms of copying, one item being a descendant of another if it is an exact copy of the former, and is so as a result of some law-like process that ensures that such a replica is forthcoming. This is too strict a definition: for one thing, it would be impossible to provide a decent conception of a malfunctioning device if all descendants had to be exact copies of their ancestors. A malfunction involves some mistake, so a device that malfunctions will not be an exact copy of a properly functioning device. One has to allow for an item being a descendant of another even though the copying process is not perfect. One way of making what is involved more precise is due to Millikan (1984, 1993), and it goes like this: individual A is a reproduction of individual B if and only if B has some determinate properties in common with A, and this correlation of properties can be explained by a natural law operating in that environment. These common properties are the reproductively established properties of B, and the devices sharing these properties form reproductively established families (refs.). (Only first order refs. reproduce in this manner. Higher order refs. (horefs.) are produced by first-order refs. My and my ancestors' two-leggedness is the result of common genes, such genes being first-order refs. They have certain properties in common, this being ensured by the laws of genetic inheritance.) Direct proper functions are functions of devices that are members of refs. or horefs.

So how do we get malfunctioning members of refs. or horefs? Millikan suggests that a malfunctioning member of a ref. can be produced provided it has been produced by a device the proper function of which is to produce members of the ref. in question, that it is in

some aspects similar to other members of the ref., and that its production has an explanation which approximates (in some undefined degree) to a normal explanation for the production of members of the ref. Because a particular reproductive process need only approximate the 'perfect' reproductive process it can produce imperfect copies, ones which malfunction.

The historical theory's emphasis on the role of reproduction and selection has given rise to the thought that functional explanation, construed in this way, cannot explain novel behavior. On the historical account it is inheritance from past (selected for) patterns of behavior that gives rise to the present functional behavior. However, to suppose that novel behavior cannot be produced by something whose function it is to produce that behavior is to underestimate the variety of functions. Millikan notes that there can be a relational proper function, which is a function to do something bearing a certain relation to something else. The chameleon's skin-color mechanism has the relational function of making the chameleon's skin match that of its environment (whatever that color may be). Given a specific color to adapt to, the mechanism then acquires an adapted proper function. So when it's sitting on a brown plant the adapted proper function of the device is to produce brown skin. Crucial here is that the brown skin may have never been produced before, so it is not a member of a ref., and the production of it is not a direct proper function of the device. That to which an adapted device adapts is its adaptor (the brown plant being the adaptor in this case). One can extend this picture to include the functions of those who may be consumers of the output of an adapted function. Consider the dance of the bees which acts as a signal to other bees, 'telling' them to go southeast, where the nectar is to be found. This dance is adapted to the location of the nectar, so has an adapted function. The response it calls forth, the flight of the bees to the southeast, is adapted to this adapted function. It is not a short step to extend this approach to the communicative linguistic and other cultural capacities of human beings. Complications abound, but the fundamentals for doing this are in place. For the extension to be seen as a realistic prospect, two important features of these extensions of the historical approach need to be noted. The first is that functional explanation includes a normative aspect: a functional property is there in order to produce an effect, the effect being one it ought to cause, given that the property is instantiated here because past instantiations have produced that effect. Malfunctioning items are not doing what they ought to do. The extent to which functional explanation can account for psychological and sociocultural properties is partly dependent upon whether the normativity intrinsic to these latter properties can be explicated by the norms of functional explanation.

The second is that the above extensions indicate the explanatory reach of that theory. Any items (including

patterns of behavior (e.g., rituals), social structures, ideas, 'memes'), can be the bearer of functions, provided that the basic conditions for this are met: variety, selection, varied reproduction. The reproduction need not be gene-centered, nor even broadly 'biological.' The transmission of bodies of knowledge and ways of behaving across generations is a form of reproduction, and so these items are possible bearers of functional properties. Whether they are or not depends on whether selection operates on these reproduced items, and here it is important to note that selection can take many forms. As behaviorists emphasized, behavior can be molded by painful experience, so pain is one selector. Emotional experience can be another. These have been called 'vicarious' or 'secondary' selectors (Campbell 1960), indicating that they themselves have been selected for by natural selection, their purpose being to be local, less severe eradicators of those types of behavior that are inimical to the flourishing of the organisms which produce the behavior. Given that reproduction and selection can be ubiquitous, it would appear that functional explanation has a much wider reach than once thought.

There have, though, been suggestions both that the explanatory scope of functional explanation has become too broad, that it is being abused by those who tend to see functionality everywhere, and that it is too narrow, not encompassing explanations of items which have no past history of selection, but are nevertheless beneficial to the organism. In the next section these criticisms of the historical theory are reviewed.

4. Difficulties and Additions

4.1 Panglossianism

The etiological theory is supposed to explain the persistence and spread of a trait in terms of its adaptive history, such explanations involving historical claims as to why the feature being explained was adaptive, and then tying this adaptation to its present function. At times it is extremely difficult to trace the relevant history, and the temptation is to ascribe functionality anyway, with the hope that the historical detail will be forthcoming. This has led to the accusation that such explanations are 'Panglossian,' presuming that this world is the 'biological best' world, one in which all features of organisms have some function or other (see Gould and Lewontin 1978). This accusation is particularly relevant to the sociobiological explanations of psychological and social behaviors, as the relevant history supporting these explanations is often unknown. The complaint is a welcome caution against epistemologically unconstrained attributions of function, but does not attack the heart of the historical theory. It simply warns against its abuse.

4.2 *Exaptations*

A related worry about attributing functions incorrectly, also voiced by Gould (Gould and Vrba 1982), concerns a trait that may be adaptive for one reason, but then change function in mid-stream, either doubling up in its functions or (because of environmental change maybe) sloughing off the old function in favor of the new, without any change of structure. Now the present persistence of the trait cannot be explained via its previous adaptive history, as that would give a misleading account of its present function. Such a change in function has been called an exaptation. An example of such an exaptation is the wing feathers of birds, the function of which we now think is to enable birds to fly. A selectionist story, however, would have to say that their function was as a thermal regulator, as that is what they were selected for originally. Clearly some complication of the original selectionist theory is called for, but the complications are hardly unexpected. Causal explanation and accounts of causality in general have difficulties dealing with causal overdetermination, and what we have in the above examples is functional overdetermination. What we need to say, in the functional case, is that the functions of an item are those fitness enhancing effects which advantage (or advantaged) the bearer of the item relative to rivals, and which have been reproduced differentially in the relevant population. With the change of function of feathers from thermal regulation to flight, the stage at which some feathers produced a flight-response would be critical in noting the acquisition of the new function. As soon as the feathered and flighted birds became preponderant, then one can say that the feathers had acquired a new function.

4.3 *Dispositional and Virtual Selection*

Once one notes that selection can be recent, a question arises as to how recent selection can be for functionality to be correctly attributed. To some extent the answer will be vague; there seems no exact stage at which one can say: before now there was no function, but now there is. But given that selection can be recent, a natural suggestion will be that present selection suffices for functionality. The radical proposal is that history can be irrelevant, that the essential component in the historical theory is the contribution of a feature to the reproductive success of the item possessing that feature, and that a trait without any history of selection can be seen to be useful for its possessor in the precise sense of making a (future, or possible) positive contribution to the survival or reproductive capacity of the organism. This 'dispositional' or 'virtual' selectionism involves the recognition that selection does play some role in the attribution of functionality, but it need not be actual selection. Differential reproduction need not have taken place. On this

account, a trait is functional because of its potential contribution to the fitness of the organism possessing the trait, given the environment in which the organism operates.

A weak version of this view can be developed by requiring the environment to engage actively with the trait, so that the trait can be seen to contribute positively to the organism's fitness. On this view, suitably developed, the trait is seen as being currently under a selection pressure. That is, it is seen as potentially contributing to the differential reproduction of those organisms with the trait, where 'potentially' only means 'barring accidents, it will contribute to the differential reproduction of those organisms.' So the weak view agrees with the historical view in all respects except that it allows for functionality prior to differential reproduction. The stronger interpretation does not require actual engagement with the selection-relevant environment; it only requires that there are factors present in the environment that could influence the organism's chances of reproduction and survival.

The crucial question facing these amendments to the historical theory is: what is the advantage in using the notion of function here? What does it add to a simple nonfunctional causal explanation of what is happening? On the historical view, part of the distinctive force of functional explanations derives from the normative elements implicated in attributions of function, the idea that the functional trait is there in order to do something. The selectionist story, cashed out in terms of differential reproduction, makes this normativity naturalistically respectable. Without the selection occurring, this distinctive feature of functional attribution is threatened.

It is clear that these accounts cannot give any functional explanation for the existence, proliferation, or shape of the trait. Take the first generation of long-necked giraffes. The existence of the long-necks (their proliferation) cannot be given an explanation relating to the reproductively beneficial effect which the long necks have had in the past. The only explanation seems to be a straightforward causal explanation in terms of mutations. A suggestion is that one change the explanandum, which is what Pettit (1996) does in his advocacy of virtual selection. The idea is that there can be an explanation of a fact that is not an explanation of its existence or actual persistence, but an explanation of why it would persist in certain nonactual conditions, an explanation of its resilience. This would be an explanation of the robustness of the fact under various hypothesized circumstances. An example given is that of equilibrium explanation, where the explanation of, say, a certain ratio of prey to predator in a given environment need not explain why this ratio (which is in fact an equilibrium) has come about—why it exists—but why, if disturbing influences were brought to bear upon it, it would return eventually to this ratio. It is in this sense that the

resilience of the pattern is being explained. With functional explanation, the thought is that the absence of the operation of the function, the absence of the 'good' effect, may itself initiate certain mechanisms or processes that will reinstate the function's effect. So independently of how the trait performing the function emerged or proliferated, so long as it has this modal property of being re-instatable (in certain conditions) when disturbed, then the attribution of function will be in order. There will be something for that attribution to explain: namely the resilience of the pattern.

The problem for the advocate of functions here is that a normal causal story (perhaps one with feedback built into it) seems adequate to the task of explaining what is going on. The notion of function adds nothing, or at least, the plausibility it has derives only from the systems idea of functions. There is a system here which works so as to maintain an item, or a pattern of behavior, and any subsystemic processes which work so as to maintain the item or behavior 'functions' to that end. There is nothing more that functional explanation does here, and this notion of functional explanation does not have the distinctive normative character that the historical account imparts to functionality.

4.4 The Explanandum

There is some debate about what it is that a functional explanation explains. One suggestion is that, given selection only 'chooses' among the traits which are in existence, favoring some at the expense of others, all it can explain is why there are (so many) items with that feature. The proposal is that selection can not explain how individuals come to have the traits that they do have. The explanation for the traits possessed by any individual would be simply the causal story instantiated in the genetic processes that produced the trait in that individual. Sober (1984) called this latter type of explanation developmental explanation, and the good point he was making was that selection operates on traits already in existence.

One can accept this point, though, and still deny that all that functional explanation can explain is the very general fact concerning the proliferation rates of the causally explained traits. Neander has pointed out that that each such trait will have an evolutionary history, the elaboration of which will show how the trait came to have the form it now has (Neander 1995). Given the falsity of saltationist doctrines, traits such as opposable thumbs or long necks did not emerge fully fledged; there were many mutations, and those which increased, say, opposable thumbness were favored. A restriction to looking at 'selection of pre-existing traits' has almost an air of saltationism about it. It is as though the traits came into existence formed as they now are, and were then selected for on that basis. Ignoring the more gradual, at times almost imperceptible, changes that took place in order to arrive at

the present trait is to ignore a part of the explanation as to why the trait is what it is. It is because functions can be performed by different mechanisms that attending to the historical detail is important in showing how and why this mechanism evolved into its present form. The central point is that although selection does only distribute characteristics, the history of successive distributions shows selection as also molding those characteristics. It is then possible to explain not only why people with thumbs were selected, but also the shape of the thumb, by citing facts about the shapes of its predecessors.

See also: Adaptation, Fitness, and Evolution; Causation: Physical, Mental, and Social; Causes and Laws: Philosophical Aspects; Evolutionary Theory, Structure of; Explanation: Conceptions in the Social Sciences; Functionalism, History of; Functionalism in Sociology; Methodological Individualism in Sociology; Natural Selection

Bibliography

- Campbell D T 1960 Blind variation and selective retention in creative thought and in other knowledge processes. *Psychological Review* 67: 380–400
- Cummins R 1975 Functional analysis. *Journal of Philosophy* 72: 741–65
- Cummins R 1983 *The Nature of Psychological Explanation*. MIT Press, Cambridge, MA
- Gould S Lewontin R 1978 The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London, series B: Biological Sciences* 205: 581–98
- Gould S J, Vrba E S 1982 Exaptation—a missing term in the science of form. *Paleobiology* 8(1): 4–15
- Hull D, Ruse M (eds.) 1998 *The Philosophy of Biology*, Part IV. Oxford University Press, Oxford, UK
- Millikan R G 1984 *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT, Cambridge, MA
- Millikan R G 1993 *White Queen Psychology and Other Essays for Alice*. MIT, Cambridge, MA
- Neander K 1995 Pruning the tree of life. *British Journal for the Philosophy of Science* 46: 59–80
- Pettit P 1996 Functional explanation and virtual selection. *British Journal for the Philosophy of Science* 47: 291–302
- Sober E 1984 *The Nature of Selection: Evolution Theory in Philosophical Focus*. MIT, Cambridge, MA

G. Macdonald

Functionalism, History of

Functionalism is a method of sociological analysis which has its roots in nineteenth-century social thought. It originated from the application of methods from biology, the rising science at that time, to the analysis of society. Three features are characteristic of

this orientation toward biology: society is more or less strictly conceived as an organism; its development is understood as an evolutionary process; special attention is given to the relationship between the parts and the whole constituted by them, particularly to functions which have to be fulfilled by the parts for the working of the whole. The parts are institutions such as forms of family life, economic production, government and religion as well as the organizational embodiments of such institutions: families, enterprises, governmental bodies, and churches. The institutions are binding rules which regulate the activities and relationships between people. The functions are, for example, biological reproduction for the family, generating resources and allocating them to needs for economic production, binding decisions for the collectivity for the government or the production of meaning for religion. It is assumed that, for its simple continuation, society needs the subserving of such functions by institutions and their organizational embodiments.

1. *Origins of Functionalism: Auguste Comte and Herbert Spencer*

It was the fundamental change from traditional agrarian to modern industrial society which constituted the background for the rise of functionalism in nineteenth century social thought, particularly in the work of Auguste Comte and Herbert Spencer. Comte (1830–42/1969) carried on the system developed by his former teacher, Saint-Simon, from whom he separated after having served him in his younger years. In his *Social Statics*, Comte conceived of society as a whole which, in its ideal state, is kept in perfect harmony by the balanced contribution of its constituent parts to its working in the process of fulfilling their specific functions. In his *Social Dynamics*, he conceptualized social change as an evolutionary process from lower to higher forms of societal organization. Because he expected science to bring about the perfect organization of social life in its final stage in modernity, he couched the evolutionary process in terms of the evolution of thinking from the theological to the metaphysical to the positive stage shaped by 'positive' science.

The other major source of functionalism in nineteenth century social thought next to Comte is Herbert Spencer (1909). Though he shared the adoption of biological and evolutionary thought with Comte, he demarcated his approach explicitly from Comte's by conceiving society and social development not as a matter of ideas, but as a matter of the objective things, namely the behavior of human individuals. The major process in the development of society is differentiation; functions become more numerous and the tendency is a function that is fulfilled by one specific institution specialized in it. The more society grows in size, the

more differentiated will be its functions and corresponding institutions. The first differentiation occurs between regulative, sustaining, and distributive functions. The corresponding institutions are government, economy, and the infrastructure of channels for transport and communication. With the differentiation of functions and institutions, society works like higher organisms, it becomes more heterogeneous, definite in the demarcation of institutions, and integrated; group life is transformed from the coexistence of groups with the same structure without coherence between them (families, clans) to greater heterogeneity but also greater coherence between groups. The evolution of social life goes from incoherent homogeneity to coherent heterogeneity in the process of specialization and growing division of labor (see *Evolutionism, Including Social Darwinism*).

2. *The Classical Framing of Functionalism: Emile Durkheim*

Among the great classical founders of sociology at the turn of the nineteenth to the twentieth century, it is Emile Durkheim (1893/1973a, 1897/1973b, 1912/1968) who was most influenced by Comte and Spencer and who made a distinctive contribution to functionalism as a method of sociological analysis. From Comte he borrowed the holistic view of society, from Spencer the idea that the growing division of labor is a core feature of social evolution. In his *Rules of the Sociological Method*, Durkheim made a clear distinction between functional and causal analysis and a clear statement that every functional analysis needs to be complemented by causal analysis. He argues that the utility of an institution in serving a need for individuals or in performing a function for society does not suffice to explain its existence.

In his study on *The Division of Labor in Society*, Durkheim points out that specialization and division of labor are tools for the self-preservation of the individual and the society and for integrating society. They help to overcome the struggle for scarce resources. This struggle is caused by the increasing material and dynamic density that results from population growth and the extending as well as denser weaving of ways and means of transport and communication. Nevertheless, serving self-preservation and social integration do not produce specialization and the division of labor by themselves. In Durkheim's eyes, this functional explanation of specialization and division of labor needs a complementation by causal analysis. In doing so, he looks for forces which lead the selection of behavior and individuals in the direction of increasing labor division. He says that emigration, suicide, and crime would have been other possible solutions of the problem besides labor division. He explains the fact that they have not been selected to the same extent by the greater resistance called for by

them. The bonds to one's home country work against emigration, the bonds to one's life against suicide, the sympathy for our fellows prevents us from committing a crime. They are on average stronger and more resistant feelings than the habits which could prevent us from further specialization. The latter had to give way to the pressure toward specialization. In doing so, Durkheim points out the 'functional equivalents' which would serve self-preservation and then gives answers to the question as to the forces which lead human behavior in the direction of one of them. This is his complementation of functional analysis by causal explanation in his study on the division of labor.

In his later work on *The Elementary Forms of Religious Life*, Durkheim draws on anthropological studies on Australian aborigines. Here he makes use of his idea of the circular link between cause and functional effect in the study of ritual practices in which the sacred is strengthened in distinction to the profane and in which the community is regularly becoming conscious of itself. We have another circular link here. People gather because they want to influence the demonic powers positively for practical reasons of a good harvest or hunt, their willingness to come together is, however, supported by their feeling of belongingness and solidarity, which is again strengthened by the ritual itself.

3. *A Paradigm of Functional Analysis: Robert K. Merton*

Durkheim's analysis of tribal societies was continued in the anthropological study of tribal societies, particularly by the leading British anthropologists, Alfred R. Radcliffe-Brown and Bronislaw Malinowski. According to Malinowski (1948), every social phenomenon has to be explained by an indispensable function it fulfills not only for the whole of society, but also for the human individual: biological needs for food, shelter, and reproduction, psychological needs for support, needs for maintaining culture and social organization. Institutions of exchange, decision-making, social control, and education should be explained by such functions. Radcliffe-Brown (1952) criticized Malinowski's emphatic functionalism and preferred to call his approach structuralism. He did not accept the presumption that every social phenomenon exists because of a necessary function it fulfills. In his view, the unity of society was not a postulate claimed for every society, but a hypothesis which allowed a functional explanation of a social phenomenon by its contribution to the persistence of society. One of his examples is the explanation of lineage systems by the fact that they help to avoid conflicts about property rights after the death of property holders and thus contribute to the maintenance of solidarity. It is obvious that such an explanation is nevertheless insufficient in the light of

Durkheim's methodological objections against purely functional explanations.

At this point, it is quite reasonable to turn to the detailed methodological analysis of functionalism contributed by Robert K. Merton (1949/1968) in his paradigm of functional analysis. Merton points out three postulates of functionalism, which all have to be doubted in his eyes. (a) Functional unity postulates that society is one consistent whole and every phenomenon contributes to its working and that there is no division of society into parts with conflicting needs. (b) Universal functionalism postulates that every social phenomenon performs a positive function for the whole of society. (c) Functional indispensability postulates that every social phenomenon performs an indispensable function for the whole of society and cannot be replaced by some other phenomenon. Merton says that all three postulates are unfulfilled by concrete societies, particularly by modern societies, so that functional explanations of social phenomena cannot be based on them.

Instead, Merton introduced a more realistic paradigm of functional analysis with a set of 12 requirements. (a) Only standardized social phenomena like institutions should be analyzed with regard to their functions. (b) Subjective dispositions of actors and objective consequences of their action in social practices should be outlined and distinguished. (c) Positive, negative (dysfunctional) and neutral effects of standardized practices should be discerned. (d) The units for which an item fulfills a function have to be exactly determined. (e) The functional requirements for maintaining a system have to be denoted. (f) The mechanism of their fulfillment must be specified. (g) The functional alternatives have to be enumerated. (h) The structural context needs description regarding the narrowing down of space for functional alternatives. (i) Dysfunctions have to be detected and analyzed with regard to effecting stress and strain with resulting pressures for change. (j) The functional explanations should be validated in comparative research. (k) Functional analysis must be neutral in ideological terms. (l) Functional analysis of a social practice needs to include; (i) the location of participants in the social structure; (ii) alternative modes of behavior excluded by the pattern; (iii) the emotive and cognitive meaning of the pattern for the participants; (iv) a distinction between motivation and objective behavior; (v) behavioral regularities that are part of the pattern but are not recognized by the participants.

A good example of Merton's idea is his functional explanation of the political machine by the various needs it fulfills for the whole of society and for a variety of groups. As an informal network with a boss at its head, it concentrates power and thus makes political decisions possible in the American political system of checks and balances; it also serves as a vehicle for helping the poor and weak, legitimate and illegitimate business, and provides channels for

upward mobility. Thus the political machine becomes stabilized because—according to Merton's requirement—the positive functions outweigh the dysfunctions by far.

4. *Functional Prerequisites and Social Stratification: Aberle et al., Davis and Moore*

In the 1940s and 1950s, functionalism achieved the dominant paradigm of sociological analysis under the name of 'structural functionalism.' The investigative strategy of structural functionalism was the discovery of the functions which have to be fulfilled in order to secure the survival of society and the discovery of the corresponding structures serving these functions (Aberle et al. 1950/1967). Another representative contribution to structural functionalism is the functional theory of social stratification outlined by Kingsley Davis and Wilbert E. Moore (1945) in their much debated article 'Some Principles of Social Stratification' published in the *American Sociological Review* in 1945. Davis and Moore argue that every society needs a system of social stratification for the following reasons. (a) The members of the society share common views on the importance of a certain position needed for the provision of public and private goods and services. (b) To perform the corresponding roles, more or less talent, training and/or effort is required. (c) Thus, positions can be more or less important, the personnel needed for filling the positions more or less scarce. In order to fill the positions there must be a differentiation of rewards in terms of money or prestige according to the importance of the position and the scarcity of the personnel available. The greater the importance and scarcity, the greater the reward. The result is that positions are always stratified. The representative criticism of this functional theory of stratification was formulated by Melvin M. Tumin (1953) in an article published in the *American Sociological Review* eight years later, in 1953. It is couched in terms of conflict theory: the class structure is based on the power of the dominant class, which defines the importance of positions; its major effect is the differentiation of access to the higher positions according to socialization advantages and training of the offspring originating from the dominant class. Rather than contributing to the working and integration of society, the stratification system has disruptive effects so that society is not stabilized by value consensus and stratification, but by the application of power.

5. *Functionalism as an Attempt at Synthesis: from Talcott Parsons' Structural Functionalism to Neofunctionalism*

Five years earlier than Davis and Moore, in 1940, Talcott Parsons published another functional theory

of social stratification, which he revised twice, in 1954 and 1975 (Parsons 1954a, 1954b, 1977). In the perspective developed by Parsons, social stratification has to be explained by the interaction of dynamic and controlling forces of action, interests, and power on the one hand, and solidarities and cultural values on the other. Interests and power determine the access to positions and the enforcement of positional differences and outcomes; particularistic solidarities of family and lineage explain the inertia of class structures because of the tendency of families to secure their offspring remaining in the same class position. The *legitimacy* of the stratification system in distinction to its factual production and reproduction is, however, an effect of the consensus on a grading scale according to which prestige is attributed to positions. Only in as much as such a consensus on a grading scale is achieved will the system of stratification be accepted as legitimate. Otherwise, the definition of the importance of positions and the access to them will be a matter of the war of all against all. In general terms, this is the problem of double contingency.

Other than Merton's deliberate plea for empirical middle range theories, Parsons' program was analytical grand theory. Merton stood for empirical functionalism, Parsons for analytical functionalism (Parsons 1951). He introduced his four-function paradigm as an analytical instrument of analyzing structures as well as dynamic processes with regard to abstract functions (Parsons 1966, 1967, 1978, Parsons and Smelser 1956). By cross-tabulating the instrumental/consummatory and the internal/external orientation of action, he defined four abstract functions: instrumental/external: adaptation (A); consummatory/external: goal attainment (G); consummatory/internal: integration (I); instrumental/internal: latent pattern maintenance (L). For further differentiation, Parsons distinguishes three levels of applying the four-function scheme: social system: economic (A), political (G), community (I), fiduciary (L) system; general action system: behavioral (A), personality (G), social (I), and cultural (L) system; human condition: physico-chemical (A), organic (G), action (I), and telic (L) system. Further differentiation results from the application of the scheme to one of the subsystems. After having developed an analytical action theory in his first work of 1937, *The Structure of Social Action* (Parsons 1937/1968), Parsons established structural functionalism as a link between analytical functionalism and systems theory in *The Social System* in the early 1950s, and extended that link with the introduction of his theory of generalized media of interchange (money (A), political power (G), influence (I) and value-commitments (L) for the social system) and his theory of evolution as instruments for the analysis of dynamic processes of inflation and deflation and social change in the 1960s. With his turn to functionalism, systems, media, and evolutionary theory he did, however, not give up the core position

of his early action theory: the idea that social order emerges from the interaction between controlling and dynamizing forces. In his systems theory, he couched this idea in the double cybernetic hierarchy of conditions and controls. From L to A energy and dynamics in the system increase, while information and control decrease, and vice versa.

Looking at one analytical level, the social system, this system is composed of subsystems which fulfill their specific functions (A G I L) with specific structures (markets A, authority G, association I, and communication L) and specific media of interchange. They build an inter-related set of systems with in- and output processes of factors and products processed by the circulation of the media from one system to the other (interpenetration). Crises in the working of the systems can be analyzed as inflation–deflation spirals of the media involved.

In his theory of evolution, Parsons tries to point out the evolutionary advantage of specific institutions. Religion, communication by language, social organization through kinship, and technology are the first conditions of building human societies. A breakthrough to higher forms of social organization is realized by social stratification according to prestige based on achievement instead of ascription by origin. The rigidity of kinship systems is pushed back in this way. A further step forward is the replacement of tradition by cultural legitimation through general values. The developmental stage of a modern society is realized by the establishment of bureaucratic organizations, markets and money, universalistic norms, and democratic association. In two volumes, he outlines the evolutionary process in four dimensions according to his AGIL scheme: A adaptive upgrading, G structural differentiation, I inclusion and L value generalization.

Evolutionary achievements require breakthroughs in all four dimensions. He points out such breakthroughs in his analysis of primitive societies, archaic societies (Egypt and Mesopotamia), intermediate historic empires (China, India, Islamic, and Roman empires), seedbed societies (Israel, Greece) and, finally, the system of modern societies from ancient cultural roots via the Renaissance and the Reformation to the Enlightenment and the building of modern citizenship, state, and capitalism. The major modern breakthroughs are the industrial, the democratic, the educational, and eventually the 'expressive' revolution of the 1960s and 1970s (Parsons 1966, 1967, 1971).

According to his position that social order always has to reside in some form of consensus and solidarity in order to avoid an unlimited power struggle, Parsons lays special emphasis on the development of citizenship as a core of a modern 'societal community.' It is a modern form of community which has the greatest potential for inclusion, it associates people of the most varied ethnic and cultural origins. In this respect, it is

the cornerstone of modern social organization. He sees the society of the United States farther developed in this respect than the ethnically more homogeneous European welfare states. The modern community of citizens is the core which is complemented by a capitalist economy, democratic government, and universalistic institutions of education, science, professional work, and voluntary religious association. This is the abstract path of sociocultural evolution, which has to be distinguished from concrete historical development that can more or less deviate from this path. There is particularly always the danger of fundamentalist movements against modernity, which originate from the severe tensions caused by the evolution of modern forms of social organization. For Parsons, fascism as well as communism were such fundamentalist reactions to modernity, which have, however, no chance of long-enduring stabilization in his eyes. Therefore, he expected that the Soviet system would crumble some time, an expectation grounded in 'abstract' evolutionary theory which turned out to be very 'realistic' eventually.

The dominant role played by Parsonian functionalism in the 1950s provoked countermovements which concentrated mostly on some particular aspect of social organization, which was relativized within the very general Parsonian scheme. The need for revitalizing the Parsonian approach has called for a new movement, which aims at continuing with the general program of Parsons on the contemporary level of development. This is the movement of neofunctionalism which arose in the 1980s. Neofunctionalism tries to advance the Parsonian legacy by way of learning from new developments in social theory and social research, for example, from advancements provided by Clifford Geertz or Jürgen Habermas. This new movement has contributed fresh interpretations of Parsons' work along with a new reading of the classics, studies on civil society and public discourse, on the evolution of the culture of modernity, on communication society, European integration and globalization (Alexander 1982/83, 1998, Münch 1982, 1998).

6. Functionalism and Modern Systems Theory: Niklas Luhmann

A comprehensive renewal of functionalism has been undertaken by Niklas Luhmann (1970, 1984, 1997). He, too, combines functionalism with systems, media, and evolutionary theory. In his early phase of theory building, Luhmann introduced a turn from structural-functional to functional-structural systems theory. To justify this turn, he claims that the question of function has to be posed in a more fundamental way. Before we ask for the structures of systems which have to fulfill certain functions, we should raise the question as to which function is served by systems-building as such.

What is the problem solved by systems building? This is how he asks his question. His answer is: world complexity which needs reduction so that it can be handled. Systems just perform this function of reducing world complexity.

In his second phase of theory building, Luhmann undertook another turn of his functionalist systems theory: the conceptualization of social systems as autopoietic, self-referential systems. Because there is always the danger of too much complexity, the major problem is systems building and the continued maintenance of systems boundaries, which are boundaries of meaning. Autopoiesis is a form of systemic operation which guarantees the maintenance of boundaries and, at the same time, allows for maximum absorption of environmental complexity. An autopoietic system is operationally closed, but open to any kind of complexity in the environment. Operational closure is guaranteed by a binary code according to which any kind of environmental complexity is processed. The concrete processing is organized by a system's program. Special links to other systems in the environment are stabilized by structural coupling. Generalized media of communication allow for most flexible forms of systemic processing: money, power, law, love, for example. Social systems are built up by communication, they originate from that and continue to exist as long as communication takes place according to the established code. Communication consists of three elements: information, message, and interpretation.

An autopoietic system produces and reproduces all its elements—communication—by itself, that is by systemic communication, but it makes use of environmental complexity. It does not take inputs from the environment to transform them into outputs, but receives impulses from the environment to which it shows more or less response. The system observes events in the environment and absorbs them according to its own code and program. Problems of maladaptation to the environment can be discovered by second-order observation, that is, self-observation and self-description.

Luhmann distinguishes three levels of social systems: interaction systems, organizational systems, and societal systems. According to his theory of evolution, more complex societal systems evolve by variation, selection, and stabilization. There are three (later four) stages of societal evolution which are defined by the form of differentiation of systems: segmentary, estate (later: center/periphery), and functional differentiation.

Modern societies are functionally differentiated and no longer have a center. The economic, political, legal, science, educational, arts systems are more and more differentiated on a world scale and thus transcend their centering around a societal community and/or a state. Thus the nation-state societies are transitional stages on the way to world society. Paying vs. not paying, having power vs. having no power, right vs.

wrong, true vs. false are the binary codes of the economy, polity, law, and science systems. The evolutionary process works toward the establishment of functionally differentiated systems on a world scale, which are linked by structural coupling and by mutually stabilizing co-evolution.

7. Future Need for Functionalism

The question finally is: what will be the future of functionalism after its renewal by neofunctionalism and by Luhmann's functionalist systems theory? Whatever the movements against functionalism have contributed to our knowledge, we cannot say that they can replace the insights of functionalism completely. They have merely emphasized some particular aspect of social reality and some particular question. But they did neither raise nor answer the questions posed by functionalism from Comte and Spencer to Parsons and Luhmann. Because their questions are still legitimate, there is still a place for functionalist theorizing, whatever form it might take in the future.

See also: Action, Theories of Social; Biology's Influence on Sociology; Human Sociobiology; Comte, Auguste (1798–1857); Differentiation: Social; Durkheim, Emile (1858–1917); Evolution, History of; Explanation: Conceptions in the Social Sciences; Functional Explanation: Philosophical Aspects; Functionalism in Anthropology; Functionalism in Sociology; Law, Autopoiesis in; Luhmann, Niklas (1927–98); Malinowski, Bronislaw (1884–1942); Parsons, Talcott (1902–79); Social Stratification; Spencer, Herbert (1820–1903); System: Social

Bibliography

- Aberle D F, Cohen A K, Davis K, Levy M J Jr., Sutton F X 1950/1967 The functional prerequisites of a society. In: Demerath N, Peterson R (eds.) *System, Change and Conflict*. Free Press, New York, pp. 317–31
- Alexander J C 1982–83 *Theoretical Logic in Sociology*, 4 vols. University of California Press, Berkeley, CA
- Alexander J C 1998 *Neofunctionalism and After*. Blackwell, Oxford, UK
- Comte A 1830–42/1969 *Cours de philosophie positive*, 6 Vols. Culture et Civilisation, Brussels (1855 *The Positive Philosophy of Auguste Comte*. Calvin Blanchard, New York)
- Davis K, Moore W E 1945 Some principles of stratification. *American Sociological Review* 10: 242–9
- Durkheim E 1893/1973a *De la division du travail social*. Presses Universitaires de France, Paris (1964 *The Division of Labor in Society* [trans. G Simpson]. Free Press, New York)
- Durkheim E 1895/1973b *Les règles de la méthode sociologique*. Presses Universitaires de France, Paris (1982 *The Rules of the Sociological Method* [ed. Lukes S, trans. Halls W D]. Macmillan, London)

- Durkheim E 1912/1968 *Les formes élémentaires de la vie religieuse*. Presses Universitaires de France, Paris (1976 *The Elementary Forms of Religious Life* [trans. J Swain W]. Allen & Unwin, London)
- Luhmann N 1970 *Soziologische Aufklärung*. Vol. 1. Westdeutscher Verlag, Opladen (1982 *The Differentiation of Society* [trans. Holmes S, Larmore C]. Columbia University Press, New York)
- Luhmann N 1984 *Soziale Systeme. Grundriß einer allgemeinen Theorie*. Suhrkamp, Frankfurt am Main
- Luhmann N 1997 *Die Gesellschaft der Gesellschaft*, 2 Vols. Suhrkamp, Frankfurt am Main
- Malinowski B 1948 *Magic, Science and Religion*. Doubleday, New York
- Merton R K 1949/1968 *Social Theory and Social Structure*. Free Press, New York, pp. 73–138
- Münch R 1982 *Theorie des Handelns. Zur Rekonstruktion der Beiträge von Talcott Parsons, Emile Durkheim und Max Weber*. Suhrkamp, Frankfurt am Main
- Münch R 1998 *Globale Dynamik, lokale Lebenswelten. Der schwierige Weg in die Weltgesellschaft*. Suhrkamp, Frankfurt am Main
- Parsons T 1937/1968 *The Structure of Social Action*. Free Press, Glencoe, IL
- Parsons T 1951 *The Social System*. Free Press, Glencoe, IL
- Parsons T 1954a *Essays in Sociological Theory*. Free Press, New York, pp. 69–88
- Parsons T 1954b *Essays in Sociological Theory*. Free Press, New York, pp. 386–439
- Parsons T 1966 *Societies. Evolutionary and Comparative Perspectives*. Prentice Hall, Englewood Cliffs, NJ
- Parsons T 1967 *Sociological Theory and Modern Society*. Free Press, New York
- Parsons T 1971 *The System of Modern Societies*. Prentice Hall, Englewood Cliffs, NJ
- Parsons T 1977 *Social Systems and the Evolution of Action Theory*. Free Press, New York, pp. 321–80
- Parsons T 1978 *Action Theory and the Human Condition*. Free Press, New York
- Parsons T, Smelser N J 1956 *Economy and Society*. Free Press, New York
- Radcliffe-Brown A R 1952 *Structure and Function in Primitive Society*. Macmillan, New York
- Spencer H 1909 *The Principles of Sociology*, 3 Vols. Appleton, New York
- Tumin M M 1953 Some principles of stratification: a critical analysis. *American Sociological Review* 18: 387–94

R. Münch

Functionalism in Anthropology

In its simplest common-sense form functionalism in anthropology describes approaches which purport to explain cultural or social phenomena in terms of the purpose they serve.

Since the 1960s, specialized research into the sociological and historical framing of anthropology has revealed the extent to which discussions of functionalism contribute to analysis of knowledge as culture. As yet, however, it falls short of meeting the criticism that it tends to treat the discipline of

anthropology as a closed discourse, despite calls that it be viewed as part of a larger discursive field in which anthropology operates (Asad, in Stocking 1991; Trouillot, in Fox 1991).

1. The Intellectual History of an Idea

The term ‘function’ was first introduced as a mathematical concept by Gottfried Wilhelm Leibniz in 1714, to refer to a variable considered in relation to one or more other variables in terms of which it might be expressed.

Two main bundles of ideas underlay the concept of ‘function’ that dominated academic anthropology in the first half of the twentieth century: (a) a relation of interdependence and (b) an orientation toward given ends. The first involved reciprocity and reversibility in terms of cause and effect. The ends might be the needs of the individual, as in the utilitarian functionalism of Malinowski (1884–1942) or the needs of society, as in the structural-functionalism of Radcliffe-Brown (1881–1955). Both adopted notions of culture or society in the modern pluralized usage as totalities with a certain unity and integrative character, a notion derived from the German philosopher Johann Gottfried Herder.

In the second half of the century, the premise remained that cultures and societies were structured wholes, now more consistently referred to as cultural or social systems. At this point functionalism’s orthodoxy came under challenge in the aftermath to the Second World War which saw the withdrawal of sovereignty claims over a plethora of colonial territories within the British Empire and the establishment of United States ‘trusteeship’ in Oceania. This spurred British functionalists to address complex ‘new nations’ and provided American anthropologists *de novo* with a laboratory of functioning colonial cultures (Bashkow, in Stocking 1991).

By the 1970s, functionalism (as an achieved, shared paradigm which defined problems and procedures) was seen to be outmoded. It was replaced, amidst repeated calls for anthropology to reinvent itself, by struggle among new modes of discourse, chief among them Levi-Straussian structuralism, Marxism, and symbolic and interpretive anthropology. Nevertheless, echoes of functionalism lingered on into the decentralized, fragmented, pluralizing, flexible, refractive, postmodern world of the new millennium.

2. Functionalism’s Multivocality

A distinction among (a) functionalism as doctrine (or theory); (b) functionalism as method; and (c) functionalism as interest or ideology was a critical feature in Bronislaw Malinowski’s modernist legacy to functionalism (Vincent 1991). This, along with the essential dualism between natural science and the humanities (particularly science and history) that entered mod-

ernist anthropology with Franz Boas's *The Study of Geography* (1897, reprinted in Stocking 1996), goes some way towards ordering the cultural complexity of functionalist knowledge. The scientific tendencies of the physicist who tries theoretically to reduce reality to simple elements and the humanist tendencies of the cosmographer who seeks to comprehend a complex phenomenon as a whole are both contained within the functionalist moment in anthropology.

2.1 Functionalism as Doctrine or Theory

Three cross-cutting anthropological trajectories may be recognized within functionalist anthropology.

2.1.1 Utilitarian functionalism (Malinowskian). Recent research has exposed the extent to which Malinowski's utilitarian functionalism, which dominated British anthropology between the wars, derived from his engagement with the neopositivism of Ernst Mach in his doctoral dissertation—particularly Mach's view that knowledge is a response to biological human needs (Thornton 1985; Thornton and Skalnik 1993). In his last, posthumously published, position paper, *The Functional Theory*, Malinowski takes his stand on two axioms: first, that every culture must satisfy the biological system of needs, and second, that every cultural achievement that implies the use of artifacts or symbols is an instrumental enhancement of human anatomy, and refers directly or indirectly to the satisfaction of a bodily need (Malinowski 1944, p. 171).

Strictly followed, these axioms would restrict cultural anthropology to the economism of the rationalizing individual (human nature), or the ecologism of selective advantage (external nature) (Sahlins 1976), but Malinowski's practice deviated considerably from his theorizing. As with Boas, his sentient functionalism lay immanent in his ethnography; neither man was deemed 'a theorist.'

2.1.2 Hybrid cultural functionalisms (Boasian/Parsonian). This complex functionalist trajectory, triggered by Franz Boas's (1858–1942) 'vaguely functionalist ideas' (Sahlins 1976, p. 71), drew explicitly on German humanist intellectual traditions. It comprised several contradictions which manifested themselves transgenerationally. Initially this took the form of Alfred Kroeber's (1876–1960) functionalist historicism and Robert Lowie's (1883–1957) neopositivist empiricism. A second generation was further divided; it was also developing in the context of the 'British' functionalist paradigm.

Thus, in 1934, Alexander Lesser (1902–82), recognizing functionalism as the most vigorous tendency extant in anthropology, set out to distinguish the particular interests and conceptions of individual

American functionalists from the broader, foundational Boasian functionalist approach. He adopted a Leibnizian mathematical definition of a functional relation, established by observation, while arguing that contemporaneous events associated by the observer might be merely coexistences. Functional historicity recognized that 'institutions, customs, beliefs, artifacts have careers in time, and that their form and character is molded more by what has happened to them in the course of that history than by what particular things they occur associated with at any one time' (Lesser in Mintz 1985, p. 60). Radcliffe-Brown's position paper 'On the concept of function in social science' (1935) was prompted by his taking Lesser's critique personally.

It was actually directed at Lesser's fellow Boasians at Columbia University in New York, Ruth Benedict (1887–1948) and Margaret Mead (1901–78). Romantic holistic sensibility figured most prominently in Benedict's cultural configurations or patterns of culture. Mead, on the other hand (after corresponding with Radcliffe-Brown), designed a framework for the comparative functional analysis of social systems. This was published in *Cooperation and Competition among Primitive Peoples* (1937).

At the same time a more scientific anthroposociological functionalism was being developed at Harvard where sociologist Talcott Parsons (1902–79) joined forces with anthropology to set up a Department of Social Relations. One of its earliest products was arguably the most thoroughgoing functionalist ethnography ever published, Conrad Arensberg and Solon Kimball's *Family and Community in Ireland* (1940). Under the direction of W. Lloyd Warner (formerly Radcliffe-Brown's student) this extended functionalist anthropology from simple/primitive to complex/modern cultures. Later critics faulted the work for presenting as eternally cultural what was no more than a well-observed historical moment in county Clare.

2.1.3 Structural-functionalism (Radcliffe-Brown). Structural-functionalism's core concepts are, in harness, structure and system. Under the auspices of Radcliffe-Brown (1881–1955) it emerged as a clear-cut alternative to hybrid Boasian functionalism and successor to Malinowski's particular brand of economism and radical individualism. It derived not from German and British precursors but mainly from Durkheimian French sociology with supplementary insights from the Russian geographer and anarchist, Peter Kropotkin (1842–1921).

Structural-functionalism emphasized the formal ordering of parts and their functional interrelations as contributing to the maintenance needs of a structured social system. The function of any institution (or 'recurrent social activity') was the part it played in the maintenance of the larger structural whole. This

assumption attributed to social systems an internal integration of parts similar to that found in organisms.

Whereas Malinowskian functionalism was seen to have outworn its usefulness by 1945, and American cultural functionalism remained factionally divided, Radcliffe-Brown's successors, particularly Meyer Fortes (1906–83), Edward Evans-Pritchard (1902–73), and Max Gluckman (1911–75), expanded structural-functionalism to address situation, selection, and history while continuing to work implicitly with equilibrium models of processes internal to the socio-cultural system. Marxisant anthropological analyses of contradictions in structure and function supplemented structural-functionalism with a compatible, historically oriented, approach to competition, conflict, and change.

2.2 Functionalism as Method

Malinowski claimed that the so-called functional method in modern anthropology constituted a specific and necessary method of knowledge production. The relationship between theory and practice was reciprocal. Practice consisted of the intensive fieldwork that had emerged in contradistinction to the scientific survey methods of the Cambridge University Torres Straits Expedition of 1898–9, and the writing up of research findings in monographic form. The 1912 edition of *Notes and Queries*, that Malinowski carried with him into the field, contained W. H. R. Rivers, (1864–1922) delineation of 'intensive' fieldwork in which 'the worker lives for a year or more among a community of perhaps four or five hundred people and studies every detail of their life and culture; in which he comes to know every member of the community personally; in which he is not content with generalized information, but studies every feature of life and custom in concrete detail and by means of the vernacular language' (Rivers 1913, p. 7, quoted in Urry 1993, p. 28). This became functionalism's ideal mode of production.

Functionalism's methods of reproduction (knowledge as culture) included grantsmanship, correspondence, seminars, public events, unpublished and published lectures, oral histories, biographies, and the writing of monographs. Commemorations and conferences held in Cracow, London, and New York to memorialize Malinowski in 1984 (Gellner 1987) and in America to memorialize Boas in 1992 exposed a more complex functionalist heritage than that commonly recognized.

Paradigmatic adjustment was essential to the production and reproduction of functionalism. Thus, the last volume of Malinowski's Trobriand ethnography, *Coral Gardens and their Magic* (1935), documents the extent to which he was beginning to question 'the wholeness of culture,' criticize his neglect of the colonial order and of the inadequacy of the concept of culture contact or acculturation. He noted the possi-

bility of placing his own functional analyses of land tenure in the same frame as political controversy over land rights in Ireland and India (Vincent 1991). But such paradigmatic adjustment was rare. More routinely, it was the thick, detailed ethnographies functionalism engendered that empowered the reader to re-evaluate functionalism long before the textualization of theory became emblematic of paradigmatic change. The perdurance of a romantic ethnographic tradition contributed significantly to the persistence of inherent functionalist notions in anthropology even when the theoretical adequacy of functionalism remained an issue (Gellner 1987).

2.3 Functionalism as Interest: The 'Functionalist Movement'

Functionalism was both a concept and a scientific community contesting the constitution and distribution of cultural capital. Throughout the first three decades of the twentieth century in both Britain and America evolutionism, diffusionism and functionalism represented warring trajectories and interests and the paradigmatic dominance of functionalism was by no means secure (Vincent 1986). Explanations of dominance, as George Stocking has consistently demonstrated, are to be found not in the valency of ideas but in academic power, control, and institutionalization—to which observation must be added strategies of reproduction. In anthropology departments where field materials were presented in seminars (as at the London School of Economics and Manchester) 'field-notes were treated as facts, as concrete things, as sources of authority, as weapons in the struggle to gain the high ground of functionalist debate' (Bond in Sanjek 1990, p. 279).

3. Methodological Issues or Problems in Functionalism

As with other -isms, functionalism has been described most clearly and simply by those with most to gain from detracting interest from it. Yet anthropological functionalism has been characterized by constant immanent critique as well as external challenge. For critics it has been something of a moveable feast. What is noted depends on whose functionalism they are addressing and at what phase in that functionalist's thinking. Broad, generic charges laid against anthropological functionalism relate to positivism and pragmatism, on the one hand, and primitivism, on the other. Specific critiques have been directed against its holism along with its assumption of interdependence of parts, synchronism, homogeneity and the notion of collective consciousness, uniformity and its transgenerational replication through custom. Re-evaluations of structural-functional classics demonstrated that even so-called simple societies/cultures were internally

differentiated and heterogeneity gave rise to conflicting concerns among different sectors of the population. As the paradigm expanded further, new charges were made that functionalism could not accommodate the analysis of power.

4. Conclusion

The radical antipositivism of the 1960s put functionalism on trial. Functional anthropologists who declined to commit anthropology to the scrap heap argued ethnography's value as social history. Others argued its continued relevance for specific subfields such as ecological and applied anthropology.

It is not likely that future theoretical and research problems will follow the trends set by functionalism except, perhaps, in re-evaluations of problems in Foucault's adoption of functional terminology in his conceptualization of 'discourse'; in the work of anthropologists who holistically find instances of cultural resistance to what they perceive to be Gramscian style 'hegemony'; and in the relationship of anthropological functionalism to other modernisms (Webster, in Manganaro 1990).

The main thrust is likely to be the further undercutting of the two pluralized concepts integral to functionalism: 'cultures' and 'societies' (Wolf 1988, Barth 1992). This new emergent paradigm, 'New Critical Theory,' retains functionalism on its agenda, albeit in a literary and historical mode. It promises to further problematize addressed by anthropological historical realism (sometimes in the guise of political economy) and postinterpretive cultural critique. In both instances, functionalism, in anthropology, lingers on (just as evolutionism and diffusionism linger on) not just as situated knowledge ever in the making but as historical ethnography of the operations of culture.

See also: Anthropology, History of; Functionalism, History of; Functionalism in Sociology; Malinowski, Bronislaw (1884–1942); Parsons, Talcott (1902–79)

Bibliography

- Barth F 1992 Towards greater naturalism in conceptualizing societies. In: Kuper A (ed.) *Conceptualizing Society*. Routledge, London and New York
- Fox R G (ed.) 1991 *Recapturing Anthropology: Working in the Present*. School of American Research Press, Santa Fe, NM
- Gellner E 1987 The Zeno of Cracow, or revolution at Nemi, or the Polish revenge. In: *Culture, Identity and Politics*. Cambridge University Press, Cambridge, UK
- Malinowski B 1944 *A Scientific Theory of Culture, and Other Essays*. University of North Carolina Press, Chapel Hill, NC
- Manganaro M (ed.) 1990 *Modernist Anthropology: From Fieldwork to Text*. Princeton University Press, Princeton, NJ
- Mead M (ed.) 1937 *Cooperation and Competition among Primitive Peoples*. McGraw-Hill, New York
- Mintz S W (ed.) 1985 *History, Evolution, and the Concept of Culture: Selected Papers by Alexander Lesser*. Cambridge University Press, Cambridge, UK

- Radcliffe-Brown A R 1935 On the concept of function in social science. *American Anthropologist* 37: 394–402 (reprinted 1952 in *Structure and Function in Primitive Society*. Free Press, Glencoe, IL, pp. 178–87)
- Sahlins M D 1976 *Culture and Practical Reason*. University of Chicago Press, Chicago
- Sanjek R (ed.) 1990 *Fieldnotes: The Makings of Anthropology*. Cornell University Press, Ithaca, NY
- Stocking G W Jr. (ed.) 1991 *Colonial Situations: Essays on the Contextualization of Ethnographic Knowledge. History of Anthropology*. University of Wisconsin Press, Madison, WI, Vol. 7
- Stocking G W Jr. 1992 *The Ethnographer's Magic and Other Essays in the History of Anthropology*. University of Wisconsin Press, Madison, WI
- Stocking G W Jr. (ed.) 1996 *Volksgeist as Method and Ethic: Essays on Boasian Ethnography and the German Anthropological Tradition. History of Anthropology*, Vol. 8. University of Wisconsin Press, Madison, WI
- Thornton R 1985 Imagine yourself set down ...: Mach, Frazer, Conrad, Malinowski and the role of imagination in ethnography. *Anthropology Today* 1(5): 7–14
- Thornton R J, Skalik P (eds.) 1993 *The Early Writings of Bronislaw Malinowski*. Cambridge University Press, Cambridge, UK
- Urry J 1993 *Before Social Anthropology. Essays on the History of British Anthropology*. Harwood, Chur, Switzerland
- Vincent J 1986 Functionalism revisited: An unsettled science. *Reviews in Anthropology* 13: 331–9
- Vincent J 1991 Engaging historicism. In: Fox R G (ed.) *Recapturing Anthropology: Working in the Present*. School of American Research Press, Santa Fe, NM
- Wolf E 1988 Inventing society. *American Ethnologist* 15(4): 752–61

J. Vincent

Functionalism in Sociology

Functionalism is a sociological research program of historical and systematic significance. Sociological explanations in terms of functions are teleological: social phenomena are to be explained by their effects and consequences—chiefly of a beneficial, morally desired or non-intended character—and/or by a circuit of equilibration (a homeostatic loop) which is maintaining a certain state of proper functioning. The notion of (mono-)functionally specialized subsystems of society can be seen as the most important precipitate of functionalist thinking in sociology.

1. Basic Extra-disciplinary Orientations of Sociological Functionalism

Beginning with the renaissance, the quest for a truly empirical analysis in the natural sciences has given the concept of function a crucial leverage in replacing the metaphysical notion of substance (Cassirer 1980). Substance has been recast as the functional interdependence of (minimally two) related variables. In

sociology functionalist thinking has spread with the works of Auguste Comte, Herbert Spencer, and Émile Durkheim, who drew on contemporary physiology and evolutionary biology as the disciplines that offered the most sophisticated notions of functional interdependence so far.

Of crucial significance for the development of sociological functionalism is the work of Harvard University's biochemist and physiologist Lawrence Joseph Henderson, who played the role of a mediator between physiology and sociology as regards equilibrating processes. In his famous seminar 'Pareto and the Methods of Scientific Investigation' (1932–42) Henderson exerted considerable influence on several generations of Harvard sociologists, among them Robert K. Merton and Talcott Parsons, who became leading figures in the movement of functionalist thinking in the USA. The basic ideas of a refined sociological functionalism originate within this context.

1.1 System, Interdependence, Equilibrium

A mixture of ice, soda water, and whisky in a tightly stoppered thermos bottle can serve as an example for an isolated system of different components: water, alcohol, and carbon dioxide are its main components. There are three phases: a solid phase (ice), a liquid phase, and a gaseous phase. All three components exist in the liquid and the gaseous phase in this system. Concentration, pressure, and temperature form a set of related variables. If one thrusts the stopper more deeply into the neck of the bottle the pressure will increase and there will be a movement of components from the gaseous phase to the liquid phase; there will also be a change in concentration and temperature. All variables that characterize the system stand in a relation of mutual dependency with one another and make up an equilibrated state (Henderson 1935). If the system exhibits a minimal complexity, its internal process cannot be conceptualized as a temporal succession of single cause–effect chains any more. The alternative is the simultaneous variation of interdependent, but relatively autonomous, variables. The two single most important sources of Henderson's analysis of systems are Claude Bernard's experimental physiology—which deeply impressed Durkheim as well—and Josiah Willard Gibbs' research *On the Equilibrium of Heterogeneous Substances* (Gibbs 1948), which decisively extends the applicability of the laws of thermodynamics. On the basis of Gibbs' insights, the biochemical equilibrium of a *milieu intérieur*—of the circulation of blood, for example—can now be given a description in terms of functional equations (Russett 1966). This 'homeostatic process,' as it has been called by Henderson's colleague Walter B. Cannon, provides the systematic starting point for functionalist thinking in sociology. The conceptual

scheme of the social system involves the variation of heterogeneous, interdependent components that can be accounted for in a model of equilibrium—in the ideal case, in the form of a system of functional equations. The facts, that the variables of a social system are not easily quantifiable, and that the system cannot be experimentally isolated, do not detract from the value and applicability of this conceptual scheme.

1.2 The Problem of Teleology

Equilibrating processes are not random processes. They are not the result of deterministic causal mechanisms, and they do not issue from a deterministic process of natural selection, from a 'survival of the fittest.' The internal integration of an organism and its adaptation to external conditions are based on *reciprocal* relations with the environment: as the organism is adapted to the environment, the environment is adapted to the organism; there is something like the 'fitness of environment' which is based on the internalization of the environment's properties within the organism—water and carbon dioxide, for example, play a very important role for the organism in both milieus: the external and the internal (Henderson 1913).

At least *ex post factum* the world seems to be a world of interdependent functional systems and can be described as a process of 'negentropy'; a process of organization. This is the basic assumption on which the teleological character of functionalism is grounded. But commitment to this assumption is not a necessary condition to make use of equilibrium models in sociological functionalism.

2. Structural Functionalism

Structural functionalism marks the phase of dominance of functionalism in sociology, especially in the USA from the 1940s to the 1960s. Structural functionalism does not constitute a single and unified research program. There are simpler versions of functionalism than the refined functionalism already characterized above by its extradisciplinary grounding in equilibrium models. A paradigmatic version of the simpler forms of structural functionalism can be found in the work of British anthropologist A. R. Radcliffe-Brown (1952). The anthropological study of institutions in the context of a social totality can be seen as a blueprint for the procedure of functionalist research: the careful description of structural patterns, i.e., of non-random interrelations between components of a society or a social system that is at the core of this procedure. The *functioning* of these structural patterns secures the adaptation and/or integration, i.e., the persistence of the social unit in question. The structural

patterns are explained in terms of their consequences. Such a structural pattern would be, for example, the repressive law in archaic societies which Durkheim saw as indicative of a mechanical solidarity. Repressive law contributes to the integration of society. Retributive acts revive the feelings on which mechanical solidarity is based. All sociological research tends to exhibit the formal characteristics of functionalism in so far as it is pursued carefully (Davis 1959).

2.1 Simple Functional Explanation

A simple functional explanation follows the following sequence:

- (a) A system S is adapted to its environment/is integrated, if the corresponding functional prerequisite P is fulfilled;
- (b) If the functional prerequisite P is fulfilled, pattern A or its functional equivalents B, C ... exist;
- (c) Patterns B, C ... do not exist in the system;
- (d) System S is adapted to its environment/is integrated, because pattern A exists in the system. This type of explanation does not imply an equilibrium model or a system of variables in the sense of Gibbs/Henderson.

Two questions have to be asked here. First, how can one explain *historically* that pattern A contributes to the fulfillment of the functional prerequisite? The functional explanation has to be linked to a theory adequate to ground the claim that pattern A would not exist at all if it did not make a difference in the fulfillment of the functional prerequisite P or in the adaptation/integration of system S, respectively (Runciman 1990). Second, what is the *origin* of functional prerequisites? Either all consequences of action are functional prerequisites, or a limited set of functional prerequisites has to be justified theoretically and/or empirically. In the latter case a more refined version of functional explanation is called for, which will be taken up in the next step. In the first case an encompassing functional unity of society cannot be presupposed. Here, functionalism advances on an inductive basis. Such a middle-range approach to functionalism is presented by Robert K. Merton, who distinguishes between manifest functions on the basis of intended and recognized consequences of action on the one hand, and latent functions on the basis of non-intended and non-recognized consequences of action on the other (Merton 1963).

Since Merton, functionalism has become associated with a concept of (mono-)functional systems that are the product of an interweaving of non-intended consequences of action and have gained substantial autonomy as against the intentions of actors. The discrepancy between actors' intentions when acting in economic systems and the actual outcome of market processes as an aggregated result of these intentional actions may serve here as an illustration. As Merton demonstrates, the definition of functions on the basis

of consequences of action does not imply any preference towards an existing social order. It is impossible to make the disruption of social order the center of a functional perspective. Merton's repudiation of the commonplace critique of conservatism proves him to be one of the liberal representatives of functionalism.

2.2 Refined Functional Explanation

The greatest danger for functional explanation lies in the arbitrary selection of consequences of action as defining a function—this holds true for both simple and refined functional explanations. A limited set of functional prerequisites could be gained inductively from the empirical comparison of singular societies. The alternative is to justify a limited set of functional prerequisites on theoretical grounds. It has to be clarified, in which sense functions such as adaptation or integration make sense, not only for the organisms of biology and physiology, but for social systems as well.

In particular, the question has to be answered, how an essential change of the system can be distinguished from mere surface variations in its operations. What is missing is a clear analogon in sociological functionalism for the concept of the 'death of an organism' (Döbert 1973).

In his version of structural functionalism Talcott Parsons has provided a theoretical, normativistic account of functional prerequisites: a cultural system of values that is shared by all members of a society constitutes the central criterion for the adaptation and integration of society. The processes of society are functional in so far as they help to realize this common system of values. Parsons cannot introduce this system of values independently from an analysis of contemporary society that makes his theoretical grounding of functional problems a mere affirmation of consequences. At the same time he has to admit, that he is not able to conceive the equilibrating process of society in terms of a system of functional variables, in terms of a Gibbs/Henderson system. In particular, it does not seem possible to explain the *change* of action systems in terms of a functionalist analysis of equilibrating processes, especially processes of a moving or even a dynamic equilibrium.

That leaves Parsons with static, normative structural categories and little difference compared with the 'no frills' strategy of simple functional explanation. The concept of function serves for him as a means for the dynamic interpretation of static categories that can eventually explain no more than a static equilibrium—the boundary maintenance of society and social systems, as they already exist. Parsons sees the concept of function as a logical, but not really satisfactory, equivalent for a system of simultaneous functional equations; structural functionalism therefore is only the 'second best type of theory,' biased for a normatively defined static equilibrium (Parsons

1954, 1959). Since an existing system of values is used here as the grounding interpretation for a limited set of functional prerequisites, the charge of conservatism that has been brought against Parsons is difficult to repudiate. However, it should be clear that the reasons for these shortcomings of Parsons' theory are not political or ideological in character, but due to theoretical deficiencies. Finally, Parsons' normative version of structural functionalism has to be put into the historical context: it is a reaffirmation of the cultural identity of American society during the dramatic developments on the eve of World War II, when the identity of American society indeed seemed to be at risk (Parsons 1973).

The guiding ideal of a Gibbs/Henderson system certainly is an admonition for Parsons to avoid simple analogies and to search thoroughly for a system of interdependent functional equations. This clearly distinguishes Parsons' effort from simple forms of functional explanation, but it is not the only difference. Formally the essential difference is marked by Parsons' use of an equilibrium model (Nagel 1957, 1961). A state of equilibrium E is dependent on the variable contributions of several functions—a case of hierarchical or vertical organization of equilibrium: E is a system's overriding concern. The range of variation of these functional contributions is large enough to endanger the stability of E. If the stability of E is endangered, certain mechanisms intervene to start a process of re-equilibration. A state of equilibrium E can be brought about by several different combinations of the values of the functional variables, but there are also combinations that characterize a disequilibrium: the loss of stability. In contrast to the hierarchical, vertical organization stands the horizontal organization of equilibrium: in a horizontally organized equilibrium functional variables are on a par with each other and constitute a mutual equilibrium.

Of crucial importance for equilibrium models and for the development of functionalist thinking are the mechanisms that regulate the processes of (re)equilibration. There are three different mechanisms:

(a) *Compensatory mechanisms.* The interdependence of the functional variables is based on direct relations between them: a decreasing contribution of the functional variable A directly causes the compensating, increased contribution of the functional variable B. The limits of compensation are given in the functional variables' range of variation. Increasing the pressure in the thermos bottle illustrates a process of re-equilibration on the basis of compensatory relations between pressure, concentration, and temperature.

(b) *Mechanisms of a central control of equilibrium.* The cybernetic self-regulation of machines can provide an example for this kind of re-equilibrating mechanism. The thermostat of a refrigerator is an instance of a central re-equilibrating control.

(c) *Mechanisms of reflexive control.* In this case the

system must be able to observe its own operations and to decide on re-equilibration on the basis of a 'conscious' or self-referential act of taking these observed operations into account.

3. Systemic Functionalism

Parsons' systemic functionalism exemplifies an equilibrium model characterized by a central control. Systemic functionalism replaces the deficient structural functionalism and develops an equivalent to the Gibbs/Henderson system. This theory claims to be able to explain the change of action systems in terms of a model of a moving or even a dynamic equilibrium. Parsons makes temporal process a constituent of his functional variables. The standard reproach directed against this type of functionalism concerns its deductive character. It may be deductive, but the theory is nevertheless empirically grounded in the empirical studies of small groups and interaction processes by Robert F. Bales, who collaborated with Parsons in developing the so-called 'four-function scheme' (adaptation, goal attainment, integration, latent pattern maintenance: AGIL). These four basic functions of social systems are derived from the analyses of processes of problem solution in small groups (Bales 1950, Parsons et al. 1981).

Parsons conceives of society as a functionally differentiated unity consisting of four subsystems: economy, polity, societal community and the cultural, fiduciary subsystem. Parsons reconstructs the interdependence of these functional subsystems in terms of communication or interchange.

Each subsystem has a specialized language of communication at its disposal—generalized media of communication or interchange: money, power, influence and moral appeals (see Fig. 1). In this scheme of

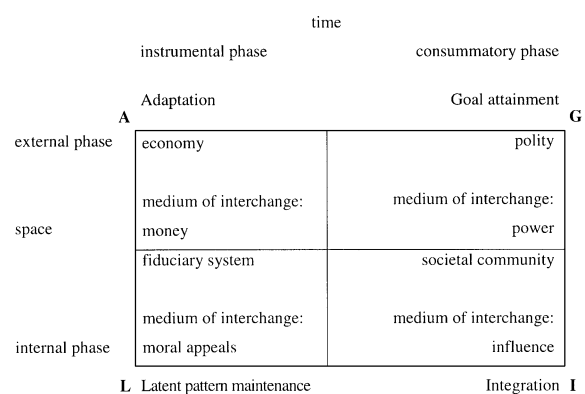


Figure 1

Parsons' four-function scheme, the functional subsystems of society, and the societal media of interchange

interchange relations between the subsystems in terms of these media, Parsons has constructed his final equivalent to a system of functional equations in the sense of Gibbs and Henderson. Moving and even dynamic equilibrium now can be explained in terms of this communicative interchange. An empirical application of this scheme can be found in the study of Parsons and Gerald Platt on the system of higher education in the USA (Parsons and Platt 1973).

All functional processes are regulated in the final instance by a *central control mechanism*: the top system of a cybernetic hierarchy which Parsons has integrated into his four-function scheme since the 1960s. This top system is culture or—on the level of societal subsystems—the fiduciary subsystem. This is a theoretical assumption that perpetuates the ‘cultural determinism’ of Parsonian theory.

4. Functional Structuralism

The theory of German sociologist Niklas Luhmann offers an outstanding example for the ‘reflexive control’ of the equilibrium process. Although Luhmann’s theory has been referred to as a ‘second encoding’ of Parsons’ functionalism, it is really an inversion of Parsons’ structural functionalism. Therefore it may be called functional structuralism: Luhmann is not—as is Parsons—primarily interested in the function of given structures. His inverted theoretical perspective is characterized by the search for equivalent solutions of problems, i.e., for equivalent structures that fulfill given functional prerequisites. The basic assumption of his theory is that for almost all problems there are also other solutions than the ones already selected; every solution is a contingent one. This perspective brings ever-new potentialities and the continuing change of systems into view. The basic functional problem is not the boundary maintenance of systems but the functioning of systems in the face of contingency. There is no culturally defined structure, no program, no code that has to be maintained, executed, or realized to preserve the identity of the system. Luhmann shifts Parsons’ problem of the boundary maintenance of systems radically into the dimension of time, shaking off all the normative connotations of Parsonian theory. This leaves only the continuity of the system over time as a problem—and as a radically purged understanding of persistence in an overly complex world of other possibilities, of other equivalent solutions (Luhmann 1970).

From the 1980s on, Luhmann began to rethink the distinction between system and environment, drawing on the works of Argentinian biologists F. J. Varela and Humberto Maturana and on the so-called ‘second cybernetics’ developed by researchers such as Heinrich von Foerster and Ernst von Glasersfeld. At the core of this process of rethinking stands the concept of the autopoiesis (self-creation) of systems. Even before this rearrangement of theoretical concepts, Luhmann de-

veloped the notion of a self-referential system. Self-referentiality took the strategically important place which the top system of the cybernetic hierarchy, culture, held in Parsons’ theory. This was a first step to introduce a reflexive control of equilibrium process. Even more significant is the concept of autopoietic systems. Autopoietic systems repeat the difference of system and environment internally, within the system. Autopoietic systems use this difference to observe (and to regulate, to condition) their own operations. The distinction between system and environment is transformed, the difference of identity (what belongs rightfully to the system makes up its identity), and difference (what does not belong to the system, what is foreign to it) forms the basis for the system’s operations.

Autopoietic systems produce the typical elements they consist of on their own by observing themselves in the process of production, by using the difference of identity and difference—in a metaphorical sense one could call this the self-consciousness of the system. Autopoietic systems are closed, and operate only under the condition of self-contact, whereas Parsons postulated open systems engaged in processes of mutual interchange. The only way for autopoietic systems to influence each other is in a manner that Luhmann calls the ‘structural coupling’ of systems: mutual irritations and resonance effects which the operations of one system have in the other. In a sense, autopoietic systems discover their own functionality and internalize the process of the search for equivalent solutions (Luhmann 1984). Luhmann’s theory can eschew Anthony Giddens’ criticism that equilibrium process, the self-regulation of systems, eludes an analysis in terms of *reflexivity* (Giddens 1977).

For Luhmann, social systems consist of communications as elements. The differentiation of society into (mono-)functionally specialized subsystems—the economic system, the political system, science, art, religion, love, etc.—is based on the development of generalized media of communication. These specialized types of communications form the elements which autopoietic systems have to reproduce endlessly. The economy, for example, is made up of money communications; payments.

It is observing itself in terms of the difference payment/non-payment (or liquid/broke). The maintenance of liquidity is the single most important functional prerequisite which has to be fulfilled to secure the continuity of the economic system’s operations (Luhmann 1988).

5. Neofunctionalism

Neofunctionalism is a research program that has been founded on the basis of a thorough critique of Parsonian theory, in particular by Jeffrey C. Alexander (Alexander 1985, 1998). Although the name suggests a

close relationship with other forms of functionalism, the central criterion for belonging to the class of functionalist approaches is clearly not satisfied: neofunctionalism does not provide simple or refined functional explanations in terms of the teleological interpretation of consequences of action or in terms of equilibrium models. The only reason that the label 'neofunctionalism' is justified at all can be seen in its renewal of Davis' assertion, that all careful sociological analysis is (neo)functional analysis. As a movement of theory and empirical research, neofunctionalism is a reorientation of sociological analysis from systems and equilibrium to concrete actors and interpretative processes.

See also: Control: Social; Differentiation: Social; Integration: Social; Labor, Division of; Luhmann, Niklas (1927–98); Parsons, Talcott (1902–79); System: Social; Values, Sociology of

Bibliography

- Alexander J C (ed.) 1985 *Neofunctionalism*. Sage Publications, Beverly Hills, CA
- Alexander J C 1998 *Neofunctionalism and after*. Blackwell, Oxford, UK
- Bales R F 1950 *Interaction Process Analysis: A Method for the Study of Small Groups*. Addison-Wesley, Cambridge, MA
- Cassirer E 1980 *Substanzbegriff und Funktionsbegriff. Untersuchungen über die Grundfragen der Erkenntniskritik*. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany
- Davis K 1959 The myth of functional analysis as a special method in sociology and anthropology. *American Sociological Review* **24**: 757–72
- Döbert R 1973 *Systemtheorie und die Entwicklung religiöser Deutungssysteme. Zur Logik des sozialwissenschaftlichen Funktionalismus*. Suhrkamp, Frankfurt, Main, Germany
- Gibbs J W 1948 *The Collected Works of J Willard Gibbs*. Yale University Press, New Haven, CT
- Giddens A 1977 *Studies in Social and Political Theory*. Hutchinson, London
- Henderson L J 1913 *The Fitness of Environment*. Macmillan, New York
- Henderson L J 1935 *Pareto's General Sociology. A Physiologist's Interpretation*. Harvard University Press, Cambridge, MA
- Luhmann N 1970 *Soziologische Aufklärung 1. Aufsätze zur Theorie der Gesellschaft*. Westdeutscher Verlag, Opladen, Germany
- Luhmann N 1984 *Soziale Systeme. Grundriß einer allgemeinen Theorie*. Suhrkamp, Frankfurt, Germany
- Luhmann N 1988 *Die Wirtschaft der Gesellschaft*. Suhrkamp, Frankfurt, Germany
- Merton R K 1963 *Social Theory and Social Structure*, rev. edn. The Free Press, Glencoe, IL
- Nagel E 1957 *Logic Without Metaphysics, and Other Essays in the Philosophy of Science*. The Free Press, Glencoe, IL
- Nagel E 1961 *The Structure of Science: Problems in the Logic of Scientific Explanation*. Harcourt, New York
- Parsons T 1954 *Essays in Sociological Theory*, rev. edn. The Free Press, New York
- Parsons T 1959 *The Social System*. The Free Press, Glencoe, IL

- Parsons T 1993 *Talcott Parsons on National Socialism*. Aldine De Gruyter, New York
- Parsons T, Bales R F, Shils E A 1981 *Working Papers in the Theory of Action*. Greenwood Press, Westport, CT
- Parsons T, Platt G M 1973 *The American University*, Harvard University Press, Cambridge, MA
- Radcliffe-Brown A R 1952 *Structure and Function in Primitive Society*. Cohen & West, London
- Runciman W G 1990 *A Treatise on Social Theory. Vol. I: The Methodology of Social Theory*. Cambridge University Press, Cambridge
- Russett C E 1966 *The Concept of Equilibrium in American Social Thought*. Yale University Press, New Haven, CT
- Sztompka P 1974 *System and Function. Toward a Theory of Society*. Academic Press, New York
- Wenzel H 1991 *Die Ordnung des Handelns. Talcott Parsons' Theorie des allgemeinen Handlungssystems*. Suhrkamp, Frankfurt am Main, Germany

H. Wenzel

Fundamental Rights and Constitutional Guarantees

'Fundamental rights' are typically the protections of individual (and sometimes group) liberties in modern constitutions. International human rights law provides a baseline of minimum protections that all members of the international community ought to observe. Many constitutional orders supplement these minimum protections with additional guarantees.

1. Types of Fundamental Rights

Fundamental rights developed in three stages. Initially governments committed themselves to protect a set of basic civil and political rights, and later began to treat social welfare rights as fundamental. In the later part of the twentieth century, rights to cultural and environmental protection came to be understood as fundamental as well.

The idea that people have constitutional rights makes sense only in the context of accounting for why some exercises of power are arbitrary and unjustified. Postmedieval Western political theory abandoned accounts in which sovereign power was by definition unlimited, and replaced them with accounts of sovereign power—usually the power of governments, but sometimes the power of individuals over others—according to which there were limits on the justified exercise of power. Initially, the accounts defined a set of rights, which came to be described as 'civil' rights, that people had simply by virtue of the fact that they lived in an organized society (rather than in a state of nature). At the core of this set were the right to own property, the right to dispose of one's property

according to one's personal preferences, and the right to enter into legally enforceable contracts. Protecting these core rights were collateral procedural rights, usually associated with judicial proceedings, to ensure that one could effectively protect property rights and enforce contracts.

Successful popular and elite struggles against arbitrary royal governments led to the creation of institutions to limit government (e.g., parliaments), and to the development of another set of rights. These rights, denominated 'political' rights, arose because people lived in particular societies, each with its own political form of regulating arbitrary power. Civil rights were universal, held by all people outside the state of nature; political rights were held by the people of particular political society, and might vary from one society to the next. As parliamentary, and more generally popular, government spread, the core political right came to be the right to vote. Some rights that became widespread in this first period checked arbitrary power by protecting the institutions of civil society; examples are the press and religious institutions (other than an established church).

Accompanying the development of civil and political rights was the expansion of their distribution among the population. Conceptually, civil rights inured to people simply because they were members of an organized society, and political rights did so because they were members of a particular political order. As the idea that people were entitled to civil and political rights took hold, the denial of those rights to certain groups—women, slaves, men with little or no property—became increasingly difficult to defend. Social movements pressed for the extension of civil and political rights to those groups. Slavery was abolished in western societies by the middle of the nineteenth century. Anglo-American law recognized the rights of married women to hold property and to enter into contracts in their own names through a series of statutes enacted during the same century. Most generally, the franchise was expanded gradually, beginning with the elimination of property-based qualifications for male voting in the United States in the 1830s and 1840s, extending to universal male suffrage from the 1860s through the early years of the twentieth century in the United States and western Europe, and culminating in the elimination of gender-based qualifications for voting from the 1920s through the late 1940s in those regions.

The idea of a generalized right to equal treatment became increasingly credible as civil and political rights were extended gradually. Equal protection of the law was implicit in the theory of civil and political rights, but the manifest denial of equality made it difficult to conceptualize equality as a fundamental right. As barriers to equal protection of the law crumbled, equality as such became a fundamental right. In the contemporary United States and in discussions elsewhere influenced by the US experience,

the term 'civil rights' now refers not to a category of rights, as it did earlier, but to the general right to equal treatment. Controversies over civil rights concern whether rules denying some benefit to specific groups violate their right to equal treatment, not whether access to the benefit is itself a right.

A second generation of fundamental rights, dealing with social welfare, began to emerge in the late-nineteenth century. The rise of socialist parties in western Europe provoked a response from conservatives, such as Germany's Chancellor Otto von Bismarck, and from the Roman Catholic Church. The socialist parties challenged the capitalist order for its failure to secure conditions of material well-being, such as minimal guarantees of shelter, food, and employment. According to the conservative perspective and Catholic social teaching, providing these second generation rights is a basic task of modern governments. The western commitment to social welfare rights deepened during World War II; in the War's immediate aftermath, socialist parties again played a large role in the constitutional restructuring of western Europe. The commitment was extended geographically as well, as former colonies gained independence and incorporated second generation social welfare rights in their constitutions.

The socialist arguments for social welfare rights rested to some extent on the fact that first generation civil and political rights were already viewed as fundamental. They were purely formal protections, because they guaranteed a right to hold property without guaranteeing that a person would actually hold property. In more modern terms, they protected opportunities without ensuring outcomes. Critics argued that first generation rights were empty guarantees. As they saw it, the right to own property was meaningful only to people who actually owned it. More generally, people could really have and exercise certain first generation rights only if they had second generation rights as well.

Social welfare rights were connected to first generation rights in another way. People who lacked guarantees of material well-being were likely to be dependent on others for their lives; they might, therefore, be under the command of those on whom they depended. They might be so consumed by their efforts to reach minimal levels of material well-being that they could not exercise rights to vote or participate in politics. Proponents of second generation rights argued that the fair exercise of formally guaranteed rights could occur only when people had social welfare rights as well.

First generation civil and political rights and second generation social welfare rights generally had a fundamentally individualistic character: They were rights that each person held simply because of his or her status as a human being or as an individual member of a particular society. A third generation of rights became important in the late 1960s and thereafter,

although there were hints of this new category of rights earlier. This group includes rights of national minorities to sustain their own cultures and languages within states dominated by other national groups and, relatedly, rights of aboriginal peoples to obtain support for their cultures and some degree of reparation for their displacement by settler and immigrant communities (see *Aboriginal Rights*). Rights to the protection of the environment may also be included in enumerations of third generation rights.

Although most first generation rights were individual, some had a collective orientation. Rights to religious liberty are individual only under a particular theological view of the relation between individuals and their god(s); they are collective rights when held by a group organized, for example, as a church. Third generation rights expanded the range of rights that were importantly collective in nature. A person can have a right to maintain the language of his or her origin only if, and to the extent that, other people similarly have and exercise that right. Environmental protection benefits society as a unit, rather than any particular member of society.

2. Permissible Limitations of Fundamental Rights

No legal system gives absolute protection to every constitutional guarantee. Doing so would sometimes substantially impair a government's ability to pursue social goals that most people regard as worthwhile. Modern formulations of the power of governments to restrict fundamental rights often resemble the one used in Canada's Charter of Rights and Freedoms (1982), which authorizes restrictions on the rights it protects when such restrictions are 'reasonable limits ... [that] can be demonstrably justified in a free and democratic society.'

Historically, governments asserted the power to limit constitutional guarantees to advance the public interest in such matters as health, safety, and (more controversially) morality. These were called the 'police power' limits on constitutional guarantees. Frequently, governments found it relatively easy to assert reasonable police power interests to support restrictions on fundamental rights. This undermined the status of those rights as truly fundamental.

Constitutional systems, often through their constitutional courts, now restrict the occasions on which ordinary police power interests can justify restricting fundamental rights. In societies characterized by a wide range of views about what constitutes morality, it is now difficult to justify restrictions by reference to morality: the claim that some restriction advances a specific moral goal is met with the response that, in a morally pluralistic society, public policy ought not seek to advance that goal. Most systems require that restrictions on fundamental rights be 'proportional' to the harms the restrictions are designed to reduce.

Many constitutional systems require that when efforts to promote uncontroversial goals such as health and safety impair fundamental rights, those efforts must be justified by showing, for example, that they infringe the fundamental rights no more than is absolutely necessary to advance the police power goal, and that they will be effective in doing so. Terms such as 'strict scrutiny' or 'substantial showing' are common in describing the limits on a government's ability to invoke its police powers to restrict a fundamental right.

Sometimes, constitutional guarantees conflict with each other. A defendant's right to a fair trial may conflict with the press's free expression right to publish information about a crime and with the public's right to know about the conduct of public proceedings such as trials. A person's rights of free expression, when they take the form of hate speech, may conflict with the ability of subordinated groups to achieve true equality.

Constitutional systems address these clashes in two ways. They may give one right substantial priority over others, allowing infringements on the preferred right only when such infringements appear to be the only way to protect the other rights. That is the solution in US constitutional law to the conflict between free speech and fair trial rights: the US Supreme Court allows restrictions on dissemination of information about criminal proceedings only when no other methods of protecting the defendant's right to a fair trial seem likely to succeed. However, the court must provide reasons for choosing the right that has the preferred status, and must explain the fact that such choices may produce intrusions on the subordinated rights that the alternative approach could mitigate.

The other solution is to balance the conflicting constitutional guarantees. Under this approach, a statute prohibiting some forms of hate speech might be acceptable if decision-makers reasonably conclude that the restrictions it imposes are outweighed by the increase in equality that it will produce. While common among constitutional orders, balancing solutions are problematic because they call upon decision-makers to compare interests that are not truly commensurable.

The emergence of third generation rights poses particular difficulties in the theory of rights, because those rights sometimes conflict quite dramatically with first and second generation rights. Some national minorities, for example, may have internal norms that subordinate women in ways incompatible with the exercise by their female members of the full set of first and second-generation rights. A right to individual mobility may undermine the ability of a linguistic minority to sustain its language, as individual members of the minority move away, or simply choose to conduct their affairs in the majority's language. Constitutional orders have found it difficult to address

these sorts of conflicts. The balancing approach seems even more inapt here than it does when there are conflicts between political rights, because the rights at stake so obviously are different. Further, the evolutionary development of rights means that first and second-generation rights have a stronger theoretical basis than third-generation rights, and commitments to first and second-generation rights usually are held more deeply, if only because more familiar. When third-generation rights conflict with others, then, the third-generation rights are likely to yield, making them legally and practically less fundamental and secure than other rights.

3. *Methods of Protecting Fundamental Rights*

Some fundamental rights are identified in written documents, such as constitutions, statutes, or international human rights treaties. Particularly in Anglo-American legal systems, others derive from the unwritten law. It is conventional to regard judicial enforcement as the necessary concomitant of fundamental rights. The means of enforcing fundamental rights are more varied, however. Notably, the institutional mechanisms for enforcing international human rights consist of quite weak judicial institutions and significantly more potent moral and political ones.

Most of the world's nations have signed treaties committing themselves to enforcing international human rights defined as minimum standards universally applicable. Some, but not all, of these treaties also require the nations to submit to adjudication by international human rights courts. Not all states have submitted themselves to these courts, however. An extremely limited number of human rights violations, such as genocide, can be enforced against wrong-doers in *any* court in the world willing to exercise its power, and the 1990s saw the creation of international criminal courts with authority to convict upon finding that a defendant had committed criminal violations of core human rights. Otherwise, the power of international tribunals to enforce international human rights is rather limited.

Fundamental rights can be effectively guaranteed even without judicial enforcement. International human rights, for example, are enforced in part by moral pressure generated by adverse publicity produced by some organs of the international community and by nongovernmental human rights organizations (NGOs), an increasingly potent participant in the international arena. The mere presence in a constitution of a fundamental guarantee may place political pressure on governments to honor the guarantee. Legislatures and chief executives can create committees or administrative units charged with reviewing legislative and administrative proposals to determine whether they are consistent with constitutional rights.

The question of enforcement is particularly pressing in connection with second generation rights. Fully guaranteeing social welfare rights would be extremely expensive in many countries, often far beyond the nation's fiscal capacity. Some constitutions address this problem by describing the guarantees of second-generation rights as 'directive principles of social policy,' a phrase first used in the Irish Constitution of 1937 and then adopted by the postwar Constitution of India. As constitutional 'directives,' second-generation rights are committed to the care of the legislature and sometimes, as in the Irish Constitution, specifically excluded from judicial enforcement. Other constitutional systems, such as the Italian and Japanese, treat second-generation rights as judicially enforceable, but allow legislatures a wide range of discretion in choosing how to go about actually providing those rights.

Modern constitutional systems typically provide some mechanisms for the judicial enforcement of first generation rights. Courts may develop the unwritten (or 'common') law or interpret statutes in ways expressing sensitivity to concerns about fundamental rights, even if they acknowledge that the legislature has the power to alter the law however it chooses (as is the case, at least in theory, in many parliamentary systems). They may find that when executive officials violate fundamental rights, they have acted outside the authority conferred on them by statute, a doctrine known as *ultra vires*.

Sometimes, the mechanisms for judicial enforcement of constitutional rights are so limited in scope or so cumbersome that individuals may be unable to raise claims that their personal constitutional rights have been violated in specific cases. In France, for example, individuals cannot challenge statutes once they have been promulgated; only legislators and a few other high officials have the power to invoke constitutional review, and then only before a statute has become binding law.

Proponents of judicial enforcement of fundamental rights often overlook a serious problem with judicial enforcement. Even when all agree that there is a fundamental right, described in general terms, there may be reasonable disagreement over the precise content of that right and, in particular, over whether a specific statute in fact violates the right. Some legal doctrines take account of the existence of reasonable disagreement of this sort. The European Court of Human Rights, for example, gives states a 'margin of appreciation,' allowing them to adopt interpretations of the agreed-upon rights that other states reject and that the judges themselves might reject if they had the power to specify the content of the right themselves.

Judicial enforcement means that the views of judges concerning the precise content of an agreed-upon right prevail over the views of legislators or of executive enforcement officials that no violation of fundamental rights has occurred. Where disagreement regarding

the precise content of a fundamental right is reasonable, the case for judicial enforcement is weaker than its proponents believe.

See also: Aboriginal Rights; Bill of Rights; Civil Liberties and Human Rights; Civil Rights; Civil Rights Movement, The; Constitutional Courts; Discrimination; Equality and Inequality: Legal Aspects; Human Rights, History of; Human Rights: Political Aspects; Rights: Legal Aspects

Bibliography

- Alston P (ed.) 1999 *Promoting Human Rights through Bills of Rights: Comparative Perspectives*. Clarendon Press, New York
- Henkin L, Neuman G L, Orentlicher D F, Leebron D W 1999 *Human Rights*. Foundation Press, New York
- Jackson V C, Tushnet M 1999 *Comparative Constitutional Law*. Foundation Press, New York
- Kull A 1992 *The Color-blind Constitution*. Harvard University Press, Cambridge, MA
- Tushnet M 1999 *Taking the Constitution Away From the Courts*. Princeton University Press, Princeton, NJ

M. Tushnet

Fundamentalism (New Christian Right)

'New Christian Right' (NCR) is the label given to a cluster of US social movement organizations and political pressure groups formed in the late 1970s and early 1980s, of which the best known was the Moral Majority, headed by Baptist fundamentalist pastor and televangelist Jerry Falwell, of Lynchburg, VA.

Unlike such ethnic and nationalist movements as right-wing Afrikanerdom and Ulster Unionism, where religion had long played a part of identifying a people and legitimating their political struggles, the political involvement of American fundamentalism has always focused on sociomoral issues and that engagement has been sporadic. Fundamentalists have alternated between periods of pietistic retreat from the world and attempts to re-assert their values on what was seen as an increasingly secular and immoral public sphere. In the 1920s and 1930s fundamentalists campaigned for temperance and against Darwinian evolution. In the 1950s they were involved in anticommunist campaigns. They were relatively quiet during the 1960s and early 1970s.

The mobilization of the NCR in 1979 was initiated by a group of right-wingers who had previously supported third-party candidates against the liberal Republican establishment. It had two novel features. It made extensive use of direct mailing (using the lists

of those who had contributed to fundamentalist radio and television programs) to raise funds, to spread information, and to create mass pressure from very small acts of commitment. For example, potentially supporters in a state with a liberal senator would be sent a pre-printed and pre-addressed card protesting about some piece of legislation and threatening to vote against the senator should he or she not vote the right way. All the recipient had to do was sign the card and mail it. The NCR also, for the first time, used white fundamentalist churches in a way that the Civil Rights movement has used black churches: to provide a network of contacts and venues for politicization. Where the law permitted a voter-registration booth would be set up in a fundamentalist church and the pastor asked to preach on the Christian's duty to vote.

Although Falwell and other NCR leaders were avowedly nonpartisan, they quickly won more friends among Republicans than Democrats and during the 1980s, the movement evolved into a pressure group within the GOP. In 1987, Marion G. 'Pat' Robertson, a religious broadcaster, Pentecostal minister, and founder of the Christian Broadcasting Network (later the Family Channel) ran for the Republican nomination for the presidency.

Specifically the NCR was opposed to an Equal Rights for Women Amendment to the constitution, gay rights, abortion, divorce, public displays of sexuality and promiscuity, and to the teaching of evolution and 'secular humanism' in schools. It was in favor of public prayer in schools and the teaching of Creationism (which it tried to recast as 'Creation Science'). More generally, it wanted public re-affirmation of the superiority of conservative Protestant culture and identity. Although this was often denied there was also a large amount of racism in the appeal of the NCR. Criticism of liberal welfare state and immigrant policies provided a respectable language for a 'nativist' reaction to Blacks and Hispanics. Support for tax tuition credits for independent schools, although presented as pro-parental choice, was also a way of opposing racial integration.

The NCR had considerable early success in gaining publicity for its cause. Falwell, Robertson, and other religious broadcasters were masters of the mass media. With the technical assistance and experience of right-wing lobbyists, they were able to create an appearance of power far beyond the reality. In this they were aided initially by the credulity of naive commentators who over-reacted to the reappearance of a large constituency which the liberal and cosmopolitan centers of the US had forgotten. Thus, the NCR's claims to have been responsible for the ousting of four leading liberal Democrat senators in 1980 was reported without the qualification that the swing against those four was the same as the swing against the Democrats in states where the NCR had not campaigned actively. The NCR also claimed the credit for the 1980 election of Ronald Reagan to the presidency, even though

Reagan's success owed far more to the appeal of his economic liberalism and his aggressive foreign policy than to his conservative rhetoric on socio-moral issues.

1. The NCR's Record

Despite the 12 years of Reagan and then George Bush in the White House, the NCR conspicuously failed to make much progress on its agenda. In areas where fundamentalists were strong, it won a number of court cases and managed to have state legislatures pass such measures as bills requiring that teaching of evolution be balanced by equal time for creationism, but such gains were always reversed once the issue had been appealed upwards. Appeal courts decided that Creation Science was religiously-inspired rather than evidence-driven science and hence that its teaching in public schools was unconstitutional. Attempts to have any form of instruction that did not specifically promote Christianity regarded as expressions of the 'religion' of secular humanism were rejected repeatedly by the courts. A landmark judgment ruled that teaching children about beliefs and values they did not share was not in itself a form of unconstitutional propaganda. On its most popular issue, opposition to abortion, the NCR made very little headway. Although the Supreme Court was willing to permit very slight constraints (for example, on public funding of information about abortion), it turned down opportunities to outlaw it repeatedly.

The NCR's electoral record was similarly unimpressive in the light of the unprecedented sums that were spent. Being able to deploy groups of highly committed activists allowed NCR groups to win power and promote candidates in some branches of the GOP but such advances were usually reversed by mainstream Republicans. When they were not, movement candidates usually were rejected by the electorate. In particular, the NCR alienated middle-class, women voters who saw the entire conservative agenda as an attempt to reverse their gains in the labor market and in public life.

Where the NCR had some success was in presenting itself as a cultural minority the rights of which had been infringed. It was thus able to reverse a number of overly zealous decisions of the separation of church and state. For example, a Christian student group won back funding that had been denied it while the University funded a variety of equally contentious political and cultural groups.

2. Explaining the NCR's Failure

Idiosyncratic causes of the NCR's failure can be identified. It certainly did Robertson's campaign no good that it coincided with two major television evangelists being exposed as frauds, perverts, and

hypocrites. However the major roots of failure lie in the dynamics of motivation and in US social structure.

Despite the title of Falwell's organization, NCR supporters were never a majority. They could only achieve their goals if they formed alliances with other conservative groups but this created motivational problems because it required alternating between two sets of values. Fundamentalists were encouraged to become politically active by various threats and insults to their socio-moral values, which were derived from their religious beliefs. But to be effective politically, they had to set aside their religious preferences and co-operate with conservative Catholics, Jews, and atheists. While the NCR leadership could separate religion and politics, many of the rank-and-file could not. They either became disillusioned at the compromises involved in pressure group politics or refused to make such compromises and by their anti-Catholicism, anti-Semitism, and racism offended, respectively, Catholics, Jews, Blacks, and Hispanics, groups which had considerable sympathy for aspects of the NCR agenda.

Potential for frustration was inadvertently built into the movement from the start by its flawed explanation of what was wrong with America. Because it saw all that offended it as the deliberate work of militant 'secular humanists,' it supposed that all that was required to 'bring America back to God' was an equal counterforce of militant socio-moral conservatives. The NCR failed to appreciate that many of the changes that concerned it were the accidental by-product of two social forces that, in general, its members supported: prosperity and individual autonomy.

The one sense in which liberals were implicated directly in the failure of the NCR was in successful counter-attack. Once leading scientists realized that they could not take their cultural dominance for granted but had to persuade people, they did so extremely effectively. Organizations such as People for the American Way mobilized very effectively to produce a compelling defence of diversity and toleration. But the liberal counterattack worked because it rested on values that many fundamentalists themselves accepted. Unlike, for example, Islamic fundamentalists, NCR sympathisers were not opposed thoroughly to the division of the public and private spheres or to the principle of toleration. Hence, they were often ambivalent about their own legislative program; while polls repeatedly showed most Americans opposed to abortion, more sensitive questioning showed considerable willingness to make exceptions and some reluctance to impose morality on others. In addition, most voters put socio-moral concerns relatively low down their list of priorities and attempts by NCR-supported candidates to stress such matters reinforced the impression that they were single-issue candidates who would not be effective politicians. The result was that the NCR failed to turn the very large constituency of born-again Christians into committed supporters.

What was clear from opinion polls and voting patterns in the southern states in Pat Robertson's election campaign is that even many who shared Robertson's religion were unenthusiastic about supporting a conservative and fundamentalist pastor for the presidency.

While it is easy to assess the NCR's record on specific agenda items and conclude that it failed, it is less easy to assess its more general and long-term effects. Given that the proportion of women in full-time employment, divorce rates, and abortion rates have continued to rise, we may conclude that the movement failed in its more general objectives. The one sense it can claim to have succeeded is in making sociomoral issues extremely contentious and emotive. Although the commonly used phrase 'culture wars' is an exaggeration, the NCR has contributed to the development of an extremely acrimonious tone in American public life signaled, for example, by murderous attacks on the staff of abortion clinics.

3. Conclusion

Although interesting in its own right, the career of the NCR allows us to understand some basic features of modern democracies. While the more ethnically and culturally homogenous and smaller states of Europe tend to have centralized and closed systems of political power and public administration, the size and diversity of the US has resulted in an open, federal structure that allows sizable minorities considerable opportunity to initiate debate and promote their agendas, and to win some local victories. However, those issues of central importance to the operation of the state are

dealt with in a centripetal structure that prevents even well organized and highly motivated minority movements carrying the day. In electoral politics and in the formulating of legislative programs, the NCR could press its case but then suffered repeated rejection at higher levels as that case was scrutinized by politicians answerable to wider constituencies. Similarly victories in lower level courts were reversed in various levels of appeal.

What is clear from the decisions of such influential elite groups as Congressional politicians and judges, and from the voting patterns of ordinary Americans, is that there is widespread implicit understanding that the protection of individual liberty and the preservation of social harmony in a culturally diverse democracy require a division of the public from the private and that, while considerable liberty can be permitted in the private sphere, the public sphere has to remain religiously neutral.

See also: American Studies: Politics; American Studies: Religion; Christian Parties: European; Creationism, Evolutionism, and Antievolutionism; Parties/Movements: Extreme Right; Protestantism and Gender; Religious Fundamentalism: Cultural Concerns; Republican Party; Televangelism; Tolerance

Bibliography

- Bruce S 1998 *Conservative Protestant Politics*. Oxford University Press, New York
Moen M C 1992 *The Transformation of the Christian Right*. University of Alabama Press, Tuscaloosa, AL
Wilcox C 1992 *God's Warriors*. Johns Hopkins University Press, Baltimore, MD

S. Bruce

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

G

Gaia

In the early 1970s, the English chemist, James Lovelock, and the American microbiologist Lynn Margulis postulated the Gaia hypothesis that life on Earth is part of a self-regulating system that binds the atmosphere to the continents and oceans in a quasi-physiological process (Lovelock and Margulis 1975). The ideas behind Gaia were developed as part of NASA's efforts to find life on other planets (Lovelock 1965). As a theory of evolution, Gaia holds that life forms have co-evolved with their environment in such a way that environmental conditions are held steady in the face of a continual influx of energy and matter. Gaia posits first that life on Earth is a planetary phenomenon, and second, that living organisms have to regulate their planetary environment otherwise the forces of physical and chemical evolution would render it uninhabitable (Lovelock 1979).

The evidence put forward for Gaia draws attention to the fact that the chemical composition of the Earth's atmosphere has stayed remarkably constant and quite unlike the composition that would be expected in the abiological steady state of a lifeless planet. The geological record suggests that neither global freezing nor boiling conditions have ever prevailed. Mean surface temperature has probably never departed from the range 5–50°C (Lovelock and Whitfield 1982). The Gaian perspective has spawned new avenues of scientific inquiry aimed at furthering the understanding of biogeochemical cycles, especially in the atmospheric sciences. For example, the importance of methane in climate change came to be taken seriously with the recognition that biological organisms play a vital role in regulating the atmospheric environment.

Lovelock (1988) coined the term geophysics to denote an emerging systems approach to earth science that might assist in furthering understanding of the implications of human intervention into biogeochemical cycles in a way that distinguishes situations which threaten the planetary operating system from those that do not. By using the term geophysics, Lovelock distanced himself from teleological criticisms and the mysticism that has come to surround the notion of Gaia. For Lovelock, Gaia is a quasi-living entity that has the capacity for global homeostasis operating according to cybernetic principles, which like a thermostat does not require will or sentience.

Gaia caught the public imagination in the 1980s, along with an emerging conviction in the scientific

community that the global environment was being seriously affected by human action (Joseph 1990). For environmentalists, Gaia became a symbol of the need for promoting an ecocentric view of the world around us, in contrast to the dominant anthropocentric view. For many, Gaia became also an argument that cooperation rather than competitiveness was needed to unite humanity in the face of threats to the global environment and an argument for alternative ways of living.

It is important to distinguish between the Gaia hypothesis of earth science, and Gaia, the popular symbol of environmentalists, which underpins a call for an ecological egalitarianism to end all forms of human domination of nature (Serafin 1988). A considerable literature has developed exploring Gaia both as a scientific concept and as an ethical, philosophical, and sociological proposition. Debates in the literature about Gaia have been set increasingly in the historical context of an emerging science of the biosphere. Gaia is seen as an elaboration of ideas put forward by the Englishman James Hutton who first proposed the Earth to be a living system in 1785 and the Russian Vladimir Vernadsky who realized the importance of life in transforming geology in the early part of the twentieth century.

See also: Cassandra/Cornucopian Debate; Ecology, Political; Environment and Development; Environmental and Resource Management; Environmental Economics; Nature, Concepts of: Environmental and Ecological; Sustainability Transition: Human–Environment Relationship

Bibliography

- Bunyard P (ed.) 1996 *Gaia in Action: Science of the Living Earth*. Floris Books, Edinburgh, UK
- Joseph L E 1990 *Gaia, the Growth of an Idea*. St. Martin's, New York
- Lovelock J E 1965 A physical basis for life detection experiments. *Nature* **207**: 568
- Lovelock J E 1979 *Gaia: A New look at Life on Earth*. Oxford University Press, Oxford, UK
- Lovelock J E 1988 *The Ages of Gaia: A Biography of Our Living Earth*. Norton, New York
- Lovelock J E, Margulis L 1975 The atmosphere as circulatory system of the biosphere—the Gaia hypothesis. *CoEvolution Quarterly* **6**: 30–41
- Lovelock J E, Whitfield M 1982 Life span of the biosphere. *Nature* **296**: 561

- Schneider S H, Boston P (eds.) 1992 *Scientists on Gaia*. MIT Press, Cambridge, MA
- Serafin R 1988 Noosphere, Gaia and the science of the biosphere. *Environmental Ethics* 10: 121

R. Serafin

Galton, Sir Francis (1822–1911)

Francis Galton was born in Birmingham on February 16, 1822, the third son of Samuel Tertius Galton, a wealthy banker whose fortune was mainly derived from his father's musket manufactories. Francis was also the grandson of Erasmus Darwin, in whose scientific and evolutionist footsteps he was later to follow. However, he was originally destined for a career in medicine. Aged 16, at his father's behest, Francis enrolled as a medical student at the Birmingham Free Hospital, from where he moved to King's College, London to study medicine in 1839, and to Trinity College, Cambridge in 1840 to read mathematics. Having achieved only a pass degree at Cambridge, largely due to the first of the mental breakdowns that punctuated his career, Galton continued with his medical instruction despite a growing aversion to it. Coming into a substantial fortune on the death of his father in 1844, he immediately abandoned medicine and spent several indolent years 'hunting, fishing and shooting.' Then, in 1850, he led and financed an exploratory expedition to South-West Africa (modern Namibia). On his return, in 1852, Galton was lionized by the scientific establishment. And having achieved this entrée into the scientific community, gradually he established himself as one of Victorian Britain's most respected men of science. He was knighted in 1909.

Francis Galton is now remembered chiefly for his innovations in the field of statistics and for advocating the practice of eugenics. But his interest in both of these areas initially arose from an attempt to provide scientific evidence for the inheritance of human mental attributes. In fact, Galton devoted most of his professional energies to investigating varied aspects of heredity, a predilection largely inspired by reading Charles Darwin's (his half-cousin) *The Origin of Species* in 1859. Throughout his career, Galton's views of mind and morality were unequivocally materialist and evolutionary. In 1865, he published an article in two parts, under the heading 'Hereditary talent and character,' that were later expanded into his first book on the subject of mental heredity, *Hereditary Genius* (1869). In each of these publications he sought to show that the pedigrees of geniuses contain considerably more forebears of high social distinction than the lineages of those with mediocre parentage. Galton was aware that any positive results might instead suggest

the importance of nepotism and unequal access to education. But on the basis of several ingenious—albeit less than convincing—arguments he insisted that for the highly intelligent neither education nor preferment are necessary for the fulfillment of their inborn intellectual potential. This permitted him to interpret any instance of relatedness among the eminent as being due to shared heredity.

Having presented detailed pedigrees for several hundred eminent men Galton calculated that 50 percent of illustrious individuals have at least one distinguished relation. On this basis he concluded that heredity is the primary determinant of intellectual ability. He also devoted a section of *Hereditary Genius* to estimating, in terms of the overall stock of hereditary intelligence, the 'comparative worth' of different nations and races. Galton reproduced the typical mid-Victorian conception of a racial hierarchy stretching from the Caucasian to the Negroid, although he argued that the ancient Athenians were appreciably more intelligent on average than any modern civilized nation.

During the 1870s, Galton continued his researches into human mental heredity. In 1874 he issued questionnaires to leading men of mid-Victorian science, hoping to shed light on the relative importance of 'nature and nurture' in their choice of vocation. Predictably, his data interpretation was prejudiced in favor of a hereditarian explanation. In 1875, although he had no conception of the modern distinction between monozygotic and dizygotic twins, Galton turned to the phenomenon of twinning as a means of delineating between heredity and environment. He collected the life histories of pairs of like-sexed twins and sought to determine if those that were similar at birth and in childhood became unlike upon leaving their similar home environments. He also asked if like-sexed twins that were dissimilar at birth subsequently became more similar in the shared home environment. In both cases Galton concluded that character had been formed prior to birth and had thereafter shown only limited flexibility in response to environmental pressures.

On the basis of his studies of heredity, Galton proposed the peremptory regulation of human breeding so as to maximize the biological quality of the nation. In 1869 he recommended the confinement of the socially undesirable in 'celibate monasteries or sisterhoods,' and the creation of incentives for the intellectual elites to reproduce more frequently. The term 'eugenics' was coined in 1883 by Galton from the Greek *eugenes* meaning 'good in stock.' Yet, eugenics was widely considered impracticable during the 1860s and it was only during Galton's last decade that he began championing the idea. By then, foreign challenges to Britain's industrial and mercantile hegemony, and fears of an increasingly volatile urban underclass had rendered eugenics an appealing notion to many *fin-de-siècle* Britons. In 1908, after initial

reluctance occasioned by his advanced age, Galton became President of the newly formed Eugenics Education Society.

Political considerations bulked large in Galton's advocacy of eugenics. Running through his work on heredity and eugenics, and becoming more pronounced with time, was a strongly elitist conviction that the vast majority of the lower classes were sexually profligate and irremediably inferior to the higher social classes. Accordingly, he expressed the fear that the lower classes were out-breeding the 'respectable' classes. Moreover, Galton's predilection for studying intelligence and his commitment to enhancing the reproductive rate of the most intelligent are symptomatic of his upbringing as a member of the Victorian professional middle class. This was a social stratum that set great store by intellectual accomplishment and produced many of the century's foremost social reformers, commentators and scientists: Galton unconsciously celebrated the core values and accomplishments of this important social group. In advocating eugenics he was also influenced by the widely held belief in the possibility of human perfectibility—of which his grandfather Erasmus Darwin had been a leading exponent—and by the potent mid-Victorian ideologies of human and social progress advanced most forcefully by his friend and colleague Herbert Spencer. Lastly, Galton's eugenic thought reflects the growing impetus among Victorian life-scientists for making their discipline practically useful to the state.

In the context of studying the inheritance of human mental and bodily traits, Galton emerged as an innovative statistician. His maxim was 'Whenever you can, count,' and his researches suggest a man with an almost obsessive compulsion to do so. In *Hereditary Talent and Hereditary Genius*, Galton became the first person to employ the Gaussian error curve in classifying human *mental* attributes. Although his use of statistics was at this stage rudimentary, he accomplished the important step of measuring—and focusing almost entirely upon—statistical variation. This aspect of the normal curve had not received much attention from his contemporaries (with the exception of the Belgian astronomer and statistician Adolphe Quetelet), many of his predecessors regarding variation from the mean as 'error' rather than a natural phenomenon worthy of investigation. All of Galton's major statistical insights stemmed from this concern for statistical spread as opposed to averages alone. In particular, his work was pivotal in the development of both the correlation co-efficient and regression analysis. Galton formulated a graphical method to measure correlation in 1886 by plotting the distributions of both variables, and two years later he devised an 'index of co-relation.' In fact, his computational procedures for 'co-relation' involved measuring the slope of the regression line and producing a measure of the regression coefficient. The mathematical underpinnings of correlation were developed more fully by

his admirer and first biographer, Karl Pearson. From Galton's idea of regression, Pearson also developed a statistical system for the linear production of two continuous variables.

Galton's role in the development of regression analysis is more ambiguous. As a means of studying the mathematics of hereditary variation, during the late 1870s Galton began to breed sweet peas. He found that the F_1 generation peas were, on average, only a third as extreme in any given quality as the parental generation. Thus, through his breeding experiments, Galton had stumbled upon the statistical principle of regression. However, he always assumed that the phenomenon of regression related to a fundamental physiological process, and his explanation for it was biological rather than statistical. Galton asserted that each individual is composed of distinct hereditary units derived both from parents and from earlier ancestors. In each generation a random selection of the ancestral and parental hereditary material is made, such that only some of the total complement of hereditary units are actually expressed in each offspring. Galton believed that, from a purely probabilistic point of view, the entire hereditary material—stretching back several generations—would on average be close to the population mean. However, because only a random selection of hereditary units gain expression, it is possible for more extreme characteristics to be produced. Regression occurs when, as is likely, the more mediocre hereditary units regain supremacy in the following generation. This conception of the nature of heredity led to Galton's formulation of the 'ancestral law of heredity,' a law stipulating that any two parents together contribute approximately one-half of the total hereditary material of the offspring, the four grandparents one-quarter, and so on until the proportions become vanishingly small. Galton believed this law to describe the actual hereditary composition of each individual and the model formed the basis of his statistical treatment of heredity in *Natural Inheritance* (Galton 1889). During the twentieth century, 'Galton's law' would be eclipsed by the advent of Mendelian genetics.

'Galton's law' may not have long survived the rediscovery of Mendel's work, but Galton has been credited with anticipating August Weismann's concept of the continuity of germplasm. Unlike Darwin and the majority of his contemporary life-scientists, from as early as 1865 Galton was largely dismissive of the notion of the inheritance of acquired characteristics. During the early 1870s he experimentally tested Darwin's theory of 'Pangenesis'—according to which bodily tissues throw off the germ cells from which offspring are produced—by transfusing blood among different varieties of rabbit. His predictable conclusion brought no pleasure to Darwin, and Galton's 1871 rejection of Pangenesis in *Nature* elicited from him a swift rebuke. Subsequently, Galton's resistance to the inheritance of acquired characteristics hardened. In

the models of hereditary transmission he published during the first half of the 1870s, he claimed that the 'patent' hereditary units that become bodily structures only rarely return to the reproductive organs to be transmitted to progeny. In contrast, the 'latent' elements remain in the testes and ovaries ready to be transmitted to the next generation. In this rather circumscribed sense his hereditarian model did foreshadow Weismannism.

However, aside from the many technical inconsistencies between Galton's and Weismann's concepts of heredity, it is important to note that after 1875 Galton departed significantly from the position outlined above. He gradually came to appreciate that his case for the continuity of traits between parents and offspring was weakened if it were assumed that the latter inherited only the former's latent elements and very few of their patent elements. As a result, by the late 1880s Galton was writing of patent and latent elements as being equally heritable. Whilst he remained only mildly sympathetic to the notion of the inheritance of acquired characteristics, he had abandoned his earlier emphasis upon the near-absolute separation of bodily from reproductive cells.

Galton's dual fascination with heredity and quantification also generated an interest in anthropometry, the measurement of bodily morphology. During the 1870s, appalled that so little was known about the physique of the British people and fearful that a biological decline was already in progress, Galton sought to obtain anthropometric information on children from different social classes. Subsequently, he became a member and later chairman of a British Association committee investigating occupational disparities in size and weight. Throughout his involvement with this committee Galton was convinced that heredity was the chief determinant of anthropometric variation. As a means of procuring raw data from which to study hereditary transmission in humans, in 1884 he set up an elaborate anthropometric laboratory at the International Health Exhibition in London. Equipment was installed to measure the height, weight, breathing-capacity and, *inter alia*, the strength of pull of members of the public. Also measured were such qualities as keenness of sight, hearing, and color discrimination.

These last tests reflect Galton's interest in what he dubbed 'psycho-physics,' the scientific investigation of mental phenomena. Generalizing his earlier conviction that individual nurture is a subordinate force to heredity, he emphasized aspects of unconscious cerebral function in which the 'will' seems to play little if any role. In his 1883 *Inquiries into Human Faculty*, Galton described and, where feasible, quantified a range of unconscious and sensory phenomena, including number forms, visual imagery, color associations, visions, and sensory awareness among the blind. He also documented the results of his introspective investigations into the power of the individual

will. In the course of these studies, Galton devised the method of 'word-association' with which to demonstrate his claim that very few thoughts or actions are ever the spontaneous product of the will but relate to desires, ideas and associations of which we have little conscious awareness.

In addition, Galton was a pioneer in the design and employment of composite photography. He hoped that through superimposing numerous photographic portraits on top of each other he would be able to discern the common hereditary features shared by families or groups. His main work in this area involved an attempt to discover the physiognomical indices of different forms of criminality from a series of photographs supplied by the Inspector of Prisons. To his disappointment, Galton found no evidence of distinctive criminal types. However, this interest in hereditary variation, measurement and their relevance to criminal investigations also culminated in his extensive study of fingerprints. Galton was not an innovator in this field but his several publications on the subject drew attention to the enormous value of fingerprints as a means of personal identification.

Although it is clear that most of Galton's scientific work stemmed from his early fascination with heredity, his scientific curiosity was nonetheless polymathic in breadth. Over the course of half-a-century's involvement with the Royal Geographical Society, he was able to make significant contributions to the expeditions of such explorers as R. Burton, J. Speke, and J. Grant, and he worked hard to place geography on a more scientific footing. As an active meteorologist he was instrumental in the construction of weather maps and was responsible for discovering the 'anti-cyclone' in 1862. Finally, he was an enthusiastic advocate of establishing scientific activity as an occupation of dedicated, professional and, if necessary, salaried, researchers. Alongside T. H. Huxley, John Tyndall, Joseph Hooker, and others, he campaigned vigorously to free science from the limitations imposed by religious orthodoxy. To this end, many of Galton's works carried thinly veiled attacks on theology and the Church, most strikingly his 1872 attempt to prove statistically that prayer is inefficacious.

In retrospect, there can be little doubt that Galton gave an important impetus to the scientific study of heredity and the development of psychometric testing. Despite the shortcomings later revealed in his own work, he was the first to subject human intellectual attributes to systematic numerical analysis. With the aid of W. F. R. Weldon and Karl Pearson, he was also intimately involved with establishing a statistical approach to the study of evolution. Perhaps Galton's most lasting contributions to science arose from his statistical breakthroughs of the 1870s and 1880s. Statistical procedures derived from his initial insights are now indispensable to both the natural and the social sciences. More practically, his private wealth enabled him to set up a Eugenics Research Office at

University College London in 1904. He honored and extended this commitment upon his death (on January 17, 1911) when he created the Galton Professorship of Eugenics and named Karl Pearson as the Chair's first holder.

Any summary of Galton's legacy must recognize that eugenic ideas were implemented after his demise with horrific consequences, and that the hereditarian ideologies he promoted have often been used as means of social, sexual and racial oppression. In the wake of the Nazi atrocities of the 1930s and 1940s, and of the civil rights movements of the 1960s and 1970s, Galton's stock as an important scientific figure has declined rapidly. However, in the interests of a proper understanding of the past, it should be pointed out that to implicate Galton in the rise of twentieth-century eugenic practices is to over-personalize a complex series of events and set of national circumstances. Eugenic measures were linked to a wide range of social, economic and political factors against which Galton's advocacy is scarcely more than trivial. Furthermore, even in 1865 the notion of selective human breeding was far from novel: Galton's originality lay only in buttressing the idea with apparently rigorous scientific research.

Bibliography

- Forrest D W 1974 *Francis Galton: The Life and Work of a Victorian Genius*. Paul Elek, London
- Galton F 1863 *Meteorographica, or Methods of Mapping the Weather*. Macmillan, London and Cambridge
- Galton F 1869 *Hereditary Genius*. Macmillan, London
- Galton F 1874 *English Men of Science: Their Nature and Nurture*. Macmillan, London
- Galton F 1883 *Inquiries into Human Faculty and its Development*. Macmillan, London
- Galton F 1889 *Natural Inheritance*. Macmillan, London
- Galton F 1908 *Memories of My Life*. 2nd edn. Methuen, London
- Kevles D J 1986 *In the Name of Eugenics: Genetics and the Uses of Human Heredity*. Penguin, Harmondsworth, UK
- Keynes M (ed.) 1993 *Sir Francis Galton, FRS: The Legacy of His Ideas*. Macmillan, Basingstoke and London
- Pearson K 1914/1930 *The Life, Letters and Labours of Francis Galton*. Cambridge University Press, Cambridge, UK
- Stigler S M 1986 *The History of Statistics: The Measurement of Uncertainty Before 1900*. The Belknap Press of Harvard University Press, Cambridge, MA

J. C. W. Waller

Game Theory

1. Introduction

Game theory is the mathematical study of situations of conflict of interest. It is applicable to military

situations, but has also been applied to economics and other social sciences, and even to biology.

We think of game theory as the fourth in a sequence of paradigms, explaining how people attempt to achieve goals; call these type I, II, III, and IV thinking: in type I thinking, the question is whether a goal can be met. In type II thinking, the decision maker, satisfied that the goal can be met, tries to meet this goal in the most efficient way possible (at minimal cost); alternatively, having decided that the goal cannot be met, he tries to come as close as possible to meeting the goal. In type III thinking, the decision maker realizes he is faced, not by nature, but by other decision makers whom he can influence. He therefore tries to influence them so that they (acting according to type II thinking) will carry out actions helpful to him. In type IV thinking, the decision maker realizes that he is faced by other individuals whose actions can affect him, and who in turn are trying to influence his actions. Thus he will generally try either to outsmart the others, or to cooperate with some or all of them. It is this last type which represents game theory.

A game is any situation in which two or more agents (players), not having identical interests, can affect the outcome. We assume that all players seek to maximize their individual utility. There may or may not be random moves.

Games can be noncooperative or cooperative. In the noncooperative case, we are mainly interested in the strategies chosen by each player. In the cooperative case, we are also interested in the process of bargaining and coalition formation. Games may also have complete or incomplete information. In case of incomplete information, there is interest in the process of learning, as players elicit information during the play of a game.

Essentially, given a game situation, game theory tries to answer the following questions:

- What strategies should each player use?
- Where there are gains to be made from co-operation, how should the players bargain?
- Where there are many players, what coalitions should form?
- What information can be deduced from other players' actions?

2. Extensive and Normal Forms

There are several ways of representing a game. The simplest perhaps is to draw a graph, with nodes representing the several positions in the games, and arcs representing moves. This is the extensive form of the game. For example, consider the simple game of matching pennies:

Example 1. Player I puts a coin on the table, with either heads (H) or tails (T) up. Player II, not seeing the coin, calls either H or T. If II has called correctly, she wins \$1 from I. If the call is incorrect, I wins \$1 from II.

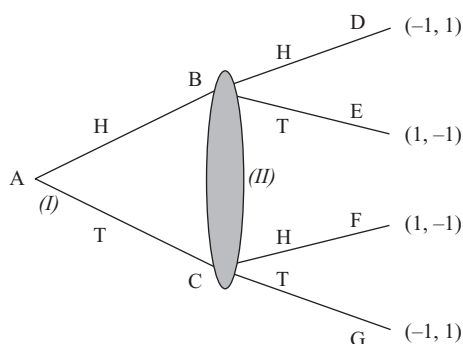


Figure 1
The game of Matching Pennies

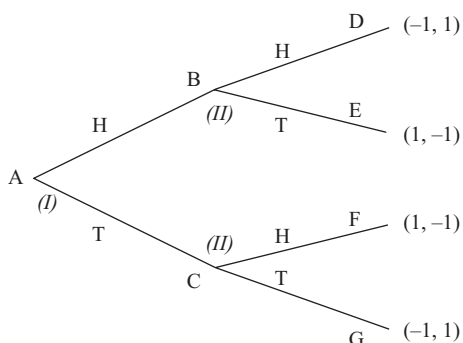


Figure 2
The game of Matching Pennies, modified

This game is represented by the graph in Fig. 1. Play starts at node A, which 'belongs' to player I (it is his turn to play there). Depending on I's choice of H or T, play then moves to either node B or node C, both of which belong to II. After II makes her choice, play terminates at one of the nodes D, E, F, or G, all of which are terminal nodes of the game. Numbers are assigned to the terminal nodes: they are payoffs for the game. Thus, at D, the payoff $(-1, 1)$ means that, since II guessed H correctly, I must pay a dollar to II. Similarly at the other terminal nodes.

Note the connection (shade) between nodes B and C. This is meant to represent II's lack of information when she moves here: she knows the position is either B or C, but cannot distinguish between them. Thus B and C are said to belong to the same information set.

Example 2. Suppose, in the previous example, II were allowed to see how I had placed the coin. Otherwise, assume the game is unchanged. In this case, we obtain the graph shown in Fig. 2. The difference between the two figures is that nodes B and C are now in different information sets. II, when moving here, can distinguish between these two positions.

Alternatively, it is frequently more convenient to represent a game by means of its strategies. A strategy is a rule that tells a player what to do in each conceivable situation, i.e., at each of his information sets. Thus, in example 1, each player has two strategies, namely H and T. In example 2, Player I still has only two strategies, H and T. Player II, however, has two information sets (vertices B and C) at each of which she has two alternatives. This gives her a total of four strategies, which we might call S(HH), S(HT), S(TH), and S(TT) (where HT means 'call H in position B, T in position C,' etc.).

For a given choice of strategies (one strategy for each player), the outcome of the game is determined except for the effect of random moves. Assuming that the probability distributions at chance moves are well defined, and known to the players, it is quite straightforward to obtain expected payoffs for each choice of strategies. The function which assigns to each n -tuple of strategies these expected payoffs is known as the normal form of the game.

For the game of example 1, above, the payoff is given in the form of a table or, more precisely, a 2×2 matrix, in which the rows correspond to Player I's choice of strategy, and the columns to II's choice. The entries are pairs of numbers, representing the payoffs:

	H	T
H	$(-1, 1)$	$(1, -1)$
T	$(1, -1)$	$(-1, 1)$

For the game of example 2, the normal form is given as a 2×4 matrix, much as in example 1:

	S(HH)	S(HT)	S(TH)	S(TT)
H	$(-1, 1)$	$(-1, 1)$	$(-1, 1)$	$(-1, 1)$
T	$(1, -1)$	$(1, -1)$	$(1, -1)$	$(1, -1)$

3. Two-person Zero-sum Games

3.1 The Game Matrix

In the simplest of cases, a game situation will have only two players, with directly opposed interests. This can be thought of as a closed system in which, whatever one player wins, the other loses. Thus the sum of the two players' utilities will be zero at all times. It follows that it is not necessary to state explicitly Player II's payoff: this is always the negative of that to Player I. The outcome of such a game can be represented, then, by a single number: I's payoff. The normal form of such a game becomes a matrix in which, as before, rows represent I's strategies, and columns, II's strategies.

tegies. The entries are payoffs to I, which I wishes to maximize, and II, to minimize. Thus, for the game of example 1, the normal form becomes the matrix:

$$\begin{array}{cc} -1 & 1 \\ 1 & -1 \end{array}$$

and, for example 2, it becomes:

$$\begin{array}{cccc} -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 \end{array}$$

3.2 Equilibria and Saddle Points

In a desire to 'solve' games, we look for equilibria, i.e., choices of strategies such that no one of the players can gain by a unilateral change of strategy. In a zero-sum matrix game, a pair of strategies will be in equilibrium if the payoff they define is simultaneously the largest in its column and smallest in its row. This is frequently called a *saddle point*, by analogy with the surface of a (riding) saddle.

Example 3. In the matrix game:

$$\begin{array}{ccc} 6 & 1 & 2 \\ -4 & 0 & 9 \\ 8 & -1 & -4 \end{array}$$

the entry in the first row, second column is a saddle point. Similarly, in example 2, above, both entries in the second column are saddle points.

When a matrix game has a saddle point, game theory recommends that the players choose the strategies giving rise to that point. Unfortunately, as in example 1 above, most matrix games will not have saddle points. Another example of a game without a saddle point is:

Example 4:

$$\begin{array}{ccc} 5 & 1 & 0 \\ -1 & 2 & 1 \\ 1 & -3 & 3 \end{array}$$

3.3 Mixed Strategies

If a game does not have a saddle point, it is important to keep one's opponents from learning one's plans. Game theory then recommends the use of a mixed strategy. In a mixed strategy, a player chooses among his strategies (the rows and columns of the matrix,

hereafter called pure strategies) by a randomization scheme. The mixed strategy is then an assignment of probabilities to the several pure strategies. Thus it may be possible for an opponent to compute one's mixed strategy (since this particular assignment of probabilities is chosen in a rational manner), but it is impossible to know the actual pure strategy used in any play of the game. We generally represent a mixed strategy (for a given player) as a vector with m components (m being the number of pure strategies that player has), non-negative and adding to 1.

Suppose, then, that in matrix game A, Player I has chosen the mixed strategy x , while II has chosen y . It is not too difficult to see that, assuming these two randomization schemes are independent, there will be an expected payoff equal to the matrix product xAy . Once again, I wishes to maximize, while II wishes to minimize, this quantity.

The question now is how the mixed strategies should be chosen. A very important theorem, the minimax theorem, discovered independently by E. Borel and J. von Neumann during the 1920s, tells us that, for any matrix game, there exist mixed strategies, x^* and y^* (optimal strategies) and a number v (the value of the game) such that, if I uses x^* , he will always have an expected payoff at least equal to v , no matter what II does. Similarly, if II uses y^* , the expected payoff will never be greater than v , no matter what I chooses. The triple $(x^*; y^*; v)$, consisting of two optimal strategies and the value, is the solution of the game.

As an example, note that, in example 1 above, the optimal strategies are $x^* = (1/2, 1/2)$ and $y^* = (1/2, 1/2)$, guaranteeing the value $v = 0$. In example 4, the optimal strategies are $x^* = (2/7, 4/7, 1/7)$ and $y^* = (1/7, 2/7, 4/7)$, guaranteeing the value $v = 1$.

4. Two-person General-sum Games

In other games, the players' interests will not be directly opposed. Thus cooperation between the two players may well be possible, unless the rules of the game (e.g., antitrust laws) specifically forbid this. We therefore consider two cases: the noncooperative and the cooperative game.

4.1 Noncooperative Games

In the noncooperative case, the emphasis has generally been on equilibrium points, (still defined as pairs of strategies such that neither player can gain by a unilateral change). Much as in the case of matrix games, it can be shown that, when mixed strategies are allowed, every finite n -person game has at least one equilibrium n -tuple of strategies. Unfortunately, equilibrium points are not generally as satisfactory here as in the case of matrix games. In fact, a game might have

several different equilibrium points; moreover, even if a game has a unique equilibrium point, it is not clear that it will be best for players to play in equilibrium.

Example 5. In the two-person (non-zero-sum) game:

(4, 1)	(0, 0)
(1, 0)	(2, 5)

the strategy pairs (row 1, column 1) and (row 2, column 2) are both in equilibrium. Note, however, that Player I clearly prefers the first of these, while II prefers the second. It is not clear how the game should then be played.

Example 6. In the game:

(5, 5)	(0, 10)
(10, 0)	(1, 1)

(usually known as the prisoner's dilemma), the pair (row 2, column 2) is the unique equilibrium point. Note, however, that both players will receive a higher payoff if (row 1, column 1) is used.

4.2 Cooperative Games

In case cooperation is allowed, it is usually possible for both players to gain by profitable deals. Since, however, many such deals may be possible, the question of bargaining comes into play. Essentially, the search in such games has been for a 'fair' arbitration rule, one that takes into account what each player can do for himself (without the other player's help) and what the two of them can accomplish together. J. Nash in 1950 developed an axiomatic model for this. Assuming a set S of possible bargains (represented as pairs of payoffs), Nash looks for an outcome (bargain) which will be Pareto-optimal (it is not possible for one player to gain more without taking something from the other) and individually rational (each player gains at least as much as he could by himself) and which moreover satisfies certain technical conditions. Nash shows that the only arbitration rule which satisfies these axioms is the rule that chooses the bargain which maximizes the product of the two players' increments in utility (over what they could get by themselves).

Example 7. Two players are to divide the sum of \$100; if they do not agree on a division of the \$100, each gets \$0. Assume player I's utility is proportional to the amount of money he receives; for Player II, however, suppose that \$ y represents a utility equal to the logarithm of $100 + y$.

In this case, any division (x, y) with $y \leq 100 - x$ is permissible. This will give rise to any utility payoffs of the form (x, w) , where $w \leq \log(200 - x)$. Nash's solution is the point which

maximizes the product $x(w - \log(100))$, subject to the constraint above. Using differential calculus, it is not difficult to see that this maximum is attained at $x = 54.4$, so that I should receive \$54.40 while II receives \$45.60.

Nash's system of axioms is not the only one possible. Several other systems of axioms, as well as nonaxiomatic approaches, have been suggested. There is, unfortunately, no generally accepted solution to the bargaining problem.

5. n-Person Games

5.1 The Characteristic Function

In case there are three or more players, the cooperative game is further complicated by the fact that each player might choose among several possible coalitions. J. von Neumann and O. Morgenstern (1947) approached these games through the *Characteristic Function*. Let $N = \{1, 2, \dots, n\}$ be the set of all players in a game. For a given coalition S (nonempty subset of N), $v(S)$ is the amount of utility which the members of S can obtain jointly. It is assumed that this utility can be divided among the members of S in any way, if they agree. Thus v , the characteristic function, assigns a number to each coalition. The problem then is to look for a division of the total amount $v(N)$ which is available in the game, in such a way that each player, i , receives at least the amount $v(\{i\})$. Any payoff vector that does this is an imputation of the game v .

Von Neumann and Morgenstern then introduce the notion of domination. Given two imputations, x and y , we say that x dominates y if there is at least one coalition S such that (a) $x_i > y_i$ for every $i \in S$ (so that all members of S prefer x), and (b) $\sum_{i \in S} x_i \leq v(S)$ (so that S is strong enough to obtain x).

5.2 The Core and Stable Sets

One possibility, now, is to look for the undominated imputations. The set of these is known as the core, and is closely allied to the economic concept of market equilibrium.

Example 8. In a three-person game, player 1 owns a horse for which she has no use. Players 2 and 3 would both like to buy the horse; player 2 values it at \$200 while 3 values it at \$100. We will equate money with utility.

It may be seen that any single player can create no utility in this game. Thus $v(\{i\}) = 0$ for $i = 1, 2, 3$. Similarly, $v(\{2, 3\}) = 0$ since the two buyers together can do nothing.

Now, if 1 sells the horse to 2 for \$ x , then 1 gains x (she has exchanged a useless horse for \$ x), while 2 gains $200 - x$ (he has obtained a horse worth \$200 for \$ x). Together, then, they gain 200, and we set $v(\{1, 2\}) = 200$. Similarly, $v(\{1, 3\}) = 100$ and $v(\{1, 2, 3\}) = 200$.

For this game, an imputation is any vector (x_1, x_2, x_3) with non-negative components adding to 200. To be in the core, the imputation must additionally satisfy the conditions

$x_1 + x_2 \geq 200$ and $x_1 + x_3 \geq 100$. This can only happen if $100 \leq x_1 \leq 200$, $x_2 = 200 - x_1$, and $x_3 = 0$. In effect, it seems that 1 will sell the horse to 2 for a price between \$100 and \$200. Player 3 receives nothing, but helps to drive the price up. (Note that, if player 3 were removed from the game, the price could be anything between \$0 and \$200.) We should point out that, in classical economics, such a situation would reach equilibrium with precisely such a price.

In many cases, unfortunately, the core is empty, and von Neumann and Morgenstern thus develop the concept of a solution (now better known as a stable set). A stable set (for a given game) is a set V of imputations such that (a) no imputation in V dominates another in V and (b) every imputation not in V is dominated by at least one in V . For some two decades, it was conjectured that every game in characteristic function form had at least one stable set, but W. F. Lucas (1967) gave an example of a ten-person game with no stable sets.

Example 9. In a three-person game, there is a trunk full of valuable objects, worth 100 units of utility. This trunk is so heavy that no one of the three players is strong enough to lift it; alternatively, any two acting together can carry the trunk and collect the utility.

We represent this game by the characteristic function u , where $u(\{i\}) = 0$ for $i = 1, 2, 3$, and $u(S) = 100$ for any coalition S with two or three members. For this game, we see that the imputations are vectors with three nonnegative components adding to 100. Now, for an imputation to be in the core, each two-person coalition must receive at least 100 units. But this means that the players must receive on average at least 50 units; i.e., the three together must receive at least 150 units. This is impossible since there are only 100 units available, and we conclude that, for this game, the core is empty.

However, it is not difficult to check that the set V , consisting of the three imputations (50, 50, 0), (50, 0, 50), and (0, 50, 50) is a stable set. (This is not the only stable set for this game.)

5.3 The Shapley Value

Both the core and the stable set solution concepts suffer from the fact that (a) the first may be empty, the latter might not exist, and (b) in many cases they contain an enormous number of imputations. Game theorists have therefore looked for other solution concepts. Foremost among these is L. S. Shapley's value.

First developed in 1953, the Shapley value or power index is based on the three axioms of efficiency (if a game is played entirely by a set of players, these players will share the total available utility), anonymity (the value depends only on the players' possibilities, and not on the names given them) and additivity (the value of the sum of two games is equal to the sum of the two games' values). Shapley proved

that there is a unique value scheme satisfying these three axioms over the set of all n -person games, and this scheme assigns to player i , in game v , the quantity $\phi_i[v]$ given by the formula

$$\phi_i[v] = \sum_{S \subset N} \gamma_{n,s} [v(S \cup \{i\}) - v(S)]$$

where the sum is taken over all coalitions, S , and for these, s is the number of players in S , while $\gamma_{n,s} = s!(n-s-1)!/n!$.

As an example, the horse game (example 8) has value $\phi[v] = (116.7, 66.7, 16.7)$. The game of example 9, being totally symmetric, has value $\phi[u] = (33.3, 33.3, 33.3)$.

The Shapley value is also very useful as a measure of voting power in political situations. A voting structure is a game v whose players are the voters, N , and such that, for any coalition S , $v(S) = 1$ if S has enough votes to win an election (or approve a motion), and $v(S) = 0$ otherwise.

Example 10. In a committee with four members, assume member 1 (the chairman) has the power to break ties. In this case, $v(S) = 1$ if S has three or four members, or if S has two members, including the chairman. For this game, the value is $\phi[v] = (0.5, 0.167, 0.167, 0.167)$.

5.4 The Bargaining Set(s)

Of other solution concepts for n -person games, we mention only the bargaining sets, developed by R. Aumann and M. Maschler (1995). Given a coalition structure, $T = \{T_1, T_2, \dots, T_m\}$ (a partition of the set of players) we consider only payoff vectors $\mathbf{x}$ which give to each of the sets T_j (the coalitions which are assumed to form) their total possibility $v(T_j)$ and to each player, i , at least $v(\{i\})$. Such pairs $\langle T; \mathbf{x} \rangle$ are individually rational payoff configurations (i.r.p.c.'s). Now, given an i.r.p.c., a player, i , can object against another player, k , if i can receive greater utility in another i.r.p.c. $\langle U; \mathbf{y} \rangle$ in which she does not have player k as a coalition partner. Of course, k can counterobject if there is yet another i.r.p.c. $\langle V; \mathbf{z} \rangle$ in which he does not need i as a partner, and such that $z_k \geq x_k$. In this case, the objection is unjustified; otherwise (i.e., if there is no counter-objection) the objection is justified.

Now, an i.r.p.c. is said to be stable, or belong to the bargaining set $\mathbf{M}$, if there are no justified objections. Because of different possible relations allowed between the objecting and objected players, i and k , there are several somewhat different bargaining sets. Aumann and Maschler prove that in one case, namely when i and k have to be partners in the original structure $\mathbf{T}$, the bargaining set—here denoted $\mathbf{M}_1^{(i)}$, is nonempty in the sense that, for any game v , and any coalition structure $\mathbf{T}$, there exists at least one $\mathbf{x}$ such that $\langle T; \mathbf{x} \rangle$ belongs to $\mathbf{M}_1^{(i)}$.

Example 11. In the three-person game of example 9, if the coalition structure is $T = \{\{1, 2\}, \{3\}\}$, the only stable payoff vector is (50, 50, 0). If, instead, the grand coalition ($T = \{N\}$) forms, then the only stable payoff vector is (33.3, 33.3, 33.3).

In the four-person game of example 10, if the coalition structure $T = \{\{1, 2\}, \{3\}, \{4\}\}$ forms, then the stable payoffs are of the form $(t, 1-t, 0, 0)$, where $0.5 \leq t \leq 0.67$.

6. Games with Incomplete Information

In some cases, players are unaware of some of the available strategies, or might be uncertain as to the payoffs which correspond to certain choices. In particular, they may be unaware of the utilities which other players assign to certain outcomes (e.g., we frequently do not know whether we are dealing with an honest or dishonest person—an honest person being one who assigns high utility to keeping an agreement). In such situations—known as games with incomplete information—repeated play will often give us further information—a person who does not keep an agreement may simply have made an honest mistake, but if this happens repeatedly, we generally conclude he is dishonest. Of course, a dishonest person, realizing this, will frequently keep some less important agreements so as to ‘lull’ others in preparation for a big scam. In some important work dating back to the 1960s, J. Harsanyi (1977) on the one hand, and Aumann and Maschler on the other, address this problem of learning and deception.

7. Conclusion

Game theory represents an important attempt to give a mathematical basis to behavior in conflict situations. Admittedly, there are elements of human activity which are not easily modeled, and this leads to certain gaps in the theory. Nevertheless, it has served to give a mathematical structure to several well-known ideas (bluffing, bargaining) and to develop new insights into human behavior. As of this writing, it is considered an important mathematical tool for economists and other social scientists.

See also: Bounded and Costly Rationality; Decision and Choice: Economic Psychology; Decision Theory: Bayesian; Game Theory and its Relation to Bayesian Theory; Game Theory: Noncooperative Games; Luce’s Choice Axiom; Mathematical Psychology

Bibliography

- Aumann R J, Maschler M B, Stearns R E 1995 *Values of Repeated Games with Incomplete Information*. MIT Press, Cambridge, MA
 Damme E van 1987 *Stability and Perfection of Nash Equilibria*. Springer, Berlin

Harsanyi J C 1977 *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press, Cambridge, UK

Luce R D, Raiffa H 1957 *Games and Decisions*. Wiley, New York

Owen G 1995 *Game Theory*, 3rd edn. Academic Press, San Diego, CA

von Neumann J, Morgenstern O 1947 *The Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ

G. Owen

Game Theory and its Relation to Bayesian Theory

This entry explores some open controversies relating to a contrast between rival decision paradigms. Alas, it cannot be a report on a unified theory of games: such a thing does not exist. The controversies between Bayesian decision theory and traditional Game Theory have a common source, however, which may be identified informally, as follows.

Bayesian decision theory comes in many varieties, Good (1971). Common to all is one rule: the principle of maximizing (subjective) conditional expected utility. Generally, an option in a decision problem is depicted as a (partial) function from possible states of affairs to outcomes, each of which has a value represented by a (cardinal) utility. Where Bayesian decision makers are uncertain about which state of affairs obtains, they assign conditional probabilities to those states, conditional upon the choice of the option. The Bayesian decision rule is to choose an option from the set of available options that maximizes subjective conditional expected utility—assuming a maximum exists, as will be done here. (Some Bayesian theories, e.g., Savage (1954), use a restricted formulation where acts and states are probabilistically independent, and dominance considerations apply. Then the Bayesian decision rule is simplified: maximize subjective expected utility.) Thus, the primary Bayesian decision rule assesses the value of an option by a weighted average of the utilities of potential outcomes, weighted according to the decision maker’s (conditional) personal probability over states (see *Decision Theory: Bayesian*).

Traditional Game Theory, by contrast, requires that rational players choose strategies that form an equilibrium—though which variety of equilibrium they adopt is also a matter of serious debate among game theorists, as noted below. However, under each sense of ‘equilibrium’ that is proposed, a player is required to assess the value of a strategy in terms of a best response that an opponent might make against it. That is, by game theoretic standards, against each available strategy that the player might select, typically

there is one state of affairs, namely the *best reply* that might be chosen by the (collective) opponent(s). That *best reply* is what determines the value of the strategy for the player. When all the players' choices jointly meet the criterion of being best replies, a *Nash equilibrium* results.

At the core of the debate, then, is this conflict. On the one hand is the Bayesian decision rule as that might be applied to games. It is a rule that specifies the value of a player's choice of a strategy in a game as a weighted average over possible replies that an opponent might make. Here, the set of potential strategies available to the (collective) opponent(s) constitute the states of affairs over which the Bayesian player is uncertain, and the weights assigned to these are the player's subjective probability distribution for what the opponent might do. On the other hand, there is the game theoretic rule that uses an opponent's best reply as the primary criterion to determine the value of a player's choice of strategy. Of course, since a distribution with all its weight concentrated on one point—corresponding to the state of affairs that is the opponent's *best reply*—is not precluded by Bayesian theory; sometimes, but usually not, the two approaches agree. But Bayesian theory generally does not privilege the Nash equilibrium solution. Before we examine games where the two approaches give different recommendations, and some recent attempts to mitigate these differences, it is worthwhile to review a few points of shared agreement.

1. Illustrations Where Expected Utility Considerations Are Welcome Within Traditional Game Theory

1.1 Decisions Under Risk

First, in their *magnum opus*, *The Theory of Games and Economic Behavior*, von Neumann and Morgenstern (1944), formalize the assumption that each player's preferences for outcomes in a game can be quantified by a (cardinal) utility. That is, they assume that a player's preferences over outcomes satisfy the expected utility rule applied to decisions under *risk*. Specifically, when options are presented as *lotteries* over outcomes, i.e., when a (simple) probability distribution for outcomes is well defined, then von Neumann and Morgenstern assume that the decision maker's preferences accord with the Bayesian rule of maximizing expected utility. This assumption is critical to the received game theoretic interpretation of *mixed strategies*, whose values are given by expectations over the numerical outcomes of pure, i.e., non-mixed strategies.

That is, the numerical values used to describe outcomes in games depend for their meaning on the application of expected utility in decisions under *risk*. This much of von Neumann and Morgenstern's theory

is in common with Bayesian decision theory, though it involves a game with one player. Of course, Bayesian decision theory, but not Game Theory, extends the rule to maximize expected utility also to decisions under *uncertainty*—where well-defined probabilities over states of affairs are not specified by the assumptions of the decision problem. In particular, Game Theory does not extend the Bayesian decision principle to problems under uncertainty with more than one player where the relevant states of affairs are decisions by an opponent!

1.2 deFinetti's Game of Coherent Previsions

In a seminal work, deFinetti (1937) proposed a game theoretic basis for using subjective expected utility. Here is a simplified version of that game—simplified to avoid important technical considerations involving deFinetti's use of merely finitely additive probability that arise with infinite state spaces.

Let $\Omega = \{\omega_1, \dots, \omega_n\}$ be a finite set of pairwise exclusive and mutually exhaustive states, i.e., Ω is a finite partition of the sure event. Call the elements of Ω the *states*. Let $\{X_i; i = 1, \dots\}$ be a set of random variables defined on Ω . That is, each $X_i: \Omega \rightarrow \mathcal{R}$ is a simple, real-valued function defined on the partition Ω . *Player*₁, whom we call the *bookie* in the game, is obliged to play first and to reveal a set of real-valued quantities, called *previsions* $\{p_i; i = 1, \dots\}$. A set of previsions constitutes a strategy for the *bookie*. (The game requires that the bookie reveals a pure strategy.) *Player*₂, whom we call the *gambler* in the game, plays second and, having seen the strategy chosen by the *bookie* is then allowed to choose a set of real-valued terms $\{\alpha_i; i = 1, \dots\}$, subject to the constraint that only *finitely* many of these terms may differ from zero. (For added simplicity in this presentation, assume that $\sum_i |\alpha_i| \leq 1$.) A set of terms constitutes a strategy for the *gambler*.

Next, consider the (simple) random variable, $Y = \sum_i \alpha_i X_i - p_i$, whose value $Y(\omega_i) = y_i$ is fixed by the state. Given a strategy choice by each player, the payoff to the *bookie* in deFinetti's game is the maximum $\max_i \{y_i\} = y^*$, while $-y^*$ is the corresponding payoff to the *gambler*. Hence, this is a 2-person, zero-sum game in which the *bookie* plays first.

Since the *gambler* may choose all the terms equal to 0, in which case $Y = 0$ for sure, the *gambler* can ensure that the value of the game to the *bookie* is bounded above by 0. Now, either the *bookie's* previsions $\{p_i; i = 1, \dots\}$ form a set of expected values for the random variables $\{X_i; i = 1, \dots\}$, based on a single probability distribution P over Ω , or they do not. deFinetti showed that one of two circumstances obtain, respectively: If there exists a probability P such that $E_P[X_i] = p_i$, then the *gambler* can do no better than securing the value 0 for the game, and the *bookie* is said to have *coherent previsions*.

Aside: This can be seen as follows. When the bookie's previsions are coherent, then $E_P[Y] = 0$ regardless of the gambler's choice of strategy. But a weighted average of terms (y_i) cannot exceed its maximum term (y^*); $0 = E_P[Y] \leq y^*$. Thus, the coherent bookie is assured a value of at least 0.

If, however, the *bookie's* chosen strategy does not agree with the expectations from a probability P , then the gambler may find a reply that insures a strictly negative outcome to the *bookie*: then the *bookie* is said to have *incoherent previsions*. That is, if and only if the *bookie* does not play as though the *previsions* are (subjective) expected values of the random variables, calculated with some subjective probability P , the *bookie* is in jeopardy to suffer a sure loss from a clever choice of strategy by the *gambler*. Thus, though the *bookie's* prudent, equilibrium strategy is to play a coherent set of previsions, just as though maximizing expected utility, which equilibrium from the continuum of coherent sets of previsions is chosen remains unspecified.

Aside: When the random variables are indicator functions for events, i.e., when they are indicators for subsets of Ω , then this result reduces to the familiar Dutch Book Theorem (see *Probability: Interpretations*).

For an original application of deFinetti's prevision game to problems of cooperative decision making among several agents, see Nau's interesting work (1999).

1.3 Admissible Statistical Decision Functions

Wald (1950) proposed a very general theory of statistical decisions where, in effect, the statistician plays a zero-sum game against Nature. The statistician chooses a statistical decision rule and Nature fixes the unknown state of affairs. In outline and simplified for presentation here, Wald's influential program for framing statistical problems uses the following structure. Let θ be a real-valued statistical parameter with parameter space Θ , a partition of the relevant states of affairs. Let X be a random value whose distribution $P_\theta(X)$ depends upon the parameter θ . The statistician anticipates observing $X = x$ and in advance of the observation chooses a statistical *decision rule*, denoted $d(\cdot)$, that may depend upon the value of X to be observed. The result to the statistician of the *terminal decision* $d(x)$ depends upon the unknown parameter θ . The value of that result is specified by a loss function, $L(d(x), \theta)$, understood as the negative (cardinal) utility of the consequence to the statistician of $d(x)$ when θ obtains. In Wald's presentation, typically the loss function is nonnegative. When, for each value of θ the minimum (over the set of terminal decision) of the loss function is 0, the utility for payoffs is then said to be formulated as a *regret* function.

Aside: In Bayesian decision theory, when acts and states are probabilistically independent and utility is bounded—as in Savage's theory—then, without loss of generality, the decision problem may be reformulated with utility represented by a regret function.

Here are key definitions in Wald's theory.

The *risk function* R for a decision rule $d(\cdot)$ is the expected loss of adopting $d(\cdot)$ as a function of the unknown parameter θ .

$$R(d(\cdot), \theta) = E_\theta[L(d(\cdot), \theta)].$$

A decision rule $d'(\cdot)$ is *inadmissible* if there exists another decision rule $d(\cdot)$ such that, for each value θ the risk for d is never greater than the risk for d' , and for some value of θ the risk for d is less. In that case, say that rule d *dominates* rule d' .

That is, for each θ $R(d(\cdot), \theta) \leq R(d'(\cdot), \theta)$, and for some θ $R(d(\cdot), \theta) < R(d'(\cdot), \theta)$.

A decision rule that is not dominated is *admissible*.

Given a probability distribution $P(\theta)$ for the parameter, the Bayes-risk r for a decision rule $d(\cdot)$ is the expected risk with respect to P of adopting $d(\cdot)$.

$$r(d(\cdot), P(\theta)) = E_P[R(d(\cdot), \theta)] = E_P[E_\theta[L(d(\cdot), \theta)]].$$

A decision rule $d(\cdot)$ is *Bayes-admissible with respect to the distribution* $P(\theta)$ if, for each dominating rule $d'(\cdot)$, the set of parameter values where $d'(\cdot)$ has strictly lower risk has P -probability 0.

A decision rule $d(\cdot)$ is minimax for the statistician if it minimizes the maximum risk over the parameter space (see *Decision Theory: Classical*).

The following result (Berger (1985, p. 254) or Schervish (1995 p. 155) helps to explain the relevance of Bayesian expected utility considerations in identifying admissible decision rules.

Theorem: Suppose that the Bayesian statistician's 'prior' distribution $P(\theta)$ gives positive probability to each open interval. Assume that the risk function $R(d(\cdot), \theta)$ is continuous in θ for each decision rule $d(\cdot)$. Then if $d^*(\cdot)$ is Bayes-admissible with respect to P , it is admissible.

The resulting connection of Bayesian decision rules to Game Theory, as first developed by Wald, is indicated by the following (see Berger (1985, p. 353) or Schervish (1995, p. 167)).

Theorem: If decision rule $d(\cdot)$ has constant risk and if it is admissible, then it is minimax.

Thus, if $d^*(\cdot)$ is Bayes-admissible with respect to the distribution P and has constant risk, it is minimax. Also, in that case P is said to be the *worst-case* prior, since it is Nature's (mixed) strategy for achieving a maximin solution in the statistical game. Though the

details are beyond the presentation here, similar results apply to certain limits of Bayes-admissible solutions, obtained by considering a sequence of (proper) priors. (See Wald (1950 Theorem 3.17) for a result showing that every admissible decision rule either is Bayes-admissible for some distribution $\mathbf{P}$ or is the limit of a sequence of Bayes-admissible solutions.) The following illustrates the more general case.

Example: Given the parameter θ , let X have a Normal $N[\theta, 1]$ distribution with $\Theta = \mathcal{R}$. The statistician is required to provide a point-estimate for θ under a squared-error loss function, $L(d(x), \theta) = (d(x) - \theta)^2$. The intuitive rule $d(x) = x$ is admissible with constant risk $E_\theta[(x - \theta)^2] = \text{Variance}[X] = 1$. For the (conjugate) Normal prior $\mathbf{P}$, with $\theta \sim N[0, \sigma^2]$, a Bayes-admissible decision rule is $d(x) = x\sigma^2/(\sigma^2 + 1)$, with risk function $R(d(\cdot), \theta) = (\sigma^4 + \theta^2)/(\sigma^2 + 1)^2$ and Bayes-risk $r(d(\cdot), \mathbf{P}(\theta)) = \sigma^2/(\sigma^2 + 1)$. Consider a sequence of such conjugate priors, $\mathbf{P}_n$, where $\theta \sim N[0, \sigma_n^2]$, with $\lim_n \sigma_n^2 \rightarrow \infty$. The corresponding Bayes-admissible rules are just, $d_n(x) = x\sigma_n^2/(\sigma_n^2 + 1)$. The limiting Bayes-rule is the intuitive rule, $d(x) = x$, which is admissible and has constant risk; hence, it is minimax. From the statistician's point of view, the *worst case prior* is the 'improper' uniform distribution for θ : that is Nature's maximin strategy. (See Schervish and Seidenfeld (1996) for a minimax theorem that applies when the worst case prior is 'improper.') Thus, the intuitive decision rule in this problem, $d(x) = x$, illustrates a case where the game theoretic equilibrium and extended Bayesian expected utility theory (using limits of Bayes-admissible strategies) coincide.

In the preceding example, each of the decision rules $d_n(x)$ is Bayes-admissible with respect to the prior $\mathbf{P}_n$, and is an admissible decision rule as well. Thus, even in this heuristic example we see the tension between the two decision paradigms. The Bayesian statistician may have prior beliefs more accurately depicted by one of the (proper) conjugate priors $\mathbf{P}_n$, rather than by the limiting, worst-case prior. In such a situation, the Bayesian statistician will not be motivated by equilibrium considerations to adopt the game theoretic solution instead of maximizing expected utility with the decision rule $d_n(x)$. This is the tension explored in the next section.

2. Equilibrium Refinements and Maximizing Expected Utility: Rationalizable Strategies, Correlated Equilibrium, and Trembling Hands

What if each player in a game is a subjective expected utility maximizer, models the opponents that way, and these assumptions are common knowledge among the players? That is, though each player is uncertain about what choices opponents will make, nonetheless each player behaves as though he or she holds a (subjective) probability over those unknown states of affairs and maximizes expected utility against that probability

distribution. But, insofar as players think that their opponents also maximize expected utility, their probabilities must be supported by the set of their strategies that maximize their expected utilities. That all the players share these assumptions leads to a modest but interesting restriction on the set of Bayesian equilibria that result.

Bernheim (1984) and Pearce (1984) consider the important cases where the Bayesian players have probability models for each other's play where, regarding simultaneous play, acts and states are probabilistically independent. That is, assume that when the game requires players to make moves in ignorance of (some of) the moves of their opponents, then from the perspective of each player, the opponents' moves are probabilistically independent of choices at that time. The result is the set of *rationalizable* strategies, which carries the following elegant description. A strategy is *rationalizable* if and only if it survives the recursive operation of deleting successively options that are *strictly* dominated by another option not already deleted.

Aside: The reader is alerted to the fact that the recursion mentioned above iteratively changes the set of states of uncertainty for a player, i.e., the set of surviving options available to the opponents, that are relevant to a player when determining whether strict dominance obtains between two available strategies.

For an examination of how rationalizability differs from the traditional criterion of Nash equilibrium, particularly for sequential games in extensive form, see the articles by DeGroot and Kadane (1983) and Kadane and Seidenfeld (1992).

The set of rationalizable strategies in a game, as judged by the common standards of traditional Game Theory, typically makes an uncomfortably large collection of permissible moves, offering too little guidance about how the game should be played. An even more radical departure from the traditional point of view is proposed by Kadane and Larkey (1982) who examine the question of what considerations become salient to a Bayesian game-player who does *not* assume Bayesian-rationality on the part of opponents. They argue that the distinctions that are most important to traditional Game Theory are of only passing relevance under their radical reformulation of the same decision problems. For the practical side of this matter, see *Negotiations: Statistical Aspects*.

In the other direction, Aumann (1974, 1987) proposes a narrowing of the rationalizable strategies by introducing additional components of common knowledge among the players. Specifically, Aumann argues for a theory of rationality in which the Bayesian players share a common 'prior' over all the states of uncertainty that are relevant to choices made in the game. The upshot is a powerful criterion of *correlated equilibrium*, where players' *mixed strategies* form a best reply against each other, while relying on what are

in effect probabilistically *dependent* events. This allows for Pareto gains, where all players do better with a correlated equilibrium compared to their expected gains from a Nash equilibrium, when the mixed strategies of the players use jointly *independent* events only.

The relevant states of affairs over which the players have a common prior in Aumann's theory include the choices of all the players. This leads to the following controversy. Can Bayesian players assign nontrivial personal probabilities to their own (current) choices? Levi 1997 examines this intricacy and argues against the coherence of Aumann's condition.

Another interesting approach to reconciling equilibrium theory with expected utility analysis is found in the important works by Harsanyi (1977), and Harsanyi and Selten (1988). Harsanyi's doctrine is that Bayesian agents ought to use special distributions—oftentimes uniform probabilities—to reflect their ignorance about whatever states of uncertainty are relevant to their choices. (Note the connection with the *worst case* prior of the *Example* given in Section 1.3.) In the joint work with Selten, cited above, their novel proposal relies on the following idea: Though Bayesian players have definite preferences over strategy choices, they can achieve only a *trembling-hand* selection, where some residual probability is invariably assigned to other options besides the one most desired. In short, the players' options are modeled as nontrivial mixed strategies with some residual probability attached to even undesirable choices, and this is common knowledge. Trembling-hand equilibrium analysis examines limits of Nash equilibrium strategies as the trembles diminish. The results underwrite the criterion that choices in extensive form games should obey the principle of *subgame perfection*, i.e., that self-defeating threats are only idle threats. This criterion is particularly significant for application of the rule of *backward induction* in the theory of sequential choices, e.g., extensive form games when these are viewed from the Bayesian perspective.

3. Additional Sources

Luce and Raiffa's (1957) work remains a first-rate source of basic ideas and results contrasting the rival decision theories discussed here. Blackwell and Girshick's (1957) book is an excellent introduction to the interplay of ideas between Bayesian decision theory and traditional Game Theory relating to the design of statistical experiments. Binmore's stimulating textbook (1992) is a rich source of heuristic examples and informed discussion illuminating rival accounts of equilibrium refinements, going well beyond the few varieties alluded to here. Binmore is not shy, commenting openly on some of the more radical attempts at applying Bayesian decision theory to games. Even a casual reading of Binmore's textbook

will convince the reader of the truth of the opening sentences in this article!

See also: Game Theory; Game Theory: Non-cooperative Games

Bibliography

- Aumann R J 1974 Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1: 67–96
- Aumann R J 1987 Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* 55: 1–18
- Berger J O 1985 *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. Springer-Verlag, New York
- Bernheim B D 1984 Rationalizable strategic behavior. *Econometrica* 52: 1007–28
- Binmore K 1992 *Fun and Games: A Text on Game Theory*. D. C. Heath and Company, Lexington, MA
- Blackwell D, Girshick M A 1954 *Theory of Games and Statistical Decisions*. John Wiley, New York
- deFinetti B 1937 La prévision, ses lois logiques, ses sources subjectives. *Annals de l'Institut Henri Poincaré* 7: 1–68
- DeGroot M H, Kadane J B 1983 Optimal sequential decisions in problems involving more than one decision maker. In: Rustagi R H, Siegmund D (eds.) *Recent Advances in Statistics*. Academic Press, New York, pp. 197–210
- Good I J 1971 46,656 varieties of Bayesians. *American Statistician* 25: 62–3
- Harsanyi J C 1977 *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press, Cambridge, UK
- Harsanyi J C, Selten R 1988 *A General Theory of Equilibrium Selection in Games*. MIT Press, Cambridge, MA
- Kadane J B, Larkey P D 1982 Subjective probability and the theory of games. *Management Science* 28: 113–20
- Kadane J B, Seidenfeld T 1992 Equilibrium, common knowledge, and optimal sequential decisions. In: Bicchieri C, dalla Chiara M L (eds.) *Knowledge, Belief, and Strategic Interactions*. Cambridge University Press, Cambridge, UK, pp. 27–45
- Levi I 1997 Prediction, deliberation, and correlated equilibrium. In: Levi I (ed.) *The Covenant of Reason: Rationality and the Commitments of Thought*. Cambridge University Press, Cambridge, UK, Chap. 5
- Luce R D, Raiffa H 1957 *Games and Decisions*. Dover, New York
- Nau R 1999 *Arbitrage, Incomplete Models, and Interactive Rationality*. Technical Report. Fuqua School of Business, Duke University, Durham, NC
- Pearce D G 1984 Rationalizable strategic behavior and the problem of perfection. *Econometrica* 52: 1029–50
- Savage L J 1954 *The Foundations of Statistics*. John Wiley, New York
- Schervish M J 1995 *Theory of Statistics*. Springer-Verlag, New York
- Schervish M J, Seidenfeld T 1996 A fair minimax theorem for two-person (zero sum) games involving finitely additive strategies. In: Berry D, Chaloner C, Geweke J (eds.) *Bayesian Analysis in Statistics and Econometrics*. John Wiley, New York, pp. 557–68
- von Neumann J, Morgenstern O 1944 *The Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ

Wald A 1950 *Statistical Decision Functions*. John Wiley, New York

T. Seidenfeld

Game Theory: Noncooperative Games

Games are mathematical models of interactive strategic decision situations: there are various actors (players) involved that jointly determine the outcome and each tries to obtain that outcome that is most favorable to him. A game is said to be noncooperative if there are no possibilities for commitment (unilateral or multilateral) outside the rules of the game. In contrast, in cooperative games, players can form coalitions with the possibilities for doing so not being explicitly modeled within the rules of the game. Nash (1953) argued that the two approaches are complementary and he proposed building noncooperative models of cooperative games, an idea that is referred to as the Nash program. The article describes noncooperative game models and solution concepts and some applications are indicated (see Aumann 1987 for a more extended overview).

1. The Problem

Imagine an interactive decision situation in which various individuals (players) are involved, whose decisions affect each other. Depending on the decisions that are taken, a certain outcome will result; players have preferences over these outcomes and typically there is a conflict of interest as different players prefer different outcomes. Each player now faces the question what decision is best. As the decisions are interdependent, the answer to the question will depend not only on the player's own preferences, but also on the decisions that the other players take. The player, hence, has to make a prediction about what the other players will do. For an outsider there is the related question of predicting the overall outcome of the game. Based on the assumption of rational behavior of all players involved, game theory offers a set of tools and concepts that provide answers to these questions.

The above mentioned problem arises in a variety of circumstances, ranging from parlor games (chess, poker, bridge) to various economic, political, military, or biological situations. Game theory offers a collection of formal models and solution concepts to analyze these situations. While the history of the field can be traced back to the analysis of the two-person card game *le Her* by James Waldegrave in the early eighteenth century, with important contributions by the French scientists Cournot and Borel in the nineteenth and twentieth centuries (Baumol and Goldfeld (1968), Weintraub (1992)), John von Neumann gen-

erally is regarded as the founder of the field. In von Neumann (1928), he showed that the above mentioned problem allows a solution in the case of two players with strictly opposed interests (zero-sum games). The book *Theory of Games and Economic Behavior* that von Neumann wrote together with the economist Morgenstern (von Neumann and Morgenstern 1944) demonstrated convincingly that the tools developed initially for parlor games could be applied successfully to a variety of social conflict situations.

2. Noncooperative Game Models

The simplest type of model is one in which there is no dynamics: each player has to make only one decision and decisions are effectively made simultaneously as no player has knowledge about the decision that has been made by another player. As an example, one might think of a sealed bid procurement auction: without knowing the bids of the competitors, each player submits their bid in a sealed envelope; the player making the lowest bid is awarded the contract and in turn receives a payment (say) equal to their bid. Such a game is called a game in normal form or also a game in strategic form.

The formal mathematical definition of an n -player normal form game is as follows. Let S_i be the (finite) set of possible decisions (also called actions or strategies) that player i might take. Furthermore, let $S = \times_{i=1}^n S_i$ be the set of strategy profiles, hence, $s \in S$ specifies a strategy for each and every player in the game. Assume that each player's preferences over the set of outcomes of the game can be described by a (von Neumann and Morgenstern 1953) utility function, hence, each player wants to maximize their utility and only cares about expected utility. Von Neumann and Morgenstern (1953) give conditions under which such a utility function can be found; some theory has been developed also without this assumption, but it is convenient to maintain it here. Each strategy profile $s \in S$ produces a certain outcome and right $u_i(s)$ for the utility of player i associated with this outcome. The normal form game is then completely specified by the strategy sets S_i and by the utility functions u_i with $u_i: S \rightarrow \mathbb{R}$, hence $G = \langle S_1, \dots, S_n, u_1, \dots, u_n \rangle$.

Games in which players move more than once, and/or in which moves of different players are sequential can be represented by means of a tree. Formally, such a game is said to be in extensive form. The nodes of the tree correspond with decision points of players, with the arcs at a node representing the decisions that are possible at that node. Each endpoint of the tree represents an outcome of the game and again it is associated with a utility for each of the players. A special type of extensive form game is one in which there is perfect information: when a player has to move, they have full information about where they are in the game and what decisions have been taken

before. Chess can be viewed as an extensive form game with perfect information. So-called information sets are used to represent limited information in a tree: when a player cannot distinguish two nodes in the tree, these are in the same information set. When a player has to move, they can condition their action only on the information set that is revealed to them, not on the actual node. The theory has been developed mainly for games with perfect recall, that is, each player is assumed to fully remember their previous actions and information. Kuhn (1953) provides a formal development. For some recent contributions dealing with imperfect recall, see Rubinstein (1998).

Von Neumann (1928) introduced the fundamental concept of strategy by means of which an extensive form game can be reduced to one in normal form. Formally, a strategy for a player is a full plan of action that specifies for each information set of this player the action that this player intends to take if and when that information set is reached. Clearly, once a strategy has been determined for each player, the outcome is determined as well. Von Neumann argued that, for rational players, there is no loss of generality in forcing them to think through the game in advance, hence, to force them to choose a strategy. Consequently, the extensive form can be reduced to the simpler normal form.

The above description abstracted from randomization. A player, however, might randomize before the start of the game which (pure) strategy they might use during the game. Alternatively, they might randomize to determine which action they will choose once a certain information set is reached. The concept of mixed strategy is used to refer to the former case, the concept of behavior strategy refers to the latter case of local randomization. Kuhn (1953) showed that both concepts are equivalent for games with perfect recall, that is, whatever a player can do with a strategy of one type they can also do with a strategy of the other, and vice versa. Note that since players are assumed to have von Neumann–Morgenstern utilities, replacing ‘payoff’ (utility) by ‘expected payoff’ creates no special difficulties.

The traditional theory has been developed for games of complete information, that is, each player is assumed to know the utility functions of all the players in the game. Frequently, a player will not have all information about their competitors. For example, in the procurement context referred to above, a player may not know how many orders a competitor has in stock, hence, they may not know how much value an opponent assigns to winning the present contract. Harsanyi (1968) showed how to incorporate incomplete information into the model. To represent asymmetric information, Harsanyi introduces an artificial chance move at the beginning of the game that determines which piece of private information each player will have. This piece of private information is also called the player’s type. It is assumed that the set

of possible types and the probability distribution over this set is common knowledge, however, which type realization results for player i is typically only known to player i himself. A normal form game with incomplete information, hence, does not only specify the players and their utility functions, but also their types and the joint probability distribution on players’ types. Formally, such a game is given by a tuple $G = \langle S_1, \dots, S_n, T_1, \dots, T_n, u_1, \dots, u_n, p \rangle$ where S_i is player i ’s strategy set, T_i is the set of possible types of player, i , $u_i: S \times T \rightarrow \mathbb{R}$ is player i ’s payoff function (where $T = \prod_{i=1}^n T_i$), and p is the probability distribution on T . The play of the game proceeds as follows: A type profile $t \in T$ is determined according to p and player i is informed about their type t_i . Based on this information, player i updates information about the types of the others, computing the posterior probability $p_i(t/t_i)$ and chooses an action to maximize the associated expected payoff. Harsanyi’s modeling technique has shown to be very powerful (see *Auctions; Information, Economics of*).

3. Solution Concepts

3.1 Individual Rationality

Concepts of ‘individual rationality’ from decision theory can also be used in a game context. If one strategy s_i of player i is (strictly) dominated by another strategy s'_i of this player (i.e., it yields a lower payoff no matter what strategy profile s_{-i} the other players use, formally $u_i(s_i, s_{-i}) < u_i(s'_i, s_{-i})$ for all s_{-i}) then a rational player i will not use s_i and it should be possible to eliminate s_i without changing the solution. Furthermore, dominated strategies can be iteratively eliminated. Alternatively, a ‘Bayesian’ player will represent their uncertainty about their opponent’s strategies by a probability distribution σ_{-i} on S_{-i} ($= \prod_{j \neq i} S_j$) and they will thus only choose strategies that are best responses against some such correlated strategy σ_{-i} . By using duality theory from linear programming it can be shown that s_i is a best response against some correlated strategy σ_{-i} of the opponents if and only if s_i is not strictly dominated, hence, the two concepts are equivalent. In most of the theory it has been assumed that the fact that players decide independently implies that a player will represent their uncertainty by a profile of mixed strategies of the opponents, that is, the components of σ_{-i} are independent. The set of strategy profiles that remain after all strategies of all players that are not best responses against independent mixtures of opponents have been iteratively eliminated is called the set of rationalizable profiles (Bernheim 1984, Pearce 1984). The set of rationalizable strategies may be smaller than the set of iteratively undominated strategies.

In his theory of the 2-person zero-sum game, von Neumann also used an individualistic rationality concept. If player i uses mixed strategy σ_i then the worst that can happen to i is to receive the expected payoff $\min_{\sigma_{-i}} u_i(\sigma_i, \sigma_{-i})$, hence, player i can guarantee the payoff $\max_{\sigma_i} \min_{\sigma_{-i}} u_i(\sigma_i, \sigma_{-i})$ by choosing their mixed strategy appropriately. Note that a mixed strategy may 'guarantee' a higher expected payoff than a pure strategy. For example, in 'matching pennies' (two players simultaneously show the side of their penny with player 1 (resp. player 2) receiving both pennies if the sides are the same (resp. different) it is obviously better to keep one's side hidden, hence, to use a mixed strategy. Von Neumann showed that, for finite 2-person zero-sum games, there exist number v_1 and v_2 with $v_1 + v_2 = 0$ such that each player i can guarantee v_i . We say that is the security level of player i : against a 'rational' opponent, player i cannot expect to get more than this level. Strategies that guarantee the security level are called optimal strategies and v_1 is also termed the value of the game. Von Neumann's theorem is also called the minmax theorem, it thus shows that there is indeed an optimal way to play a finite 2-person zero-sum game.

3.2 Equilibrium

Nash (1950, 1951) was the first to break away from the individualistic rationality concepts. Rather than focusing on the direct question 'what should a player do' (or 'what should a game theorist advise a player to do'), he approached the problem indirectly and asked 'if the game has a solution, what properties would it have.' In fact, a suggestion to proceed in this way was already contained in von Neumann and Morgenstern (1953), but the authors of that book had not followed it through. Nash assumes that a satisfactory solution of a game exists. He further assumes that the solution is unique. Hence, the solution takes the form of a (mixed) strategy profile σ , recommending to each player i a course of behavior σ_i . Rational players will know this solution and a player will be willing to behave according to it only if it is in their interest to do so, that is, if the strategy σ_i is a best response against the strategy profile of the opponents, $u_i(\sigma) = \max_{\tau_i} u_i(\tau_i, \sigma_{-i})$. A Nash equilibrium is a profile of strategies σ such that this condition is satisfied for each player i . To summarize the argument, a necessary condition for σ to be a satisfactory solution of the game is that σ be a Nash equilibrium.

By relying on fixed point theorems (either the one of Brouwer or the one of Kakutani can be used), Nash proved that every finite game admits at least one Nash equilibrium, albeit possibly in mixed strategies. Indeed, in 'matching pennies' each player should choose each side of the coin with 50 percent probability in the unique equilibrium. In a 2-person zero-sum game, σ is

a Nash equilibrium if and only if it is a pair of optimal strategies, hence, Nash's concept generalizes the solution provided by von Neumann and Morgenstern.

Pursuing the above rationale leading to Nash's solution concept, three questions remain: (a) why does a player represent their uncertainty about the behavior of their opponents by an independent mixed strategy profile? (b) do the requirements determine a unique solution? and (c) is being a Nash equilibrium sufficient to qualify as a satisfactory solution? The first issue is taken up in Aumann (1974) where an alternative concept of correlated equilibrium is developed that allows for more general beliefs. One interpretation is that players discuss before the game and construct a common randomization device, a correlated strategy σ . Upon hearing what pure strategy the device selected for player i , this player updates their beliefs on their opponent's actions; σ is a correlated equilibrium if each player is always willing to follow the recommendation of σ . The answer to the second question is a clear no, it is easy to construct games with multiple Nash equilibria and many games, also some with practical relevance, do have multiple equilibria. Hence, the rationale that has been given for the equilibrium concept appears incomplete, and the question is, whether, in a game with multiple equilibria there is any argument for focusing on any of these. Alternatively, should one look in an entirely different direction for a satisfactory solution, for example, by giving up the assumption that the solution be unique? We discuss these questions below after first having taken up the third question and having shown that that one also has a negative answer: not every Nash equilibrium can be considered as a satisfactory solution.

Consider the following extensive form ultimatum game. Player 1 divides \$10 in any (integer) way that they want; player 2 observes the division and decides whether or not to accept it; if player 2 accepts, each player receives the amount that was proposed, otherwise neither player receives anything. In addition, assume that each player cares only about the amount of money that they receive and that each prefers more money to less, say $u_i(x) = x_i$. Player 1 then knows that player 2 is sure to accept as long as they are offered at least \$1, hence, there seem only two possible solutions of the game—10; accept all—and—9, accept iff $x_2 \geq 1$ (the first number is the amount that player 1 asks for himself). Indeed these two strategy profiles are Nash equilibria, however, there are other equilibria as well, for example, 3, accept iff $x_2 \geq 7$ is a Nash equilibrium. (If player 1 demands 3 for themselves, then player 2's strategy prescribes to accept and that is a best response; on the other hand, if player 2 indeed accepts only if offered at least 7, then it is optimal for player 1 to offer exactly 7.) In this equilibrium player 2 'threatens' to reject a positive amount that is offered to, such as \$5, even though strictly preferring to accept in case \$5 would be offered. The reason the profile is in

equilibrium is that an amount such as \$5 is not offered and, hence, the threat is not called. Player 1 behaves as if player 2 is committed to this strategy, in the extensive form game, however, such commitments are not possible and facing the *fait accompli* that \$5 is offered, player 2 will accept. Player 1 can neglect the incredible threat.

Starting with Selten (1965) a literature developed dealing with the question of how to eliminate equilibria involving incredible threats. Selten (1965) proposed the concept of subgame perfect equilibrium that is based on the idea of 'persistent rationality': everywhere in the game tree, no matter what happened in the past, the player that moves will play a strategy that, from that point on, is a best response against the strategies of the opponents. Formally, a subgame perfect equilibrium is a strategy profile that constitutes a Nash equilibrium in every subgame of the game, a subgame being a part of the game that constitutes a game in itself. For games with perfect information, subgame perfect equilibria can be found by the backward induction (dynamic programming) procedure that was used already in Zermelo (1913) to show that the game of chess is determined completely: starting at the end of the game one work backwards each time reducing the game by substituting a decision set of a player by an optimal decision and its associated payoff. It is worthwhile to remark that this assumption of 'persistent rationality' was criticized in von Neumann and Morgenstern (1944).

Selten (1975) noted that requiring subgame perfection was not sufficient to eliminate all equilibria involving incredible threats and proposed a different solution. He reasoned that, if the problem is caused by nonmaximizing behavior at unreached information sets, it can be eliminated by ensuring that all information sets are reached, if only with small probability. Formally, Selten (1975) defines an equilibrium to be (trembling hand) perfect if it satisfies a robustness test: each player should be still willing to play their equilibrium strategy if each player with some small probability makes mistakes and, as a consequence, every information set is reached with positive probability. In the ultimatum game above, if player 1 makes mistakes, player 2 is forced to accept any positive amount and player 1 will not offer more than \$1 to player 2 in a perfect equilibrium. The robustness test proposed is weak (the strategies are only required to be optimal for some sequence of small mistakes that converges to zero) and this guarantees that each game admits at least one perfect equilibrium. However, at the same time this implies that some equilibria are perfect only because of the fact that the mistake sequence can be artificially chosen, hence, the perfectness concept does not eliminate all equilibria that are considered as 'counterintuitive.' The same remark applies to the closely related concept of sequential equilibria that has been proposed in Kreps and Wilson (1982). This concept requires that, at each information

set, a player constructs beliefs over where they are in the set and optimizes their payoff, given these beliefs and taking the strategies of the opponents as fixed. Kreps and Wilson require that the beliefs be explainable in terms of small mistakes, hence, the close connection between their concept and Selten's.

Intuitively, one would like an equilibrium not to be robust against just some trembles but rather against 'plausible' ones: if there is a good explanation for why the information set is reached after all, then preference should be given to such an explanation. Plausibility is hard to define formally, but, if an equilibrium is robust against all trembles, then clearly it will also be robust against the plausible ones. Unfortunately, equilibria satisfying this strict robustness test need not exist. However, Kohlberg and Mertens (1984) have shown that sets of equilibria with this desirable property exist always. They, furthermore, argue that it is natural to look at sets of 'equivalent' equilibria. For example, if a player is fully indifferent between two of their strategies, there is no reason to make them choose a specific one. Similarly, if the choice only makes a difference if and after another player has played a strictly dominated strategy, one might be satisfied to leave the choice undetermined.

Sets of stable equilibria satisfy several desirable properties, such as being robust against elimination of strategies that are dominated or that are not a best response against any element of the set. The latter property is also called 'forward induction' and it has proved powerful to reduce the number of equilibrium outcomes in signaling games (see *Information, Economics of*). The original stability concept proposed in Kohlberg and Mertens (1984) was of a preliminary nature and was not completely satisfactory. The ideas have been further developed in Mertens (1989) but that concept has not seen too many applications yet as it is difficult to handle. In addition, part of the profession considers this a too demanding rationality concept (see van Damme 1991 for further discussion).

The normative, 'rationalistic' interpretation of Nash equilibrium that has been considered thus far relies on the assumption that the solution to a game is unique, yet many games admit multiple equilibria, even multiple stable ones. The question thus arises about equilibrium selection: how will players coordinate on an equilibrium and on which one? Consider the following simple normal form stag hunt game as an example. Two players simultaneously choose between the numbers 1 and 2. If both choose the same number, they each receive a payoff equal to the number chosen, if they choose different numbers the player choosing the lower number receives 1 from the other. The game has two equilibria in pure strategies: (1,1) and (2,2). The latter yields both players higher payoffs than the former, however, choosing 2 is also more risky than choosing 1: while the latter guarantees the payoff 1, the former might result in a loss of 1. How to resolve the problem?

Harsanyi and Selten (1988) provide a general theory of equilibrium selection in noncooperative games that can be used in these circumstances. The theory can be seen as an extension of Nash's (1953) equilibrium selection theory for bargaining games. The authors start by formulating general properties that a theory should satisfy (such as symmetry, ordinality, efficiency, monotonicity, consistency) but they discover quickly that not all of these can be satisfied at the same time. Indeed, the stag hunt game given in the above paragraph has the same best reply structure as the game in which payoffs are 2 (resp. 1) if each player chooses 1 (resp. 2) and where payoffs are 0 if different numbers are chosen, and in this game the efficient equilibrium involves both players choosing 1. Hence, choices have to be made and, as Harsanyi and Selten admit, different selection theories are possible. One of the major innovations of Harsanyi and Selten (that in some form will probably play a role in these alternative theories as well) is the concept of risk dominance. Intuitively, one equilibrium is said to risk dominate another if, in a situation where attention is confined to this pair, players eventually come to coordinate on the first as they consider the second to be more risky. The formal definition makes use of the tracing procedure, an algorithm (homotopy) that transforms any mixed strategy profile (which represents the players' initial assessment about how the game will be played) into an equilibrium of the game. This tracing procedure is supposed to capture the reasoning process of rational players in the game. If, in a situation of uncertainty about equilibria σ and σ' , the tracing procedure produces equilibrium σ , then σ is said to risk dominate σ' . The concept allows a simple characterization for 2-player 2×2 games with 2 strict equilibria. Let $p_i(\sigma, \sigma')$ be the probability that player i has to assign to σ_i in order to make their opponent j indifferent between σ_j and σ'_j . Intuitively, the larger $p_i(\sigma, \sigma')$, the more attractive σ_j is for player j . In this case, σ risk dominates σ' if and only if $p_1(\sigma, \sigma') + p_2(\sigma, \sigma') < 1$. In the above example, the equilibrium (1,1) risk dominates the equilibrium (2,2).

3.3 Evolution and Learning

The above discussion of Nash equilibrium (and its variants) has been based entirely on the rationalistic interpretation of this concept, viz. that a self-enforcing theory advising players how to play should prescribe a Nash equilibrium in every game. Nash (1950), however, provided also a second interpretation of his concept. When the game is played repeatedly by myopic players who best respond against the current situation, then, if a stable state is reached, that state has to be a Nash equilibrium. In this second interpretation it is thus unnecessary to assume that players know the full structure of the game, or are able to perform complex reasoning processes. On the

contrary, the interpretation applies in a context of limited information where players behave in a boundedly rational way. After Shapley (1964) had published an example showing that a certain dynamic adjustment process need not converge at all, interest in this second approach dwindled for some time, but it was revived in the 1990s. Since then a systematic study has begun under which circumstances a variety of processes will converge and, if they converge, to which equilibria. It is too early to draw broad conclusions, instead the reader is referred to Fudenberg and Levine (1998) for a partial overview. Interestingly, some of the concepts that were introduced in the 'rationalistic' branch reappear in this branch of the literature, in particular this holds for the concept of risk dominance.

Consider the stag hunt coordination game described above and imagine that it is played repeatedly. When a player is called upon to play they 'ask around' to find out what other players in the population that have already played have been confronted with and then plays a best response against the resulting sample. Young (1998) shows that the process will converge to one of the two equilibria. Now imagine that sometimes players make mistakes in implementing their desired action and choose the other one instead. In this case the process may move from one equilibrium to the other through a sequence of consecutive mistakes. Young (1998), however, shows that, if the probability of making mistakes tends to zero, the limit distribution will put all mass on the risk dominant equilibrium as that equilibrium is more robust to mistakes than the other one. Hence, in the long run when mistake probabilities are negligible the process will end up in the risk dominant equilibrium, irrespective where it starts out. The process not only produces an equilibrium, it leads to a very specific one. Research effort is under way to investigate the robustness of this result with respect to both the underlying game and the dynamic process under consideration.

Related to the above is the application of the Nash equilibrium concept in biology. As Maynard Smith (1982, p. vii) writes 'Paradoxically, it has turned out that game theory is more readily applied to biology than to the field of economic behavior for which it was originally designed.' In biological applications the concept of utility is replaced by that of Darwinian fitness (expected number of offspring), human rationality is replaced by evolutionary stability and, according to Maynard Smith, both contribute to making the theory more applicable. Note that in this context it is assumed that there is no conscious choice of strategy at all; strategies exist and individuals are 'programmed' to play certain strategies, an evolutionary process selects those strategies that have highest fitness and in the end an equilibrium results.

To illustrate, one may consider the so called Hawk-Dove game in which two animals are contesting a certain valuable resource. There is a (finite) repertoire of possible behaviors, S , and when an individual

programmed to play strategy s meets another individual playing strategy s' the gain in fitness to the first is given by $u(s, s')$. The situation may then be represented by a game in which there are 2 players, and each has strategy set S and payoff function u , that is, the game is symmetric. The question is which strategy will be selected, that is, which one will survive in the long run. Obviously, if $u(s', s) > u(s, s)$ then s' can displace s so that a necessary condition for survival of s is that (s, s) constitutes a Nash equilibrium of the game. A second necessary condition for selection of s is that, once s has been established, random mutations that enter into the population are selected against, that is, these are driven out again. This will be the case if they cannot spread, that is, if the entrant s' is an alternative best response against s , then the incumbent strategy obtains a higher payoff against the mutant than the mutant gets against itself (if $u(s', s) = u(s, s)$, then $u(s, s') > u(s', s')$). A strategy s satisfying both conditions is called an evolutionarily stable strategy or ESS. One sees that an ESS corresponds to a Nash equilibrium satisfying an additional stability requirement (see Hammerstein and Selten 1994 for further discussion).

4. Behavior

Game theory, at least its traditional variant, relies on the assumption of rational behavior, that is, behavior that is consistent and is motivated solely by the desire to obtain well-defined goals, with, furthermore, no cognitive constraints standing in the way. In contrast, the biological branch of the theory relies on no rationality at all, but assumes long time periods for selection to work. The question, hence, is how the theory is relevant for humans that are boundedly rational and have relatively little time to learn and adjust behavior. Real players have only limited rationality: they face cognitive limitations which may make it impossible to get a full overview of the situation and to consistently evaluate and rank all alternatives, they may decide not to think and instead rely on automated procedures that have proved satisfactory in the past or in other related contexts, and even if rational deliberation suggests a decision, emotional factors may override and produce an alternative decision.

The question, hence, is to what extent a theory that is based on strong rationality assumptions is relevant for actual interactive decision making. For sure, a normative solution can serve as a benchmark to better understand actual behavior, it might point to systematic differences and it might even be that the differences are negligible in certain contexts, such as when players have had sufficient time to familiarize themselves with the situation. Unfortunately, a discussion of the issue of relevance is made difficult by the fact that alternative theories are still undeveloped and since relatively little is known yet about the actual

structure of human decision making (see Selten 1991). Nevertheless, some systematic deviations from the theory that are found in the experimental laboratory may be described (see Camerer 1997 for more details).

First of all, a remark that does not apply that much to the solution concepts, but rather to the modeling aspect. Game theory analyses the consequences of selfish behavior, where selfishness is interpreted as the individual following their preferences. However, it is not assumed necessarily that individuals are selfish in the more narrow sense of being only interested in their own material well-being. The latter assumption would indeed be too narrow. Experiments have shown that, in the ultimatum game referred to above, responding players generally dislike being treated 'unfairly,' hence, they are willing to reject proposals in which they get positive amounts but the distribution is uneven. A relevant game model of the situation should take this into account and as a consequence it may well advise to player 1 to allocate a substantial proportion of the cake to player 2 (see Kagel and Roth 1995 for further discussion on this issue). Suffice it to note here that tests of game theoretic predictions are always combined tests of the underlying game and the solution concept applied, that it is difficult to control the players' preferences, hence, that one should be careful in drawing strong inferences from results observed in the experimental laboratory.

Second, when confronted with an interactive decision situation, real-life players typically construct a simplified model of the situation. The 'mental model' may be an oversimplification and, as a consequence, important strategic aspects may be missed. Furthermore, a player needs to take into account the models constructed by other players: if these do not incorporate certain elements, it does not make sense for the player to signal along these dimensions. Selten (1998) argues that the reliance on superficial analysis may explain presentation or framing effects, that is, the way in which a game is presented to the players may have a strong influence on actual behavior. The basic reason is that the presentation may trigger a certain reasoning process or may make some outcomes more focal than others. Related is the fact that human players do not analyze each game in isolation but rather make use of analogies; a principle that has proved useful in a certain context may also be used in another one. In experiments, for example, it is observed that 2-player games with sequential moves in which the second mover receives no information about the first move may be played as if there was such information, where according to the theory the game is strategically equivalent to the game with simultaneous moves. The discussion of the presentation effect in strategic contexts originates with Schelling (1963), but his concept of 'focal points' still largely awaits formalization.

Where standard game theory assumes that all players are equally rational and, hence, homogeneous, experiments have revealed considerable player het-

erogeneity. Some players are more of the adaptive type and do not approach a game analytically. Instead, in repeated contexts, they make use of *ex post* rationality considerations and move in the direction of better responses against the previous outcome. Players who rely on analytic approaches construct simplified models, as discussed above, and they may limit the 'depth' of their analysis to a couple of steps and, hence, may not reach a game theoretic equilibrium. For example, in the 'beauty contest game' in which a set of players choose numbers between 1 and 100 and in which the winner is the one that chooses the number closest to half of the average of all numbers, the equilibrium is to choose 1 and that equilibrium is reached through iterative elimination of dominated strategies. However, most players choose larger numbers and those that choose the equilibrium number do not win the game. Relatedly, in extensive form games players do not necessarily 'look ahead and reason backwards,' but instead may use a myopic somewhat forward looking process and, as a consequence, the outcome obtained may be different from the subgame perfect equilibrium one.

At present, research is under way to construct an empirically-based behavioral game theory that steers a middle ground between over-rational equilibrium analyses and under-rational adaptive analyses, but it is too early to survey this field (see *Behavioral Economics*; *Experimental Economics*).

5. Conclusion

The space allotted has not allowed a discussion of any applications in depth, however, as Aumann (1987) has argued forcefully, the models and solution concepts from game theory should be judged by the insights that they yield. In addition to the entries of this Encyclopedia already referred to above, we may point to the areas of finance (market microstructure), industrial organization, and antitrust and regulation, in which ideas from noncooperative game theory have been influential and are used widely.

Bibliography

- Aumann R J 1974 Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* **1**: 67–96
- Aumann R J 1987 Game theory. In: Eatwell J, Milgate M, Newman P (eds.) *The New Palgrave Dictionary of Economics*. Macmillan, London, pp. 460–82
- Aumann R J 1985 What is game theory trying to accomplish? In: Arrow K, Honkapohja S (eds.) *Frontiers of Economics*. Blacksburg, VA, pp. 28–100
- Baumol W J, Goldfeld S 1968 *Precursors in Mathematical Economics: An Anthology*. London School of Economics and Political Science, London

- Bernheim B D 1984 Rationalizable strategic behavior. *Econometrica* **52**: 1007–28
- Camerer C 1997 Progress in behavioral game theory. *Journal of Economic Perspectives* **11**: 167–88
- Damme E E C van 1991 *Stability and Perfection of Nash Equilibria*, 2nd edn. Springer Verlag, Berlin
- Fudenberg D, Levine D 1998 *The Theory of Learning in Games*. MIT, Cambridge, MA
- Hammerstein P, Selten R 1994 Game theory and evolutionary biology. In: Aumann R J, Hart S (eds.) *Handbook of Game Theory*, Vol. II. Elsevier, Amsterdam
- Harsanyi J 1967–1968 Games with incomplete information played by 'Bayesian' players, parts I, II and III. *Management Science* **14**: 159–82, 320–34, 486–502
- Harsanyi J, Selten R 1988 *A General Theory of Equilibrium Selection in Games*. MIT, Cambridge, MA
- Kagel J, Roth A 1995 *Handbook of Experimental Economics*. Princeton University Press, Princeton, NJ
- Kohlberg E, Mertens J F 1986 On the strategic stability of equilibria. *Econometrica* **54**: 1003–39
- Kreps D, Wilson R 1982 Sequential equilibria. *Econometrica* **50**: 863–94
- Kuhn H 1953 Extensive games and the problem of information. In: Kuhn H, Tucker A W (eds.) *Contributions to the Theory of Games II*. Princeton University Press, Princeton, NJ, pp. 193–216
- Maynard Smith J 1982 *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, UK
- Mertens J F 1989 Stable equilibrium: A reformation. *Mathematics of Operations Research* **14**: 575–625
- Nash J (1950) Non-cooperative games. Ph.D. thesis, Princeton University
- Nash J 1951 Non-cooperative games. *Annals of Mathematics* **54**: 298–95
- Nash J 1953 Two-person cooperative games. *Econometrica* **21**: 128–40
- Neumann J von 1928 Zur theorie der gesellschaftsspielen. *Mathematische Annalen* **100**: 295–320
- Neumann J von, Morgenstern O 1944/1953 *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ
- Pearce D G 1984 Rationalizable strategic behavior and the problem of perfection. *Econometrica* **52**: 1029–50
- Rubinstein A 1998 *Modeling Bounded Rationality*. MIT, Cambridge, MA
- Schelling T 1963 *The Strategy of Conflict*. Harvard University Press, Cambridge, MA
- Selten R 1965 Spieltheoretische behandlung eines oligopolmodells mit nachfrageträgheit. *Zeitschrift für die gesamte Staatswissenschaft* **121**: 301–24, 667–89
- Selten R 1975 Re-examination of the perfectness concept for extensive form games. *International Journal of Game Theory* **4**: 25–55
- Selten R 1991 Evolution, learning and economic behavior. *Games and Economic Behavior* **3**: 3–24
- Selten R 1998 Features of experimentally observed bounded rationality. *European Economic Review* **42**: 413–36
- Shapley L 1953 Some topics in 2-person games. In: Kuhn H, Tucker A (eds.) *Contributions to the Theory of Games II*. Princeton University Press, Princeton, NJ, pp. 305–17
- Weintraub E R 1992 *Toward a History of Game Theory*. Duke University Press, Durham, NC
- Young P 1998 *Individual Strategy and Social Structure*. Princeton University Press, Princeton, NJ

Zermelo E 1913 Über eine anwendung der mengenlehre auf die theorie des schachspiel. In: *Proc. 5th Int. Congress of Mathematicians*. Vol. II, pp. 501–4

E. van Damme

Gangs, Sociology of

The term, ‘gang’ is both a theoretical construct and an object of varying definitions in legal statutes, legal and social agency policy, and common discourse. Although the term defies precise measurement, it is the subject of a large scholarly and popular literature.

1. Historical Background of Scholarly Research on Gangs

‘Gang’ and other terms (mob, syndicate, outfit, etc.) have been applied to many types of groups, including those associated with organized and professional crime and incarcerated felons. Scholarly attention, however, has focused primarily on youth gangs. Frederic Thrasher’s *The Gang: A Study of 1,313 Gangs in Chicago* (1927) was the first attempt to survey the extent of youth gang activity in a major city, and perhaps the first in any jurisdiction. The project involved census and court records, personal observations, and personal documents collected from gang boys and from persons who had observed gangs in many contexts. Studies by Clifford R. Shaw and his collaborators provided even more information about the collective nature of youthful delinquency, offering consistent documentation that most boys who were brought before the juvenile court committed their delinquent acts in the company of others. Case studies documented patterns of friendship, the association of younger with older offenders, and the influence of organized crime and other forms of adult criminality in communities with high rates of juvenile delinquency (Shaw et al. 1929, Shaw 1930, Shaw and McKay 1931, Shaw and Moore 1931). Ecological studies located delinquency in space and in relationship to urban development, documenting economic and institutional contexts within which urban lives were lived, and identifying forces that shaped the ability of communities to aid in the socialization of children and exercise control over misbehavior.

Theoretical insights from these studies continue to be important in thinking about youth crime. Fore-shadowing ‘labeling theory,’ for example, Shaw’s Jack-Roller—reflecting on his confinement in a reformatory—noted that he was no longer ‘just a mischievous lad, a poor city waif, a petty thief, a habitual runaway,’ but ‘a criminal’ (Shaw 1930, p. 103). Conversely, both Shaw and Thrasher noted that incarceration and criminal notoriety often had positive value among gang members, as it does today: ‘... res-

idents in the vicinity south of the stock yards were startled one morning by a number of placards bearing the inscription ‘The Murderers, 10,000 Strong, 48th and Ada.’ In this way attention was attracted to a gang of thirty Polish boys, who hang out in a district known as the Bush’ (Thrasher 1927, pp. 62–3).

Thrasher’s Murderers were involved in a good deal of criminal activity, but their primary pastimes ‘were loafing, smoking, chewing, crap-shooting, card-playing, pool, and bowling.’ Their ‘rudeness and thievery’ were ‘an awful nuisance’ to local shop keepers and neighbors (Thrasher 1927, pp. 62–3). The arcane language notwithstanding, these excerpts highlight similar patterns among contemporary and earlier gangs involving ‘hanging out,’ minor and more serious criminal behavior.

The most important impact of earlier gang research was its contribution to social disorganization theory. Youth gangs were found overwhelmingly in communities with high rates of crime and delinquency, poverty, and population heterogeneity and turnover. Social disorganization theory, hypothesizing that such communities lack both effective institutions and local informal means of control, continues to be refined by recent research. Thus, Sampson and his colleagues (2000) find that intergenerational closure (the linkage between adults and children), reciprocal local exchange (interfamily and adult interaction with respect to children), and expectations that community residents will share child control responsibilities, are associated with the ability of local communities to exert effective control over violent and other forms of criminal behavior.

Midway between the publication of Thrasher’s classic work and the turn of the century, the book was abridged and reissued. Social changes had vastly altered the gang landscape, and gang research, heretofore largely descriptive, was changing in response to seminal theoretical proposals by Cohen (1955), Cloward and Ohlin (1960), and Miller (1958). Competing explanations for the origins of the delinquent subculture (Cohen), variations in delinquent subcultures (Cloward and Ohlin), and the role of lower-class culture in producing gang delinquency (Miller) stimulated a large body of empirical research and subsequent theorizing.

By the last decade of the twentieth century, however, events that could hardly have been anticipated by earlier researchers had overtaken both theory and research. Whereas midcentury theories had outstripped available data, modest theoretical advances and related empirical research were overwhelmed by the rapid proliferation of gangs in the USA and their spread to many other countries (Klein 1995, Klein et al. 2000, Moore and Terrett 1999). In the USA sophisticated firearms became more readily available to young people (Blumstein 1995), often turning what previously had been ‘non-zero-sum’ contests between gangs into lethal confrontations (Short and

Strodtbeck 1965, Anderson 1999). Weakened conventional economic, political, and family institutions lost legitimacy, contributing to rising crime rates (LaFree 1998); and the technological revolution in electronic communication and entertainment media created heretofore unknown problems for institutions of social control.

2. *Change and Continuity*

It is possible, nevertheless, to note both changes and continuities in knowledge concerning the collective forms taken by the delinquent and criminal behavior engaged in by the young.

(a) Youth gangs are more prevalent in many countries than previously. Law enforcement agencies in all large US cities (over 250,000) reported the presence of youth gangs, as did many suburban counties, small cities and rural counties, where gang prevalence is quite recent (National Youth Gang Center 1998). European and UK researchers have identified variations in youth gang organization and behavior, some of them patterned after US gangs (Klein et al. 2000).

(b) Gangs in the US, Continental Europe, and the UK are comprised predominantly of youth who are marginalized in mainstream society by virtue of race, ethnicity, or economic status. In the USA these are predominantly African-American and Hispanic, but with increasing numbers of new Asian immigrant children and children of immigrant parents. In Continental Europe and the UK they tend to be the children of refugees from other countries or migrant workers. Gangs comprised of majority population youth, nearly always white, sometimes form in opposition to these marginalized groups, thus becoming marginalized themselves.

Youth gangs are comprised overwhelmingly of young males, and the experience of females who associate with male-dominated gangs, or in all-female gangs, is quite different from the gang experience of males (Chesney-Lind and Hagedorn 1999).

(c) The distribution and patterning of collective youth crime reflect the ecology of the places where they are located. In the USA and many other cities, these are inner-city areas, much as they were in Thrasher's and Shaw's day. Where public housing projects are located outside central cities, gangs of loosely structured, ethnically mixed groups often 'commute' to central city areas where they engage in carousing, 'smash-and-grab' hits on stores, rolling drunks, and robberies.

Large aggregations of gang 'nations' emerged in a few large US cities, most notably Chicago and Los Angeles, during the 1960s. Local gangs—wherever they are found—are primarily 'home grown,' but the influence of gang nations is widespread through mimicry and migration of gang members. Many—

probably most, systematic data are lacking—gangs begin as nondelinquent, friendship groups, 'hanging' groups, or groups formed around particular non-delinquent activities (sports, singing, dancing, etc.) (see, e.g., Schwendinger and Schwendinger 1985, Hagedorn 1988, Sullivan 1989, Schwartz 1987, Short 1997). The primary distinguishing characteristic of all youth gangs is that they are self-determining with respect to membership criteria, group structure and behavior, and—most importantly—they are unsupervised by adults.

(d) A variety of studies suggest that such unsupervised youth groups become violent, and otherwise delinquent, as a result of one or more of the following processes:

(i) escalation of the natural rough-and-tumble punching and wrestling that occurs among most male groups, and the association of status with fighting prowess;

(ii) competition with rival groups involving status-enhancing activities or economic motivations;

(iii) conflicts based on racial/ethnic differences;

(iv) definitions of groups as gangs, by others, and behavior toward them that pushes a violent identity on the group;

(v) group processes that create or reinforce group cohesion based on violent or other delinquent behavior.

Street gangs are diverse in behavior, organization, continuity across generations, and in their relationships with other gangs and with their communities. They change, as members age and move into adulthood. The age range of many gangs has expanded in many countries, in response to the lack of attractive legitimate jobs. This change often is accompanied by increased violence, as gang members have turned to selling drugs and other illegal activities in order to support themselves, and often their families (see Fagan 1996, Fagan and Wilkinson 1998, Hagedorn 1998).

These changes tend to blur distinctions between street gangs and a variety of types of youth and young adult collectivities, such as 'wilding' groups, 'skin-heads,' and perhaps the mass shootings that have occurred on school properties. Although differing among themselves, these groups have in common their hate of 'outsiders,' based on racial, ethnic, or other exclusionary criteria, as, also, do crowds that gather to attack community outsiders, usually members of minority groups. Violence and other delinquent activities often occur, also, in 'staging' areas in which young people gather with friends and others simply in order to escape adult supervision (Schwartz 1987, Anderson 1999).

Drug and other 'working' gangs are organized primarily around revenue-producing activities, such as drug selling, stealing, fencing stolen goods—all activities requiring greater organizational skills and structure than is characteristic of traditional street gangs. Also unlike street gangs, 'bikers,' 'punks,'

and sports fanatics and 'hooligans' are organized around their specialized interests. Except for the latter (most of whom appear to be disadvantaged socially, rather than economically), these groups are comprised primarily of youth who have been marginalized by minority ethnic and/or economic status, and who have poor prospects for integration or social mobility within dominant majority societies. The excessive behavior of bikers and sports hooligans also appears to be motivated in some measure by resentment of minorities.

3. *Youth Culture*

Virtually all youth are heavily influenced by the development, rapid growth, and diffusion of youth culture. Peddled by media advertising, and augmented by large social and economic forces that fail to provide meaningful roles for young people but cater to their distinctive fads, appetites, and currency, youth culture and gang culture together underly the excesses of consumption and violence perpetrated by gangs.

Youth culture developed rapidly during the 1960s, especially in the USA. Young people born during the 'baby boom' following World War II shared in the general level of affluence in the country, thereby creating a youth market with great economic power; the period of education prior to entering the labor market was extended; increasing numbers of women joined the workforce, further separating mothers from youth in the home and in the neighborhood; increasingly, adults worked in large organizations away from young people; and the mass media greatly expanded and focused their attention on the youth market (see Coleman et al. 1974, pp. 114–19). These changes accelerated as the century wore on, expanding their influence throughout the world.

Field researchers document the pernicious seductions of media-advertised products among less affluent young people, leading often to thefts and assaults. Anderson (1999) describes the 'zero-sum quality' that pervades the 'code of the streets,' in the search for respect—often associated with items of clothing such as shoes and jackets—among young African-American men in the ghetto. Sullivan identifies 'mutually valorizing cultural symbols' (clothing, drugs, alcohol, and recreational artifacts and activities) that cut across racial and ethnic boundaries, often resulting in invidious comparison, confrontation, theft, and assault (Sullivan 1989, pp. 248–9).

Observers of youth groups everywhere note that status differences within and between them are extremely variable and highly refined. Status criteria within a group, or between one group and another, may be based on race or ethnicity, relative economic affluence, skills in valued activities, public appearance, school performance, or most importantly, on lifestyle differences. These bases of stratification typically also

become criteria of inclusiveness and exclusiveness which, in turn, create opportunities for both friendship and rejection.

Adolescence is a period of especially intense identity formation, intense relationships, and shared feelings of friendship, acceptance, and respect. Conversely, feelings of rejection and disrespect also are especially intense, and often the basis for group and subcultural formation.

Next to families, schools are perhaps the most important contexts for adolescent friendship, achievement, and recognition. Because of this, school contexts (including the journey to and from schools) are the settings for much adolescent behavior, including delinquent and criminal behavior. On occasion they have also been the setting for the most extreme forms of violence (e.g., mass killings by students or others alienated from their fellows or from mainstream institutions in general). Although the specific causes of such extreme alienation are complex, it seems clear that schools have been especially targeted precisely because of their importance in the lives of adolescents, as symbols of rejection by both peers and an adult world that seems far removed from adolescent concerns. The ready availability of guns at times transforms normal adolescent turmoil and conflict into deadly confrontation.

More than ever before in history, young people, targeted for commercial exploitation and isolated from mainstream adult roles and institutions, confront economic conditions beyond their control. Economic decline, severe unemployment, and the unavailability of 'good jobs,' are associated not only with street gangs, but with their transformation into 'economic gangs' (including drug gangs), and with ethnic, racial, and class-related identities and antagonisms that lead to other types of collective violence (Hagedorn 1998 Pitts 2000). These same forces alter both intergang relationships and relationships between gangs and their communities. Relationships between adults and young people, shaped by generational backgrounds and community cultures, 'translate' these and other macrolevel forces at the local community level.

4. *Research Needs*

Research and theory concerning gangs, and effective means of controlling the violence and crime associated with gangs, are hampered by the lack of a viable typology of youth groups and groups comprised of young adults that would situate the many observed variations of gangs and gang-like groups within even broader collective forms of behavior. Comparative research in many countries suggests that distinctions among and between such groups vary a great deal, and that the meaning and the function of behavioral variations lack universality.

Varying definitions, meanings, and behaviors have consequences. To the extent that gangs and gang membership confer 'master status' on groups and individuals—among law enforcement personnel, in media portrayals, and among the general public—they contribute to the deterioration of relationships between economic, racial, and ethnic groupings and between the generations. Reification and demonization of gangs, by media, law enforcement, the public, and occasionally by scholars, for example, may contribute to gang problems rather than to their solution.

See also: Control: Social; Crime, Sociology of; Delinquency, Sociology of; Norms; Violence: Public; Youth Gangs

Bibliography

- Anderson E 1999 *Code of the Streets: Decency, Violence, and the Moral Life of the Inner City*. W.W. Norton, New York
- Blumstein A 1995 Violence by young people: Why the deadly nexus? *National Institute of Justice Journal* August: 2–9
- Chesney-Lind M, Hagedorn M (eds.) 1999 *Female Gangs in America*. Lake View Press, Chicago
- Cloward R A, Ohlin L E 1960 *Delinquency and Opportunity*. Free Press, New York
- Cohen A K 1955 *Delinquent Boys*. Free Press, New York
- Coleman J S, Bremner R H, Clark B R, Davis J B, Eichorn D H, Griliches Z, Kett J F, Ryder N B, Doering Z B, Mays J M 1974 *Youth: Transition to Adulthood. Report of the Panel of the President's Science Advisory Committee*. University of Chicago Press, Chicago
- Fagan J 1996 Gangs, drugs, and community change. In: Huff C R (ed.) *Gangs in America*, 2nd edn. Sage, Thousand Oaks, CA, pp. 39–74
- Fagan J, Wilkinson D L 1998 Guns, youth violence, and social identity in inner cities. In: Tonry M, Moore M (eds.) *Youth Violence. Crime and Justice*. University of Chicago Press, Chicago, Vol. 24, pp. 105–88
- Hagedorn J M 1988 *People and Folks: Gangs, Crime, and the Underclass in a Rustbelt City*. Lake View Press, Chicago
- Hagedorn J M 1998 Gang violence in the postindustrial era. In: Tonry M, Moore M (eds.) *Youth Violence. Crime and Justice*. University of Chicago Press, Chicago, Vol. 24, pp. 365–419
- Klein M 1995 *The American Street Gang*. Oxford University Press, UK
- Klein M, Kerner H-J, Maxson C L, Weitekamp E G M (eds.) 2000 *The Eurogang Paradox: Street Gangs and Youth Groups in the US and Europe*. Kluwer-Plenum, Amsterdam
- LaFree G 1998 *Losing Legitimacy: Street Crime and the Decline of Social Institutions in America*. Westview Press, Boulder, CO
- Miller W B 1958 Lower class culture as a generating milieu of gang delinquency. *Journal of Social Issues* 23(3): 5–19
- Moore J P, Terrett C P 1999 *Highlights of the 1997 National Youth Gang Survey*. OJJDP Fact Sheet, No. 97. Office of Juvenile Justice and Delinquency Prevention, Washington, DC
- National Youth Gang Center 1998 *1996 National Youth Gang Survey*. Office of Juvenile Justice and Delinquency Prevention, Washington, DC
- Pitts J 2000 *Discipline or Solidarity: The New Politics of Youth Crime*. Macmillan, Basingstoke, UK
- Sampson R J, Morenoff J D, Earls F 2000 Beyond social capital: Structural sources and spatial embeddedness of collective efficacy for children. *American Sociological Review* 64: 633–60
- Schwartz G 1987 *Beyond Conformity or Rebellion: Youth and Authority in America*. University of Chicago Press, Chicago
- Schwendinger H, Schwendinger J 1985 *Adolescent Subcultures and Delinquency*. Praeger, New York
- Shaw C R 1930 *The Jack-Roller: A Delinquent Boy's Own Story*. University of Chicago Press, Chicago
- Shaw C R, McKay H D 1931 *Report on the Causes of Crime*. National Commission on Law Observance and Enforcement Report, Vol. 2. USGPO, Washington, DC
- Shaw C R, McKay H D [1942] 1969 *Juvenile Delinquency and Urban Areas*. University of Chicago Press, Chicago
- Shaw C R, Moore M E 1931 *The Natural History of a Delinquent Career*. University of Chicago Press, Chicago
- Shaw C R, Zorbaugh F, McKay H D, Cottrell L S 1929 *Delinquency Areas*. University of Chicago Press, Chicago
- Short J F Jr 1997 *Poverty, Ethnicity, and Violent Crime*. Westview Press, Boulder, CO
- Short J F Jr, Strodtbeck F L 1965 *Group Process and Gang Delinquency*. University of Chicago Press, Chicago
- Sullivan M 1989 *'Getting Paid': Youth Crime and Work in the Inner City*. Cornell University Press, Ithaca, NY
- Thrasher F M 1927 *The Gang: A Study of 1,313 Gangs in Chicago*. University of Chicago Press, Chicago

J. F. Short Jr.

Garbage Can Model of Behavior

The garbage can (GC) model is a framework for analyzing decision making in 'organized anarchies'—organizations characterized by problematic preferences, unclear technologies, and fluid participation (Cohen et al. 1972). The model conceives of organizations as conglomerates of semiautonomous decision arenas or 'garbage cans' through which problems and solutions flow as independent streams. Decision outcomes are sensitive to the precise mix of problems and solutions represented in a garbage can at the moment of decision. This mix, in turn, depends on the number of decision arenas, the structure of access to them, the overall organizational load of problems and solutions, and the allocation of energy and attention across arenas.

1. A Challenge to the Rational Model of Organization

The GC model may be understood as an important step in a developing critique of the rational model of organization by the so-called 'Carnegie School.' This school builds on Herbert Simon's treatment of organizations as information-processing systems and his argument about cognitive limits on individual

information-processing (see *Bounded Rationality*). Simon (1985) argued that just as individuals could partially overcome their cognitive limits by decomposing complex decisions into simpler choices and routines, so too could organizations decompose complex collective decision making by means of nested hierarchical structures ('nearly decomposable systems'). The subsystems of these nested hierarchies represented relatively autonomous bundles of routines. While Simon's analysis offers a critique of the rational model, he ultimately seeks to demonstrate that problem decomposition and routinization permit cognitively limited individuals and organizations to approximate, if only imperfectly, rational organization. The GC model, in contrast, poses a more fundamental challenge to the rational model.

The GC model goes beyond Simon's analysis by focusing on the ambiguity and loosely coupled nature of individual and organizational decision making. Whereas Simon's model of bounded rationality emphasizes cognitive limits on evaluating alternative courses of action, the GC model argues that the ambiguity of decision making stems from the ill-defined and inconsistent nature of individual and organizational goals (problematic preferences), and the indeterminacy of knowledge and of techniques for achieving goals (unclear technologies). The GC model also problematizes Simon's notion of organizations as bundles of semiautonomous subsystems. The metaphor of the organization as a collection of 'garbage cans' implies a set of semiautonomous decision-making arenas where problems are neither so neatly decomposed nor so well integrated as conceptualized by Simon. Furthermore, the GC model argues that participation in these decision arenas is unstable rather than routine (fluid participation). In making these arguments, the GC model incorporates and extends many of the earlier arguments of the Carnegie school, including insights into the political and symbolic character of decision making, the coalitional character of organization, the sequential nature of decision making, the role of organizational slack, and the tendency of subunits to develop narrow and parochial decision premises.

In seeking to turn the rational model 'on its head,' the authors of the garbage can model sometimes state their alternative assumptions in fanciful ways. For example, they argue that organized anarchies can be understood as 'collections of choices looking for problems, issues and feelings looking for decision situations in which they might be aired, solutions looking for issues to which they might answer, and decisionmakers looking for work' (Cohen et al. 1972, p. 1). This ironic tone may be read as either usefully provocative or as a hindrance to understanding. Stated in the extreme, the assumptions of the model often appear as unrealistic as the assumptions of the rational model. The argument that problems are decoupled from solutions and flow through the organization as

independent streams is difficult to grasp, because it is intuitive to think that the people who formulate problems are typically the ones best positioned to devise solutions to them. Yet if we were to dismiss the GC model because of its implausibility *in extremis*, we would also miss its powerful insights. Shorn of its irony, it is easy to recognize situations in which people become advocates for certain types of solutions and promote them whenever and wherever they can, often without clearly understanding the problems which are being addressed. It is easy to see that people often bring issues and concerns to a decision arena extraneous to the ostensible decision at hand. And it is obvious that decision outcomes are highly sensitive to the mix of issues brought to bear in any decision arena, that this mix is often highly dependent on who participates, and on the range of other decisions being made at any given time. One value of the GC model is that it treats these very intuitive and commonplace circumstances seriously.

2. The Simulation of Chaotic Decision Making

Part of the purpose of stating their assumptions in exaggerated form is that it allows the authors (brilliantly) to transform their analysis of the essentially *ad hoc* nature organizational behavior into a systemic analysis of organizational decision making. They do this by linking together the following ideas. As already mentioned, the model emphasizes the *loose coupling* or *decoupling* of intentions, actions, and constraints. It argues that individual intentions are not tightly coupled to individual action, that collective goals are not tightly coupled to collective action, and that organizations are not tightly coupled to their environments. By conceptualizing problems and solutions as flowing independently through organizations, the GC model imagines organizations to be *demographic systems* in which problems and solutions (via their attachment to individuals) enter and exit particular decision-making arenas. Decisions or nondecisions can then be understood as dependent upon the composition of problems and solutions in a decision arena at any point in time. At the systemic level, this composition is expected to be very sensitive to the issues to which actors happen to be paying attention at any given moment (attention is treated as a scarce resource) and to the structure of access into decision-making arenas. The model is *processual* in that it suggests that outcomes are not easily predicted from starting conditions, but depend rather heavily on the precise interactions between elements of the system. In general, processual theories suggest that interactions are complex, nonlinear, and characterized by the prevalence of unintended consequences. Taken together, this characterization of organized anarchies as loosely coupled, demographic, and processual allows the authors to treat decision making as a *stochastic*

process, introducing elements of probability and randomness into the model. Beginning with the original paper of Cohen et al. (1972), the GC model has encouraged a strategy of simulation modeling in order to capture the dynamics of these stochastic, demographic processes (e.g., Padgett 1980, Carley in March and Weissinger-Baylon 1986).

3. *The Garbage Can Model and Political Institutionalism*

If organizational decision making is often disorderly and fraught with ambiguity, we rarely experience the extreme versions of chaos the GC model may conjure. Yet the model also suggests how disorder and ambiguity may be transformed into relative order and certainty. It is very possible, for instance, that disorder and ambiguity may be tidied up on the output side of decision-making processes: order and certainty then represent *post hoc* rationalizations of essentially *ad hoc* decisions. Yet order and certainty may also be produced during decision making through a process known in current parlance as *social construction*. Ambiguity is transformed into relative certainty by social processes in which individuals develop and institutionalize shared understandings and judgments. March and Olsen elaborate on this constructivist mechanism in their book, *Rediscovering Institutions* (1989), where they situate garbage can theory in a larger theoretical framework of political institutionalism (see *Institutionalism*). Arguing that politics can be understood as the collective interpretation and enactment of meaning, they suggest that institutionalization leads individuals and organizations to act according to a 'logic of appropriateness.' March and Olsen (1989, pp. 69–116) provide a nice example of how garbage can processes, political institutionalism, and social constructivism can complement one another in their historical analysis of the failure of comprehensive administrative reforms.

The GC literature also suggests at least two other mechanisms through which anarchies organize: leadership and learning. With respect to leadership, for instance, this literature suggests that unobtrusive leadership which embraces rather than rejects the disorder and ambiguity of decision making may be successful in steering organized anarchies toward collective purposes (Cohen and March 1974, Kingdon 1995, Padgett 1980). Learning in organized anarchies is problematic because the lessons we should draw from experience are themselves ambiguous. While learning under conditions of ambiguity may lead to superstitious beliefs, the GC literature suggests that organized anarchies may adapt and learn successfully (Cohen and March 1974, March and Olsen 1979). In fact, it is currently fashionable to see a bit of chaos as necessary for producing creativity and innovation. Hood (1999) notes the interesting parallels between

the garbage can model and Mintzberg's model of 'adhocracy,' which is widely promoted as an organizational design for promoting creativity and flexibility.

The GC model has been applied in numerous contexts. As March and Olsen have noted, the theory probably applies to many organizations some of the time and some organizations most of the time. It is often seen as a good fit for educational institutions, especially universities. Sproull et al. (1978) argue that the characteristics of organized anarchies characterize 'most new government organizations, irrespective of their mission, and most agencies operating in social policy domains, irrespective of their age' (Sproull et al. 1978, p. 5). In addition to their book on the National Institute of Education, other studies of political organization inspired by the GC model include analyses of presidential management (Padgett 1980), military organizations and operations (March and Weissinger-Baylon 1986), public-health management in the context of an 'interorganizational garbage can' (Clarke 1989), policy making and agenda setting in the federal government (Kingdon 1995), and foreign policy decision making (Newmann 1998). While these differ in their details, the GC model should also be seen as similar in spirit to Graham Allison's model of bureaucratic politics, to Charles Lindbloom's model of disjointed incrementalism (see *Decision-making Systems: Personal and Collective*), to Norton Long's notion of politics as an 'ecology of games,' and to Jeffrey Pressman and Aaron Wildavsky's theory of implementation (see *Implementation: Political*). Taken together, these models have made a major contribution to our understanding of the limitations of the rational model of decision making and to the possibility of alternative rationalities.

See also: Bounded and Costly Rationality; Decision and Choice: Economic Psychology; Organizational Behavior, Psychology of; Organizations, Sociology of; Rational Choice and Organization Theory; Rational Choice in Politics; Rational Choice Theory in Sociology

Bibliography

- Clarke L 1989 *Acceptable Risk? Making Decisions in a Toxic Environment*. University of California Press, Berkeley, CA
- Cohen M D, March J G 1974 *Leadership and Ambiguity: The American College President*. McGraw-Hill, New York
- Cohen M D, March J G, Olsen J P 1972 A garbage can model of organizational choice. *Administrative Science Quarterly* 17: 1–25
- Hood C 1999 The garbage can model of organization: Describing a condition or a prescriptive design principle. In: Egeberg M, Lægreid P (eds.) *Organizing Political Institutions: Essays for Johan P. Olsen*. Scandinavian University Press, Oslo, Norway, pp. 59–78
- Kingdon J W 1995 *Agendas, Alternatives, and Public Policies*, 2nd edn. HarperCollins, New York

- March J G, Olsen J P 1979 *Ambiguity and Choice in Organizations*, 2nd edn. Universitetsforlaget, Bergen, Norway
- March J G, Olsen J P 1989 *Rediscovering Institutions: The Organizational Basis of Politics*. Free Press, New York
- March J G, Weissinger-Baylon R 1986 *Ambiguity and Command: Organizational Perspectives on Military Decisionmaking*. Pitman, Boston
- Newmann W W 1998 Foreign policy decision making, garbage cans, and policy shifts: The Eisenhower administration and the 'Chances for Peace' speech. *American Review of Public Administration* **28**: 187–212
- Padgett J F 1980 Managing garbage can hierarchies. *Administrative Science Quarterly* **25**: 583–604
- Simon H 1985 *The Sciences of the Artificial*. MIT Press, Cambridge, MA
- Sproull L, Weiner S, Wolf D 1978 *Organizing an Anarchy: Belief, Bureaucracy, and Politics in the National Institute of Education*. University of Chicago Press, Chicago

C. K. Ansell

Gastrointestinal Diseases: Psychosocial Aspects

Research in the area of gastrointestinal diseases has focused primarily on biological and medical aspects of these disorders. Little consideration has been given to psychosocial aspects of these medical conditions. However, there has been one general area, namely irritable bowel syndrome (IBS) that has received increased attention in the area of psychosocial aspects over the past few decades. More specifically, IBS has been increasingly conceptualized using a biopsychosocial perspective (Drossman et al. 2000). This perspective provides a framework for understanding the biological and psychosocial components of IBS, and can provide a model for other areas within the digestive system.

The purpose of this article is to offer an understanding of the psychosocial aspects of gastrointestinal illness. Because it is beyond the scope of this article to supply an extensive and detailed discussion of all the specific diseases of the digestive system, we have limited our discussion to psychosocial aspects of IBS since this is where most of the theoretical and empirical literature has focused. The first section of this article will furnish an overview of IBS. The second section will address how a biopsychosocial perspective relates to IBS. The third section will summarize current psychosocial treatments for IBS. The final section will briefly address future directions.

1. Definition and Symptoms of IBS

Functional gastrointestinal disorders (FGID) are characterized by recurrent or current gastrointestinal symptoms that have no identifiable structural or

biochemical basis. The most common functional gastrointestinal disorder is IBS. Recent evidence supports the belief that dysregulation of the brain–gut neuroenteric systems, rather than the presence of structural disease, promotes the development and persistence of IBS (Drossman et al. 2000). IBS is associated with abdominal pain or discomfort for at least 12 weeks or more, that has two or three of the following features: (a) relieved with defecation; and/or (b) associated with a change in frequency of stool; and/or (c) associated with a change in form (appearance) of stool. Symptoms that are associated with IBS include: abnormal stool frequency (more than three bowel movements per day and less than three bowel movements per week), abnormal stool form (hard or loose/watery stool), abnormal stool passage (straining or urgency, feeling of incomplete evacuation), passage of mucus, bloating, or feeling of abdominal distension (Thompson et al. 2000).

2. Epidemiology

IBS is a very common disorder, estimated to affect 9–22 percent of the Western population. However, of those affected, very few seek medical consultation or treatment for their gastrointestinal symptoms (Whitehead et al. 1988). According to Drossman et al. (1997), of those who seek treatment, IBS accounts for 28 percent of gastroenterological practice in Western societies and 12 percent of primary care. IBS is often associated with unnecessary procedures and surgeries. IBS has considerable economic impact since IBS results in decreased work capacity. Next to the common cold, it ranks as the second most common cause of work absenteeism.

IBS is a disorder that is diagnosed mostly in women (Toner et al. 2000). To date, most of our information about IBS has been drawn from female participants. Accordingly, to date, we know little about IBS in men. A review of the literature by Toner and Akman (2000) noted that the majority of studies investigating IBS use only women in their samples. Moreover, of those studies which do include men, few perform a gender difference analysis. Since a major percentage of IBS patients are women, the issue of gender must be integrated into the conceptualization and treatment of these disorders (Toner 1994).

3. Conceptualization

IBS is best conceptualized using a biopsychosocial model. As summarized by Drossman et al. (1999), the biopsychosocial model provides a useful framework to help us begin to understand the connections among psychosocial factors, gastrointestinal function, disease susceptibility, and clinical outcome. Symptoms may be generated from physiological disturbances (enhanced motility and visceral sensation), which are closely

connected to central nervous system activity (via the central nervous system–enteric nervous system axis). The clinical expression of these symptoms (e.g., the decision to take medication or seek health care) is strongly influenced by psychosocial factors. For this reason, the high frequency of psychosocial factors (e.g., psychiatric diagnoses, sexual abuse history) in the absence of modulating factors (e.g., social support, coping strategies) reported among IBS patients may, in part, relate to their self-selection into referral practices. Accordingly, while psychosocial factors are not part of the diagnosis of IBS, they have important effects on gut motility, symptom experience, and clinical outcome.

In light of this information, it is no longer logical that clinicians and researchers try to assess whether IBS symptoms are caused by physiological or psychosocial factors since both are operative and interactive. Rather, the goal is to determine to what extent these factors contribute to IBS and are remediable (Drossman et al. 1999).

4. Psychosocial Aspects Discussed in the Context of Myths Associated with IBS

It is difficult to discuss psychosocial aspects without acknowledging the stigma and myths that have been associated with IBS. Functional gastrointestinal disorders such as IBS are associated with morally pejorative connotations. Patients who have functional gastrointestinal disorders often feel that their problems are treated as ‘not real’ and due to a moral and/or psychological weakness or defect. Thus, when referred to health care providers, persons with IBS may believe that caregivers think that their symptoms are ‘all in their heads’ and not ‘real’ or ‘serious.’ It is important that health care professionals validate the reality of the person’s symptoms. In doing so, health care professionals can challenge the artificial dualism between functional and organic components of illness in general and pain in particular imposed by our Western society (Toner et al. 2000). Dualism leads to a prominent myth which suggests that if pain is severe, there must be an organic cause. This is a common myth that persists as a function of our Western society’s false conceptualization of pain (i.e., if the pain is severe, there must be a structural cause). However, the experience of pain is a consequence of a complex interaction among physical, cognitive, emotional, and behavioral components.

A related myth is that IBS is caused by stress. We now know that IBS is not caused by stress, but that stress may further aggravate bowel symptoms. We also know that stress is associated with bowel symptoms in most people. In fact, over 70 percent of people under stress will exhibit some GI symptoms. However, studies have found that people with IBS are more reactive since they already have a hypersensitive gut

and therefore are more likely to experience more symptoms under any kind of stress including psychological, physical, dietary, or hormonal.

An additional myth is that IBS is a psychiatric disorder or that it is masked depression. Several studies have found that a substantial percentage of individuals presenting with IBS, particularly in tertiary care settings, also have an associated depressive or anxiety disorder. It is unclear why such an association exists, but the following explanations have been proposed. (a) People who also have an associated anxiety/depressive disorder may have more difficulty coping with gastrointestinal problems and may seek specialized help for their pain at higher rates than people with IBS who do not have an associated anxiety or depressive disorder. (b) Anxiety and depression are common in the general population, and their co-occurrence may simply be due to high frequency. (c) Depression/anxiety may be a consequence of living with a chronic, incapacitating disorder such as IBS (Toner et al. 2000).

A final myth is that people with IBS are sometimes described as difficult patients. Physicians and other health care providers are not adequately trained in the conceptualization and/or treatment of IBS. The lack of information coupled with the shame and trivialization associated with having a so-called functional or psychosomatic disorder leads to frustration, further distressing patients. For physicians, such patients may illicit uncomfortable feelings that might lead some health care professionals to feel powerless and frustrated in the situation. Rather than conceptualizing IBS patients as difficult, it is more helpful to conceptualize the disorder as difficult, especially in light of the fact that there is little information and a great deal of stigma associated with IBS (Toner et al. 2000). An open and honest discussion acknowledging the frustration associated with this chronic and debilitating disorder is beneficial. To date, there are no effective medication or cure for IBS and thus the focus of treatment is symptom management and improving coping strategies. As such, it is important to establish a collaborative environment whereby both the person with a diagnosis of IBS and health care provider work together in attempting to understand the most effective strategies in managing these chronic, painful symptoms.

5. Empirical Support for Psychosocial Treatments

Psychosocial treatments for IBS have included dynamic psychotherapy, cognitive-behavioral therapy, relaxation, biofeedback, and hypnosis. In a recent overview of this area, a multinational team including Drossman et al. (2000) reviewed 13 controlled studies to compare psychosocial treatment with conventional medical treatment. Ten of 13 studies showed signifi-

cant superiority of psychosocial over conventional medical treatment in terms of reducing bowel symptoms. Follow-up data were available for only nine studies and of these, eight showed psychosocial treatment to be superior to medical treatment.

However, it is difficult to determine the efficacy of a specific theoretical approach because most of the controlled studies have used multicomponent treatment packages (e.g., cognitive-behavioral, psychodynamic, biofeedback, and relaxation approaches). As such, it is difficult to assess which aspects of the treatments were beneficial and to what extent. Moreover, therapeutic techniques and procedures are not generally described adequately and in specific detail. This makes replications of treatment protocols difficult. Nevertheless, the data thus far seem to indicate that when compared to conventional medical treatment, psychosocial interventions are consistently more effective.

6. Summary and Future Directions

This article has provided a brief overview of IBS with a focus on associated psychosocial aspects. While psychosocial factors are not part of the diagnosis of IBS, clearly, psychosocial factors play an important role in symptom expression and clinical outcome of IBS. Given the significance of these factors, it is important that clinicians respond to psychosocial issues with the same attention and legitimacy that is given to the physical and biological factors associated with gastrointestinal disorders. Conceptualizing IBS from a biopsychosocial perspective allows clinicians to better understand and treat individuals who are diagnosed with IBS in a more comprehensive manner. Future research must examine how biopsychosocial factors influence other diseases of the digestive system.

Research has suggested that psychosocial treatments are effective interventions for IBS. Further research is needed to determine the usefulness of psychosocial interventions for other diseases of the digestive system. Moreover, future studies are required to determine which aspects of psychosocial treatment packages (e.g., relaxation, cognitive restructuring, etc.) account for their effectiveness and determine the patient characteristics which predict response to specific psychosocial treatments. Since this is a chronic disorder, long-term studies must also be conducted. The role of sociocultural factors and gender on IBS must be further explored in both our understanding and treatment of these chronic debilitating disorders.

See also: Behavioral Medicine; Biopsychology and Health; Chronic Illness, Psychosocial Coping with; Chronic Illness: Quality of Life; Chronic Pain: Models and Treatment Approaches; Emotions and Health; Illness Behavior and Care Seeking; Pain, Health Psychology of; Pain, Management of; Pain, Neural

Basis of; Psychosomatic Medicine; Stress and Coping Theories; Stress and Health Research; Stress Management Programs; Stress, Neural Basis of; Stress: Psychological Perspectives; Symptom Awareness and Interpretation

Bibliography

- Drossman D A, Whitehead W E, Camilleri M 1997 Medical position statement: Irritable Bowel Syndrome. *Gastroenterology* **112**: 2118–9
- Drossman D A, Creed F H, Olden K W, Svedlund J, Toner B B, Whitehead W E 1999 Psychosocial aspects of the functional gastrointestinal disorders. *Gut* **45**(Suppl. 2): 1125–30
- Drossman D A, Creed F H, Olden K W, Toner B B, Whitehead W E 2000 Psychosocial aspects of the functional gastrointestinal disorders. In: Drossman D A, Corazziari E, Talley N J, Thompson W G, Whitehead W (eds.) *The Functional Gastrointestinal Disorders*, 2nd edn. Allen Press, Lawrence, KS, pp. 157–245
- Thompson W G, Longstreth G, Drossman D A, Heaton K, Irvine E J, Muller-Lissner S 2000 Functional bowel disorders and functional abdominal pain. In: Drossman D A, Corazziari E, Talley N J, Thompson W G, Whitehead W (eds.) *The Functional Gastrointestinal Disorders*, 2nd edn. Allen Press, Lawrence, KS, pp. 360–1
- Toner B B 1994 Cognitive-behavioral treatment of functional somatic syndromes: Integrating gender issues. *Cognitive and Behavioral Practice* **1**: 157–78
- Toner B B, Akman D 2000 Gender role in Irritable Bowel Syndrome: Literature review and hypothesis. *American Journal of Gastroenterology* **95**(1): 11–6
- Toner B B, Segal Z V, Emmott S D, Myran D 2000 *Cognitive-Behavioral Treatment of Irritable Bowel Syndrome: The Brain-Gut Connection*. Guilford, New York
- Whitehead W E, Bosmajian L, Zonderman A B, Costa P T Jr, Schuster M M 1988 Symptoms of psychologic distress associated with Irritable Bowel Syndrome: Comparison of community and medical clinic samples. *Gastroenterology* **95**: 709–14

B. Toner

Gauss, Carl Friedrich (1777–1855)

1. Introduction

The name of Gauss will already be known to many readers for its use in such phrases as Gaussian distribution, Gauss linear model, and the Gauss-Markov theorem, and for its use in the titles of the Gaussian elimination, Gauss-Jordan, Gauss-Doolittle, Gauss-Newton, and Gauss-Seidel procedures, if not for the *Système International* (SI) unit of magnetic induction and the associated terms gaussage and degauss.

Gauss contributed to all areas of science known in his day, notably to astronomy, physics, and geodesy,

but in this article we shall concentrate on his contributions to statistics and numerical methods. For more wide-ranging surveys of his work, see Eisenhart (1968) and May (1972).

2. *Life of Gauss*

Carl Friedrich Gauss (originally Johann Friedrich Carl Gauss) was born on 30 April, 1777 in Brunswick (Germany). He was the son of a bricklayer and gardener who also kept the accounts of a local insurance company. Gauss received his elementary education at St. Katherine's School in Brunswick. However, his humble origins were not allowed to hamper his intellectual development, for, in 1791, his talent for mathematics was brought to the attention of the Duke of Brunswick, who undertook to support his continuing education, at first at the *Collegium Carolinum* in Brunswick and later at the University of Göttingen. Gauss did not graduate from the University of Göttingen but was awarded a doctoral degree *in absentia* by the University of Helmstedt in 1799 for his work on the fundamental theorem of algebra.

In October 1806, the French land forces consolidated the command of German territory that they had established in 1805 in the battle of Austerlitz by defeating a Prussian army under the command of the Duke of Brunswick at Auerstedt (and mortally wounding the Duke). As a consequence, when in 1807 Gauss submitted the manuscript of his future *Theoria Motus* for publication in German, the political situation was such that his publisher felt obliged to insist on its translation into Latin. Consequently, the manuscript which Gauss had completed in 1806 was not published until 1809. (Fortunately, Gauss had an excellent command of Latin, so that this imposition may justly be compared with requiring a modern German scholar to translate his work into English before publication.)

In 1807 Gauss was appointed director of the University of Göttingen observatory and professor of mathematics at the University of Göttingen. He continued to hold both of these positions until his death in Göttingen on 23 February, 1855.

3. *Elementary Fitting Procedures*

Given any number of observations on a lesser number of variables, the linear fitting problem is concerned with establishing the relationship between one of these variables and the remaining variables. Before the middle of the eighteenth century, practitioners were obliged to employ a variety of more or less arbitrary procedures to determine the parameters of such relationships. But between 1750 and 1800 four algebraic procedures were proposed for the linear fitting

problem, namely the method of averages, the method of selected points, the method of least absolute deviations, and the minimax absolute deviation procedure.

In 1805 Adrien-Marie Legendre (1752–1833) published an alternative algebraic procedure which he named the method of least squares. This procedure chooses the parameters of the relationship to minimize the sum of the squared deviations between the observed values and the fitted relationship. For a detailed history of these five algebraic procedures, see Farebrother (1999). Also note that the relevant pages of Legendre's original statement have been reproduced by Stigler (1986, p. 56).

Legendre's approach to the fitting problem was entirely algebraic and had no statistical content. Further, he did not prove many of the optimality properties he claimed for the method of least squares. Indeed, he had not even shown that the values determined by his algebraic procedure actually minimize the sum of the squared deviations function (except in the unlikely case when the fitted relationship is exact). Such a proof was subsequently supplied by Gauss (1809).

4. *Most Probable Value Criterion*

Gauss' first statistical derivation of the method of least squares was based on a combination of the normal or Gaussian distribution with a Bayesian variant of the maximum likelihood or maximum probability optimality criterion. The normal distribution itself was obtained from a preliminary analysis in which the arithmetic mean of a set of direct observations is associated with the maximum value of this optimality criterion.

Gauss was working in a Bayesian context, so there was no need for him to estimate the disturbance variance, but he employed a simple algebraic procedure that we would now associate with the *LDL'* decomposition of a symmetric matrix to obtain measures of relative precision of each of the parameter estimate.

5. *Priority Disputes*

In his 1809 book, Gauss referred to the method of least squares as 'our principle, which we have made use of since the year 1795' and thus sparked off a serious priority dispute with Legendre that ran on for several years. Although Legendre had certainly been the first to publish a description of the method, it is now generally agreed that Gauss had employed a similar procedure in his unpublished work, notably in his 1801 analysis of the geodetic measurements made in France, see Plackett (1972) and Stigler (1999, Chap. 17).

In this context, we should also mention that the distribution named for Gauss had previously been discussed by Abraham de Moivre (1667–1754) and Pierre Simon Laplace (1749–1827); see *Husserl, Edmund* (1859–1938), Stigler (1999, Chaps. 14 and 22).

6. Asymptotic Results

In his *Théorie analytique des probabilités* of 1812, Laplace had made use of a nonrigorous asymptotic argument to establish the optimality of the method of least squares. Gauss (1816) employed a similar technique to establish the relative efficiency of the scaled sums of powers of the absolute disturbance terms when employed as estimates of the variance of their common distribution function. In particular, he found that the scaled sum of the squared disturbances is a more efficient estimator of the variance parameter than is the scaled sum of their absolute values and significantly better than the scaled sums of their third and higher powers.

7. The Most Plausible Value Criterion

In 1818–19 Gauss abandoned his earlier Bayesian derivation of the method of least squares and adopted an alternative approach which he later published as Gauss (1823–28). Farebrother (1999) has suggested that the principal features of Gauss' second approach to the fitting problem may be traced to a careful recombination of certain elements in Laplace's earlier work. Gauss adopted Laplace's decision theoretic approach to the fitting problem, and his restriction to what we would call a class of linear unbiased estimators, but he employed finite sample theory rather than asymptotic analysis, and chose a value for the parameter to minimize the expected value of the squared error of estimation where Laplace had chosen to minimize its expected absolute value.

The principle result to this new theory, now known as the Gauss–Markov theorem, identifies the least squares estimator as the member of the class of linear unbiased estimators with minimum mean squared error.

In this context, it is interesting to note that Gauss' restriction to the class of linear unbiased estimators was an accidental by-product of Laplace's earlier concern with estimators defined by the method of averages, see Stigler (1986) or Farebrother (1999). However, Gauss also showed that the scaled sum of squared deviations is an unbiased estimator of the disturbance variance, but without making any strong claims for the optimality of the result.

Gauss (1823–8) gave a very full account of the subject. In particular, he extended his earlier analysis by considering the effect on the fitting problem of

adding a further observation, or of imposing a linear constraint on the parameters of the model. The resulting updating formulas now form the basis of the Gauss–Doolittle fitting procedure and the Kalman filter of times series analysis. They also define the set of recursive residuals employed in some statistical tests.

8. Solution of Linear Equations

The solution technique now known as Gaussian elimination is a development of the procedures known to Tobias Mayer (1723–62) and the ancient Chinese for solving systems of linear equations with the same number of unknowns. In his 1809 worked example, Gauss employed a solution technique which may possibly be associated with the ancient method, see Farebrother (1999), but in 1811 he outlined the steps of a solution technique that identified the elimination procedure which now goes by his name. This procedure has two phases: in the forward elimination phase the first equation is used to eliminate one of the unknown constants from the other equations; then the second equation is used to eliminate a second unknown constant from the remaining equations; and so on. This yields an upper triangular system which is then solved for the unknown constants in reverse order.

Gauss illustrated this computational technique by a worked example concerned with the least squares determination of the orbit of the newly-discovered asteroid Pallas.

The so called Gauss–Jordan procedure differs from the original procedure in that the array of numbers is diagonalized in the forward elimination phase. This minor modification (which is familiar from textbook descriptions of the simplex procedure of linear programming) leads to a considerable simplification of the second phase, but usually at the cost of some loss of numerical accuracy, see Farebrother (1988).

9. Matrix Algebra

As a by-product of his work in number theory and the calculus of observations, Gauss (1801, 1809, 1823–8) made use of several concepts now recognized as belonging to the theory of matrices. These include linear dependence; linear transformations; the Lagrange LDL' and Cholesky LL' decompositions of a symmetric matrix; the rank-one adjustment of a symmetric matrix; the inverse of a triangular matrix; the inverse of a symmetric matrix; and, possibly, the generalized inverse of a matrix of full rank. He also explicitly proved that the inverse of a symmetric matrix is itself symmetric.

However, despite these early beginnings, the familiar matrix formulation of the least squares problem did not materialize until 1935, and it did not become a

common feature of statistical textbooks until the second half of the twentieth century, see Farebrother (1999).

10. Iterative Fitting Procedures

Gauss (1809) also explained how nonlinear relationships may be fitted by means of an iterative variant of the method of least squares, now known as the Gauss–Newton procedure, in which the fitted relationship is linearized by a Taylor series expansion and fitted by the method of least squares. This process being repeated until a satisfactory level of convergence is obtained.

Despite its early origin, and the publication by Gauss of detailed worked examples in astronomy and geodesy, the Gauss–Newton procedure was not widely adopted for practical work in other areas until the necessary computing equipment became readily available in the second half of the twentieth century. Instead, practitioners avoided the computational burden imposed by the iterative procedure by transforming their nonlinear relationships to a more linear form before estimating the parameters of the resulting equations by a *single* application of the method of selected points, the method of averages, or the method of least squares.

A second iterative procedure usually named for Gauss is the Gauss–Seidel procedure, although Gauss and Seidel were actually not concerned with the procedure that usually goes by this name, but with a distinct fitting procedure known as the method of *relaxation*; see Gauss (1823) for details.

11. Concluding Remarks

The method of least squares and the related statistical procedures known as linear regression analysis and the analysis of variance are widely used in all areas of science. Thus there can be few numerate practitioners of the social and behavioral sciences who have not read research reports that employ these methods to fit linear or curvilinear relationships to data on the variables of interest.

The method of least squares was developed by Legendre, Gauss, and Laplace in the first quarter of the nineteenth century in the context of important practical problems in astronomy and geodesy. But it was not until later in the century that Quetelet (see *Quetelet, Adolphe (1796–1874)*), Galton (see *Galton, Sir Francis (1822–1911)*), and Edgeworth (see *Edgeworth, Francis Ysidro (1845–1926)*) among others adjusted this technique to the needs of the social and behavioral sciences. There was no great shortage of practitioners trained in the use of advanced statistical techniques, but such individuals as were active in this

area were not able to treat their observations with the necessary level of detachment, so that the original source of the observations remained a significant feature of their subsequent analyses.

The method of least squares was absorbed relatively rapidly into psychology, but its rate of infusion into the economic, social, and behavioral sciences (other than psychology) was more gradual. Indeed, applications of this technique were not at all common in these areas until well into the twentieth century. See Stigler (1986, 1999) for a detailed account of the gradual transformation of these subject areas (see also *Quetelet, Adolphe (1796–1874)*; *Edgeworth, Francis Ysidro (1845–1926)*; *Thorndike, Edward Lee (1874–1949)*).

The history of the method of least squares clearly demonstrates that practical scientists with little or no interest in the social and behavioral sciences can make significant contributions to this subject area by developing general analytical procedures that can be adjusted to the particular needs of social and behavioral scientists.

Bibliography

- Bertrand J L F 1855 *Méthode des moindres carrés: mémoires sur la combinaison des observations par Ch.-Fr. Gauss*. Mallet-Bachelier, Paris [1957 Trotter H F Gauss's work (1803–1826) on the theory of least squares, *Technical Report 5*, Statistical Techniques Research Group, Princeton University]
- Börsch A, Simon P (1887) 1964 *Abhandlungen zur Methode der kleinsten Quadrate von Carl Friedrich Gauss*, Physica-Verlag, Würzburg, Germany
- Eisenhart C 1968 Carl Friedrich Gauss (1777–1855). In: Sills D (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York, Vol. 6, pp. 75–81
- Farebrother R W 1988 *Linear Least Squares Computations*. Marcel Dekker, New York
- Farebrother R W 1999 *Fitting Linear Relationships: A History of the Calculus of Observations 1750–1900*. Springer-Verlag, New York
- Gauss C F (1801) 1870 *Disquisitiones Arithmeticae*. Vol. 1 of *Carl Friedrich Gauss Werke*. Dieterichsche Universitäts-Druckerei, Göttingen, Germany [1966 English trans. Clarke A A *Disquisitiones Arithmeticae*. Yale University Press, New Haven, CT]
- Gauss C F (1809) 1871 *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. Vol. 7 of *Carl Friedrich Gauss Werke*. F. Perthes, Gotha [(1857) 1963 Davis C H *Theory of the Motion of the Heavenly Bodies Moving About the Sun in Conic Sections*. Dover, New York]
- Gauss C F (1811) 1874 *Disquisitio de elementis ellipticis Palladis ex oppositionibus annorum 1803, 1804, 1805, 1807, 1808, 1809*. In: *Carl Friedrich Gauss Werke*. Dieterichsche Universitäts-Druckerei, Göttingen, Vol. 6, pp. 1–24
- Gauss C F (1816) 1880 *Bestimmung der Genauigkeit der Beobachtungen*. *Carl Friedrich Gauss Werke*. Dieterichsche Universitäts-Druckerei, Göttingen, Vol. 4, pp. 109–217
- Gauss C F (1823–8) 1880 *Theoria combinationis observationum erroribus minimis obnoxiae* (in three parts.) *Carl Friedrich Gauss Werke*. Dieterichsche Universitäts-Druckerei, Göt-

- tingen, Vol. 4, pp. 3–93 [1995 Stewart G W. *Theory of the Combination of Observations Least Subject to Errors*. Society for Industrial and Applied Mathematics, Philadelphia, PA]
- Gauss C F (1823) 1903 Letter to Gerling, Göttingen 26 December 1823. *Carl Friedrich Gauss Werke*. Dieterichsche Universitäts-Druckerei, Göttingen, Vol. 9, pp. 278–81 [1951 Forsythe G E Gauss to Gerling on relaxation, *Mathematical Tables and Other Aids to Computation* 5: 255–8]
- May K O 1972 Carl Friedrich Gauss. In: Gillispie C C (ed.) *Dictionary of Scientific Biography*. Scribners, New York, Vol. 5, pp. 298–315
- Plackett R L 1972 The discovery of the method of least squares. *Biometrika* 59: 239–51
- Stigler S M 1986 *The History of Statistics: The Measurement of Uncertainty Before 1990*. Belknap Press of Harvard University Press, Cambridge, MA
- Stigler S M 1999 *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press, Cambridge, MA

R. W. Farebrother

Gay, Lesbian, and Bisexual Youth

1. Introduction

Youths who identify as gay, lesbian, or bisexual have become increasingly visible during the 1990s. Although constituting only about 2 percent of the adolescent population, these self-identified youths are, in turn, a subset of the much larger population of sexual-minority youths with same-sex attractions who do not adopt a gay, lesbian, or bisexual label. It is this larger group of youths that has been neglected by social and behavioral scientists. By focusing only on ‘out and visible’ youths, researchers, educators, and mental health professionals may have unwittingly created and perpetuated myths about the behavior, attitudes, and mental health of gay youths. Indeed, more recent developmental research that includes the larger category of sexual-minority youths indicate that these youths are often indistinguishable from their heterosexual peers, except in the nature of their sexual desires.

2. Definitions and Prevalence

The term *sexual minority* characterizes a broader class of individuals than those who identify as lesbian, bisexual, or gay. This is a particularly important consideration because during adolescence, youths frequently explore issues related to their sexuality that extend beyond traditional identity labels. Their sexual attractions, fantasies, and behavior may have no easy categorization. Thus, for researchers and lay people alike, the issue of *sexual identity* labels is a difficult

one. Most people are familiar with the labels ‘heterosexual,’ ‘gay,’ ‘lesbian,’ and ‘bisexual,’ but not ‘unlabeled,’ ‘bi-lesbian,’ and ‘not straight’—terms that current cohorts of youth, particularly young women, are likely to assume (Savin-Williams and Diamond 2000).

Included under the rubric of sexual minority are individuals who reject, for personal or political reasons, cultural definitions of any sexual category. Perhaps they believe that no term can adequately describe their attractions (‘unlabeled’), are uncertain as to the nature of their sexual attractions (‘questioning’), have a mixture of bisexual and lesbian attractions (‘bi-lesbian’), are without sexual attractions (‘asexuals’), use markers other than sex of person as the basis for their sexual attractions (e.g., personality, age, sexual behavior), or do not consider themselves to be traditionally heterosexual (‘not straight’).

Another contemporary sexual-minority group increasing in visibility is comprised of transgender individuals, who may be of any sexual orientation or identity. *Transgender* is an umbrella term for those who believe that they do not meet traditional definitions of masculinity or femininity consistent with their biological sex.

One simple resolution to defining who is ‘gay’ would be to eschew sexual labels altogether and rely on descriptions of behaviors, desires, or attractions. Rather than focusing on a particular sexual identity label, researchers should be asking questions about the existence, number, and proportion of sexual behaviors, attractions, or romantic relationships that an individual has with males and females.

Given cultural prohibitions against sexual minorities claiming their identity, it is not known with certainty how many youths experience same-sex attractions. Yet this number is likely to be far greater than the number of youths who identify as gay, lesbian, or bisexual. Three recent surveys illustrate this point. In a national, representative sample of adults, Laumann et al. (1994) reported that 1.4 percent of women and 2.8 percent of men identified as lesbian, bisexual, or gay. Far more, however, reported having thought about having sex with someone of their own gender—5.5 percent of women and 6 percent of men—or having experienced same-sex attractions at least once—7.5 percent of women and 7.7 percent of men. Second, a Youth Risk Behavior Survey of Massachusetts high schools (Goodenow and Hack 1998) reported that 2 percent of students self-identified as gay, lesbian, or bisexual; however, an additional 2 percent reported having engaged in same-sex behavior despite eschewing a sexual-minority label. Finally, in the National Longitudinal Study of Adolescent Health, a comprehensive, representative study of adolescents, 8 percent of males and 6.6 percent of females reported that they had same-sex romantic attractions or had been in a same-sex romantic relationship (Russell and Joyner 1998).

These findings indicate that far more youths experience same-sex attractions, romantic desires, sexual behavior, and romantic relationships than report that they are lesbian, gay, or bisexual. Thus, many 'potential' sexual-minority youths do not identify as such during adolescence and therefore fail to be included in research studies of sexual minorities. Hence, the overwhelming body of research on youths with same-sex attractions samples a select, likely nonrepresentative minority of this population. This conclusion is consistent with McConaghy's (1999) recent assertion about the most important 'unresolved issues in scientific sexology.' Too often researchers assume that sexual attractions are distributed categorically rather than dimensionally, rendering invisible the 'nongay majority with homosexual feelings.'

Although the reasons many youths with same-sex attractions decline to adopt a lesbian, gay, or bisexual label have not been investigated empirically, it is likely linked to negative cultural attitudes and beliefs about same-sex attracted individuals. The associated stigma causes many youths to forego altogether or to delay identification as a sexual minority. It is not known the ways in which those who adopt a sexual identity label differ from the much larger population of same-sex attracted individuals; however, some evidence suggests that they may be psychologically less healthy (see below).

Clearly, many youths who know that they are not heterosexual are unprepared, either developmentally or politically, to attach themselves to a culturally defined sexual category that, as currently constructed, carries profound psychological and social ramifications. To be gay implies belonging to a class of individuals who are subject to hate crimes, prejudice, and stereotypes. Unfortunately, youth often do not realize that embracing this label also opens the door to a community of like-minded individuals who can provide understanding, acceptance, and support for one's central sense of self. Given the perceived consequences of a gay identification, it is to be expected that some adolescents prefer to describe their sexual desires or attractions (e.g., 'I'm attracted to women.') rather than embrace a specific, stigmatized identity label.

3. Myths and Counter Evidence

Relatively little is known about this class of individuals who experience significant and persistent sexual attractions and fantasies for same-sex others. Until recently, researchers have ignored such individuals and behaviors, in part due to *heterocentrism* (the assumption that the universe is heterosexual) and *homonegativity* or *homophobia* (negative attitudes, beliefs, and behaviors toward those with same-sex attractions). Other reasons for this inattention include the belief that because youths are fickle with yet unformed sexuality, as evidenced by the frequency of

adolescent sexual exploration, it is not worthwhile investigating the correlates and consequences of adolescent same-sex experiences. Finally, there is the common belief that researchers should not ask sexual questions because sexuality is highly private and that inquires about sexuality would not be permitted by parents, research institutions, or funding agencies. Traditional social and religious organizations sometimes fear that by raising the issue of homosexuality, they increase the 'risk' of youths engaging in same-sex behavior and identifying as gay. As a result of this negligence, many myths about sexual-minority youths have taken root. Three prevalent myths are reviewed briefly.

3.1 Youths Develop Their Sexual Identity Through a Series of Coming-out Stages

During the 1970s and 1980s, social scientists and clinicians made sense of these young lives by proposing a number of ideal or typical 'coming-out' models (e.g., Cass 1979). These structural models were cast in a somewhat 'one-size-fits-all' mold that outlined an orderly series of stages based on a particular theoretical understanding of human development. The life histories of sexual-minority youths were proposed to follow a series of stages, from first awareness of being different, sexual exploration, and self-labeling; to first disclosures, romantic relationships, and self-acceptance; and to political/social activism and integration of a sexual identity with the rest of one's life. With few exceptions, the empirical basis for these models remains extraordinarily weak, largely because researchers have not investigated the complexities of these models, studied youths over time, developed appropriate research tools to assess youths' 'fit' with the models, or included diverse, under-represented populations of youths. Given these limitations, it should come as no surprise that recent studies found that only 10 percent of young men followed these developmental stages and that the number of young women adhering to this hierarchy is even less (Savin-Williams 1998, Savin-Williams and Diamond 2000).

3.2 Parents React With Horror and Reject Their Sexual-minority Child

Popular advice books for parents often portray the child's act of disclosure as necessarily creating a crisis within the family that may never be fully resolved. Lesions in family relationships are inflicted, wounds that may not withstand later attempts to amend or overcome, and parents encounter unique developmental issues. This literature promotes the view that the disclosure of same-sex desires by children to their parents inevitably and uniformly places youths at psychological and physical risk. Disclosure may result

in youths being disowned, thrown out of the home, or emotionally or physically harassed. Because parents may feel that their child has rejected them and everything the family stands for, family relationships may be irrevocably damaged.

Empirical research, however, does not support these dire predictions. Around 5 percent of youths suffer physical harm from parents or are thrown out of the home once they are open about their sexuality (Savin-Williams 2000). For example, in a sample of Chicago support group youths, only 3 percent reported that they had been thrown out of the home and were living in shelters or with friends (Herdt and Boxer 1993). In another study, parental rejection and anger were the two least common reactions to the youth's disclosure of same-sex sexuality, both one week and six months after the event (Ben-Ari 1995).

3.3 Sexual-minority Youths Experience Mental Health Problems, Exemplified by High Rates of Attempted Suicide

The prevalent cultural image of gay youths is that they are suicide attempting, prostituting, drug-abusing runaways who are HIV positive. Encouraging outcomes of this depiction include sympathy and support from mainstream society, financing of community support groups and research, and conversations about homosexuality in the public school system (usually in health classes). However, the long-term effect may be disastrous—convincing some youths that the only way to be young and gay is to be disordered, unhappy, and problematic. The pervasiveness of this erroneous, degrading representation has re-pathologized gay youths, almost three decades after mental health professionals removed homosexuality from the annals of psychiatric illness.

Empirically, it is clear that some youths, especially young men, who identify as a sexual-minority early in life are at risk for mental health problems. This is particularly the case for those who seek the assistance of support groups (e.g., Hershberger and D'Augelli 1995, Remafedi et al. 1998). However, these researchers have noted that support group attending youths are also the individuals most likely to be raised in dysfunctional families, subjected to barrages of peer ridicule, and to be sex-atypical in their behavior and interests. More importantly, it is often overlooked that even among these high-risk populations, the vast majority of youths who identify as gay, lesbian, or bisexual did not attempt suicide, abuse drugs, pierce their bodies, or have unsafe sex.

Here the distinction between claiming a sexual identity label and experiencing same-sex sexual attractions or behavior becomes particularly important. Although many reports in the literature document that gay youths are three to 10 times more likely to attempt suicide and that 30 percent or more of all adolescent suicides are by gay youths (Fergusson et al.

1999, Garofalo et al. 1998, Goodenow and Hack 1998, Hershberger and D'Augelli 1995, Remafedi et al. 1998), there are reasons to suspect that these numbers overstate the prevalence of suicidality among the larger population of same-sex attracted individuals. For example, Russell and Joyner (1998) found that youths who reported romantic attractions and relationships with same-sex others—a much more inclusive category of sexual minorities than those who identify as gay—were twice, rather than three or 10, times more likely to report suicide attempts as heterosexual youths, and that these sexual-minority youths constituted 15 percent, rather than 30–40 percent, of reported suicide attempts. A study of college students indicated that the proportion of young women (10 percent) and men (11 percent) with same-sex attractions who attempted suicide was no greater than a matched sample of heterosexual women (8 percent) and men (2 percent) (Savin-Williams 2000). This sample is likely to be more representative of the general population of same-sex attracted individuals because it included individuals who identified after high school and because recruitment was based on sexual attractions, not sexual identity.

4. Conclusion

Social and behavioral scientists should consider two facts about the lives of sexual-minority youths. First, they are essentially similar to all youths, independent of sexual desires, in their biology, behavior, beliefs, attitudes, socialization, and mental health. This should be a basic assumption in empirical investigations, educational programs, and clinical interventions. For example, youths of all sexualities want parental acceptance and support, are preoccupied with peer status, desire love and sex, and wonder about their future. Although these concerns can be flavored by sexual orientation—it is far more difficult to date a same- than an opposite-sex peer in most high schools—they are nevertheless adolescent tasks, ultimately to be negotiated by all youth.

A second point to be recognized—and expected given basic developmental theory—is that sexual-minority youths also vary among themselves based on particular distinguishing characteristics. Indeed, on some domains a subset of sexual-minority youths mirror heterosexual youths more than they do other sexual-minority youths. For example, young women with same-sex attractions are often more akin to heterosexual young women than they are to gay male youths. They are less genitally oriented and more likely to value and pursue romantic relationships and to use these occasions to define their sexual attractions (Diamond 1998, Savin-Williams and Diamond 2000).

A dual focus on the ways in which sexual-minority youths are similar to all other adolescents and how they vary among themselves as the result of their

sexuality, personality characteristics, biological heritage, and other developmental processes should infuse all considerations of their lives. This more comprehensive perspective is now possible because a larger and more diverse sample of contemporary youths is willing to be counted as 'not heterosexual.' As these youths become increasingly visible it will become easier for scientists to include them in research that documents their diverse and similar developmental trajectories. Although investigators of sexual-minority youths have not yet adequately sampled from this newly visible population of youths with same-sex attractions, this will most likely occur when definitions of what constitutes a sexual minority extend beyond sexual identity labels to include all youths with same-sex attractions and behavior.

This dual focus can also be seen in the lives of sexual-minority youths. From the perspective of this contemporary generation of youths, a new world has emerged regarding what it means to be a sexual minority. Consequently, it is no longer rare to disclose the nature of one's sexuality earlier than in any previous generation—while still living at home and attending the local high school. Neither is it mandatory to partake in the 'counter culture'—the 'misfits' and disenfranchised youths—of the local school or community in order to receive support and a sense of community. Whereas 'gay liberation' once characterized the only available social or political organizations for teens, Gay/Straight Alliances, school-based clubs consisting of sexual-minority youths and their allies, now populate hundreds of high schools throughout the US. These allow a youth opportunities to discover others with same-sex attractions; a place to 'hide' until ready to self-disclose; and a means to obtain support, respect, and acceptance. Even though these youths tend to blend in with many others in their junior and senior high schools and in their communities, their issues must not be neglected.

See also: Feminist Theory: Radical Lesbian; Homosexuality and Psychiatry; Identity in Childhood and Adolescence; Sex-role Development and Education; Sexual Orientation: Biological Influences; Sexual Orientation: Historical and Social Construction; Sexual Preference: Genetic Aspects; Sexual Risk Behaviors; Teen Sexuality; Youth Culture, Sociology of

Bibliography

- Ben-Ari A 1995 The discovery that an offspring is gay: Parents', gay men's, and lesbians' perspectives. *Journal of Homosexuality* **30**: 89–112
- Cass V C 1979 Homosexual identity formation: A theoretical model. *Journal of Homosexuality* **4**: 219–35
- Diamond L M 1998 Development of sexual orientation among adolescent and young adult women. *Developmental Psychology* **34**: 1085–95
- Fergusson D M, Horwood L J, Beautrais A L 1999 Is sexual orientation related to mental health problems and suicidality in young people? *Archives of General Psychiatry* **56**: 876–80
- Garofalo R, Wolf R C, Kessel S, Palfrey J, DuRant R H 1998 The association between health risk behaviors and sexual orientation among a school-based sample of adolescents. *Pediatrics* **101**: 895–902
- Goodenow C, Hack T 1998 Risks Facing Gay, Lesbian, and Bisexual High School Adolescents: The Massachusetts Youth Risk Behavior Survey. Paper presented at the American Psychological Association Convention, San Francisco
- Herd G, Boxer A 1993 *Children of Horizons: How Gay and Lesbian Teens Are Leading a New Way Out of the Closet*. Beacon Press, Boston, MA
- Hershberger S L, D'Augelli A R 1995 The impact of victimization on the mental health and suicidality of lesbians, gay and bisexual youth. *Developmental Psychology* **31**: 65–74
- Laumann E O, Gagnon J H, Michael R T, Michaels S 1994 *The Social Organization of Sexuality: Sexual Practices in the United States*. University of Chicago Press, Chicago
- McConaghy N 1999 Unresolved issues in scientific sexology. *Archives of Sexual Behavior* **28**: 285–318
- Remafedi G, French S, Story M, Resnick M D, Blum R 1998 The relationship between suicide risk and sexual orientation: Results of a population-based study. *American Journal of Public Health* **88**: 57–60
- Russell S T, Joyner K 1998 August Adolescent Sexual Orientation and Suicide Risk: Evidence from a National Study. Paper presented at the American Sociological Association, San Francisco
- Savin-Williams R C 1998 *And Then I Became Gay: Young Men's Stories*. Routledge, New York
- Savin-Williams R C 2000 Suicide attempts among sexual-minority youths: Controversy re-ignited. Unpublished manuscript available from the author, Cornell University Press, Ithaca, NY
- Savin-Williams R C, Diamond L M 2000 Sexual identity trajectories among sexual-minority youths: Gender comparisons. *Archives of Sexual Behavior* **29**: 419–40
- Savin-Williams R C 2001 "Mom, Dad, I'm Gay". *How families Negotiate Coming Out*, 1st edn. American Psychological Association Press, Washington, DC

R. C. Savin-Williams and L. M. Diamond

Gay/Lesbian Movements

For a movement to occur there needs to be a sense of common identity or purpose and sufficient political and economic space to organize. It is not surprising that gay/lesbian movements are very much the products of modern liberal democratic and capitalist societies, nor that the globalization of these concepts had led to a rapid expansion of gay/lesbian identities across the world.

1. The Emergence of 'the Modern Homosexual'

While homosexual behavior has existed in most societies at most times in history, the creation of a

specific identity based upon homosexual behavior is a far more recent and limited phenomenon. The term 'homosexual' was coined in 1869 by the Hungarian doctor Karoly Benkert, but there is some evidence that people defined by a shared sexual attraction to the same gender had already formed loose social groupings in several European cities. More common however was a range of understandings of sexuality and gender which saw homosexual behavior as part of an undifferentiated range of possible sexual behaviors, and an equal range of acceptance and condemnation of such behaviors. Such understandings were always heavily gendered, and often ignored the potential for female homosexuality.

By the end of the nineteenth century the idea of the homosexual existed to a sufficient extent that it became possible to imagine the creation of both social and political organizations for those people who identified with the term. The first overtly political organization, the Scientific-Humanitarian Committee, was founded by the German Magnus Hirschfeld in 1897 and campaigned for the decriminalization of homosexuality.

Hirschfeld believed that homosexuals constituted a 'third sex,' and therefore were inherently different, and no more deserving of condemnation than, say, someone born left-handed. However, against the idea of the 'third sex' Freud postulated a universal 'polymorphous perversity,' so that childhood experiences determined whether or not one developed a primarily heterosexual or homosexual orientation. While many psychoanalysts developed a strong antipathy to homosexuality, seeing it as a pathology to be cured through therapy, another tradition, which would emerge in the 1960s, used Freud's works to argue against the idea of a heterosexual norm.

After World War I both Europe and the USA saw the development of small homosexual social groups in major cities, most famously Paris and Berlin. Hirschfeld's organization was ended brutally by the Nazis, and it was not until the aftermath of World War II that new organizations emerged—COC (originally the Shakespeare Club) in the Netherlands; *Arcadie* in France; *der Kreis* in Switzerland. COC remains, by far the longest lived gay/lesbian organization in the world. In the USA the first groups were established in the 1950s (the Mattachine Society in 1951, Daughters of Bilitis in 1955) but these groups remained largely underground, although some overt political protest began in the early 1960s (d'Emilio 1983).

In Europe and North America the dislocations of World War II meant that millions of people came in contact with underground gay and lesbian worlds, and although the post-war organizations remained very small, homosexual communities developed in most major western cities through the 1950s and 1960s (d'Emilio 1983, Kennedy and Davis 1993). It was the presence of both social and commercial networks

which allowed for the rapid emergence of a new sort of gay/lesbian movement in the aftermath of the major social and political upheavals in most liberal democracies from the end of the 1960s.

2. *The Development of the Gay/Lesbian Movement*

The contemporary movement was in a sense born out of the political, social, and cultural changes of 'the sixties,' symbolized by the growing freedom which surrounded discussion of sexuality, the rebirth of feminism, and the student movements of 1968. In France and Italy it was the events of May 1968 which led to the emergence of a radical gay movement; in the USA the turning point is usually assumed to be the riots which followed a raid on the Stonewall Inn in Greenwich Village, out of which the New York Gay Liberation Front was born. The gay/lesbian movement of this period saw itself as part of a larger radical assault on hegemonic cultural and political ideas and institutions, and was closely linked to various elements of New Left theory (Altman 1971). Small gay liberation groups also appeared in Canada, the UK (Weeks 1990), Scandinavia, Australasia, and some Latin American countries: Argentina in 1969; Mexico in 1971; Puerto Rico in 1974. The first Israeli gay/lesbian organization was established in 1976.

The post-1968 movement differed from its predecessors in its willingness to assert a new sense of gay/lesbian identity which demanded complete equality, at both an individual and a communal level. Although the original use of the term 'gay' included both women and men, tensions quickly emerged along gender lines, in part because of the strong interconnection between radical lesbianism and feminism, which had no equivalent for gay men (Faderman 1981, Rich 1980). During the 1970s there was increasing divergence between lesbian and gay male organizations, as a growing commercial gay world opened up new ways of living as a homosexual, particularly for men. As one novelist wrote: 'Almost through an act of will, I had made myself embrace this new identity of mine and never look back. I had gay friends. I ate at gay restaurants. I went to gay bars. I had my apartment near DuPont Circle' (Russell 1994, p. 231).

While 'gay ghettos' such as that described here only developed in a few cities outside the USA—or, as in the case of Paris's Marais, developed a decade later—the growth of gay and lesbian commercial space has been a feature of almost all western countries from the 1970s on. In many capitalist societies the 'pink dollar' has become seen as an important niche market, and this has made possible a flourishing gay press and an expansion of businesses and professional services aimed at a specifically gay and lesbian market. By the 1990s government authorities in a number of countries

were promoting services for lesbian and gay travelers, and Sydney's Gay and Lesbian *Mardi Gras* was recognized as one of Australia's most successful tourist attractions. The gay/lesbian press, originally an expression of the political movement, has become increasingly preoccupied by 'lifestyle' issues and the need to attract 'mainstream' advertizers.

Meanwhile women developed a parallel if sometimes overlapping world, and developed a set of social institutions which allowed room for a 'women-centered' politics, sometimes extended to a rejection of anything that could be seen as male-dominated (Stein 1997). In some countries links were maintained between women and men through common political organisations such as the National Gay (later Gay and Lesbian) Task Force in the USA. The Task Force's commitment to working through mainstream institutions underlines the various splits around gender, political style, and class which became apparent in the movement during the 1970s and 1980s. As the radical energies of the early 1970s dissipated, so too did organized lesbian and gay politics, though the trajectory of the movements is rather different in different countries (Adam et al. 1998). Even in the western world there were significant differences between the English-speaking democracies, where decriminalization of homosexual behavior seemed the top priority; the countries of northern Europe, especially the Netherlands and Denmark, where social acceptance seemed most assured; and those of southern Europe whose laws were less repressive and where a language of universal rights was often used to argue against any sort of identity politics.

The 1970s and 1980s saw a rapid shift in most western countries in attitudes towards homosexuality, with growing recognition of gay men and lesbians as constituting legitimate communities with a certain political influence. Openly lesbian and gay legislators were elected in Norway, Canada, and the USA, and almost all western countries abolished remaining laws which decriminalized homosexual activity. (The great exception was the USA, where homosexual acts remain illegal in a number of states, in part because of a decision by the Supreme Court in 1986 which upheld Georgia's sodomy laws.) The rate of change was particularly rapid in cases such as Spain which democratized during this period, a shift which would be echoed in Latin America in the late 1980s and some parts of Eastern Europe in the 1990s (Long 1998).

The rapidity of changes in attitudes towards homosexuality meant a corresponding backlash from moral conservatives, which was most pronounced in the USA. The mobilization of religious conservatives by groups such as the Moral Majority led to bitter attacks on homosexuality, both in the political arena and, not infrequently, in direct violence. While there were examples of similar reactions elsewhere, such as the Thatcher government's prohibitions on 'the promotion of homosexuality,' it is probably true that the

gains for gay and lesbian rights was rapid and irreversible in most western democracies.

3. *The Impact of HIV/AIDS*

The first reports of what was to be named 'AIDS' (acquired immune deficiency syndrome) came with severe illnesses among young homosexual men on both American coasts, and for a short time the new disease was known as 'gay-related immune deficiency syndrome.' Within a few years two major discoveries were made: the syndrome was transmitted through sexual and blood contact, and in many parts of the world the bulk of sexual transmission was heterosexual. Nonetheless the epidemiological link to homosexuals in the rich world has a continuing impact on how the epidemic is perceived, and has had a major impact upon gay organization. The first responses to the new disease came from gay communities in major western cities, with the Gay Men's Health Crisis in New York (founded in 1982) becoming a model for groups such as the various Australian state AIDS Councils, the Terrence Higgins Trust in Britain and the *Deutsche AIDS Hilfe*.

During the 1980s the need to lobby governments, to provide information and preventive education, and to develop home-care and emotional support for those with HIV/AIDS, dominated gay movements in most western countries, and sometimes involved considerable numbers of lesbians as well. As the epidemic was discovered to be growing very rapidly in Africa and parts of Latin America, AIDS was to be the focus for a new wave of homosexual organizing in a number of 'developing' countries. Groups such as Pink Triangle in Malaysia or Triangulo Rosa in Costa Rica came into existence as *de facto* homosexual organizations because of the epidemic (Altman 1994). The development of People With AIDS groups built on the earlier gay movement concept of 'coming out.'

The impact of AIDS on the movement varied from country to country, but overall it meant a much closer relationship between the state and gay organizations, particularly in countries such as Denmark, Switzerland, Canada, and Australia where the epidemic remained largely due to homosexual transmission and national strategies incorporated community responses. At a global level the development of international responses through the Global Program on AIDS and then UNAIDS saw some institutional support for community organizations, including gay ones.

4. *The 'Queer Movement'*

One of the successes of the movement was the development of 'gay and lesbian studies,' particularly in the Netherlands and the USA. In the case of the latter the development of post-modern theory led to

new ways of conceptualizing homosexuality which sought to move beyond fixed notions of identity and to destabilize assumptions about the relations between sexuality and gender. At an academic level this was expressed in the notion of 'queer theory,' which was played out politically in two major ways: the development of a new radicalism and the extension of the movement to include issues of bisexuality and transgenderism. One might note these were both elements of the gay liberation movement of the early 1970s, which had tended to vanish in the mainstreaming of the movement in the 1980s. (See *Masculinities and Femininities; Queer Theory*.)

In the USA, and to a lesser extent in the UK, the movement has extended itself to recognize 'bisexual and trans-gendered' people as well as 'gay and lesbian.' Elsewhere this move has been more hotly contested, as in battles around membership criteria for Sydney's Gay and Lesbian *Mardi Gras*. Yet despite the rhetoric, the gay and lesbian movement has continued to function as do most pressure groups in western societies, using a judicious mix of popular protest, lobbying, media propaganda and in some cases direct participation in the electoral process.

In the USA two issues dominated mainstream gay and lesbian politics in the 1990s: the place of gay men and lesbians in the military and arguments about same-sex marriage. The first became a major embarrassment for the Clinton Administration in 1993, when the President's pledge to remove discrimination from the military ran into considerable Congressional opposition. The resulting 'don't ask, don't tell' policy was far weaker than the policies in a number of other western countries.

The first national recognition of same-sex partnerships came in Denmark in 1989 and arguments about state recognition of homosexual relationships have now extended to almost all of Europe. A law proposed in France in 1998 would have recognized cohabiting couples (both hetero- and homosexual) 'who cannot or do not wish to marry,' and was bitterly opposed by the right. The following year the Senate amended the proposal to cover only heterosexual couples. The idea of 'gay marriage' became a major issue in the USA in 1996, following a case in the Hawaii Supreme Court which seemed likely to recognize same-sex marriage as constitutional, following which Congress passed—and President Clinton signed—the Defense of Marriage Act, which would have refused recognition in other states of such a ruling in Hawaii.

5. *The Development of an International Gay/Lesbian Movement*

In 1978 the International Gay (later Gay and Lesbian) Association (ILGA) was formed at a conference in Coventry, England. While ILGA has largely been

driven by northern Europeans, it now has member groups from over 70 countries and has organized meetings in several southern cities. Other networks, often linked to feminist and AIDS organizing, have been created in the past two decades, and emerging lesbian and gay movements are increasingly likely to be in constant contact with groups across the world. Such networks argue for a certain universality of homosexual identity, and their major strategy has involved appeals to universal norms of human rights to oppose persecution and discrimination against homosexuals across the world. It is difficult to avoid a certain amount of neo-colonialism in such universalist arguments, as Ann Ferguson has pointed out (Ferguson 1990, pp. 82–4). What is perhaps striking is that while the dominant image of gay and lesbian identity and culture is undoubtedly American, the development of international networks has owed comparatively little to American leadership. On the other hand, the two largest international gay/lesbian 'networks' are probably those based around the Metropolitan Community Church and the Gay Games, both American in origin.

It may be that a political movement is the least likely part of western concepts of homosexual identity to be adopted in many parts of the world, even as they enthusiastically embrace the mores and imagery of western queendom. The particular form of identity politics which allowed for the mobilization of a lesbian/gay electoral pressure in countries like the USA, the Netherlands or even France may not be appropriate elsewhere, even if western-style liberal democracy triumphs. The need of western lesbians to engage in identity politics as a means of enhancing self-esteem may not be felt in other societies. Nonetheless there has been a rapid growth of lesbian and gay organizing during the 1990s in countries as dissimilar as Nicaragua, Slovakia, India, and the Philippines.

The western lesbian/gay movement emerged in conditions of affluence and liberal democracy, where despite other large social issues it was possible to develop a politics around sexuality which is more difficult in countries where the structures of political life are constantly contested. Writing of contemporary South Africa, Mark Gevisser notes: 'Race-identification overpowers everything else—class, gender and sexuality' (Gevisser 1999). In the same way basic questions of political economy and democratization will impact on the future development of gay/lesbian movements in much of Asia and Africa. Yet in Latin America and Eastern Europe gay/lesbian movements have grown considerably in the past decade, and there are now signs of their emergence in some parts of Africa, for example in Kenya and Zimbabwe where President Mugabe has consistently attacked homosexuality as the product of colonialism. The twin impact of globalization and the discourses of international human rights are likely to see a strengthening of both gay/lesbian identities and rights in the future.

See also: Aboriginal Rights; Feminist Theory: Radical Lesbian; Gay, Lesbian, and Bisexual Youth; Gender and Feminist Studies; Identity Movements; Lesbians: Historical Perspectives; Lesbians: Social and Economic Situation; Masculinities and Femininities; Queer Theory; Rationality and Feminist Thought; Regulation: Sexual Behavior; Sexual Orientation: Biological Influences; Sexuality and Gender; Social Movements and Gender; Social Movements, History of: General; Social Movements: Psychological Perspectives; Social Movements, Sociology of; Transsexuality, Transvestism, and Transgender

Bibliography

- Adam B, Duyvendak J W, Krouwel A (eds.) 1998 *The Global Emergence of Gay and Lesbian Politics*. Temple University Press, Philadelphia, PA
- Altman D 1971 *Homosexual: Oppression and Liberation*. NY. Outerbridge and Dienstfrey (new edition New York University Press 1996)
- Altman D 1994 *Power and Community*. Taylor & Francis, London
- Blasius M, Phelan S (eds.) 1997 *We Are Everywhere: A Historical Sourcebook of Gay and Lesbian Politics*. Routledge, NY
- d'Emilio J 1983 *Sexual Politics. Sexual Communities*. University of Chicago Press, Chicago
- Faderman L 1981 *Surpassing the Love of Men*. Morrow, NY
- Ferguson A 1990 Is there a lesbian culture? In: Allen J (ed.) *Lesbian Philosophies and Culture*. State University of New York Press, NY
- Gevisser M 1999 Gay life in South Africa. In: Drucker P (ed.) *Different Rainbows*. GMP, London
- Hendriks A, Tielman R, van der Veen E 1993 *The Third Pink Book*. Prometheus Books, Buffalo, NY
- Kennedy E, Davis M 1993 *Boots of Leather. Slippers of Gold*. Routledge, New York
- Long S 1998 Gay and lesbian movements in Eastern Europe. In: Adam B, Duyvendak J W, Krouwel A (eds.) *The Global Emergence of Gay and Lesbian Politics*. Temple University Press, Philadelphia, PA
- Plummer K (ed.) 1992 *Modern Homosexualities*. Routledge, London
- Rich A 1980 Compulsory heterosexuality and lesbian existence. *Signs* 5: 4
- Russell P 1994 *Sea of Tranquillity*. Dutton, NY
- Stein A 1997 *Sex and Sensibility*. University of California Press, CA
- Weeks J 1990 *Coming Out: Homosexual Politics in Britain from the Nineteenth Century to the Present*. Rev. edn. Quartet, London

D. Altman

Gellner, Ernest (1925–95)

Ernest Gellner was born in Paris on December 9, 1925 and spent his childhood in Prague. The family moved to England in 1939 ahead of the Nazi invasion. At Oxford, where his studies were interrupted by military

service in the Czech Brigade, he specialized in philosophy. After teaching philosophy for two years at Edinburgh, he joined the Department of Sociology at the London School of Economics (LSE) in 1954. Here, stimulated also by his discovery of social anthropology in the department established by Malinowski, he formulated in *Words and Things* (1959) a withering and controversial critique of the linguistic philosophy that was dominant in Oxford at the time. In 1962 he obtained a doctorate in anthropology, based on fieldwork in Morocco. *Thought and Change* (1965) outlined many of the themes that he followed up in later years, including his theory of nationalism. In 1984 Gellner succeeded Jack Goody in the Chair of Social Anthropology at Cambridge, where he was elected to a Fellowship at King's College. In 1991 he returned to Prague to launch the Centre for the Study of Nationalism at the newly established Central European University. He died in Prague on November 5, 1995.

Gellner's interests ranged well beyond the conventional boundaries of the subjects in which he held academic posts. His prolific and diverse writing was securely anchored in his early philosophical studies, particularly in Hume and Kant, augmented later by his reading of Weber, Durkheim, and Malinowski. In short, he drew on classical sources in several disciplines. Negative influences were equally important. Marxism was rejected because of its economic determinism, though much of Gellner's own work emphasized a materialist standpoint and he maintained an enthusiastic curiosity for Soviet Marxist anthropology. Above all, he rejected the relativism of the later Wittgenstein.

Gellner's combination of philosophy and anthropology was far from the continental genre of philosophical anthropology. Rather, he hailed Malinowski as the innovator who made empirical observation in the course of long-term fieldwork the revolutionary basis of modern anthropological knowledge. His fascination with Malinowski led Gellner, late in his career, to explore the founder's intellectual roots in late Habsburg Cracow and to an illuminating (posthumously published) contrast with Wittgenstein (Gellner 1998).

Alongside the functionalist holism of Malinowski, Gellner was also influenced by A. R. Radcliffe-Brown's positivist insistence on the comparison of social structures. His own work prioritized pattern and system, rather than close-up ethnographic analysis of the flux of events. He was interested in all kinds of systems of ideas, and especially in linking ideologies to long-term evolutionary change and to the distribution of economic and political power. Gellner's manner of explicating societies by means of elegant structural models or ideal types was at odds with most of his contemporaries, who stressed interpretation and local knowledge. More controversially, he insisted that *one* culture, that of the modern west, based ultimately upon premises of individualism and ration-

ality, was technologically and ‘cognitively’ more powerful than all others. This anthropological philosophy (1975, 1985a) was complemented by a three-stage philosophy of history (1991). How to explain the dominance of the west and how best to spread the benefits of that superiority were central themes throughout his work.

Among the more specific research themes addressed by Gellner, the most significant in his lifetime was his theory of nationalism. In reaction to his LSE colleague Elie Kedourie, Gellner emphasized not the dissemination of the nationalist idea, but the functional needs of a modern industrial society, which requires a highly mobile labor force and educational institutions to promote literacy in a single ‘high culture.’ This model, elaborated in several variations over more than three decades, was not primarily supported with fieldwork materials but with vivid imaginary sketches, notably that of ‘Ruritania’ (1983). These were evidently much influenced by his own Central European roots, and perhaps by earlier scholars with a similar background, such as Hans Kohn and Karl Deutsch. Gellner’s ‘modernist’ emphasis on the uneven spread of industrialism does not offer a fully satisfactory explanation for the actual spread of nationalism outside Europe, or even within plausible Ruritania. It offers little help in predicting the circumstances in which national sentiment becomes ‘hot.’ It remains, however, as even critics have acknowledged, the best general theory of nationalism that any twentieth-century social scientist has devised (Hall 1998).

In his studies of Islam, Gellner supplemented an abstract model with data from many years of fieldwork in the High Atlas Mountains. His monograph, *Saints of the Atlas* (1969), an analysis of the role played by charismatic religious specialists in the maintenance of political order in the absence of state power, is a classic of the British school. Intellectual inspiration was drawn not only from the segmentary lineage theory of the French pioneers and from Evans-Pritchard, but also from the cyclical models of tribal–urban interaction developed centuries earlier by Ibn Khaldun. Gellner’s generalization of the ‘pendulum swing model’ to the wider Moslem world (1981) has not been widely accepted, while even in the Moroccan case his application of segmentary theory has been criticized by later fieldworkers (see Munson, in Hall and Jarvie 1996). However, his schematic view of Islam, with its text-based high culture, as the most modern of the great religions, gave perceptive insight both into the relative absence of nationalism in these countries and the rise of ‘fundamentalism.’

The Marxist–Leninist régimes of twentieth-century Eurasia were another form of social organization of abiding concern to Gellner, for reasons that connect closely with his biography. Although he visited the Soviet Union frequently and spent a year in Moscow at the height of *Perestroika*, most of his fieldwork in this case was textual. He delighted in reading between

the lines of ethnographers’ adaptations of Marxist theories in order to see what light these might throw on Soviet society (1988b). Though he did not predict the demise of the socialist system, his first-hand observations of socialist disenchantment helped him to get to the heart of the matter: no ideology that promised salvation here on this earth was likely to endure if the utopians could neither satisfy accelerating consumer expectations nor control corruption and widening inequalities. In his last years he found the concept of ‘civil society’ particularly useful in defining a blueprint for a modern democracy that would avoid both the disasters of totalitarian central planning and the tribal ‘tyranny of cousins’ (1994).

Testimony to Gellner’s influence in his lifetime can be found in a remarkable multidisciplinary collection of essays, to which he was able to reply shortly before his death (Hall and Jarvie 1996). Part of the explanation for this fame lies in the topicality of the subjects he treated: nationalism, Islamic fundamentalism and socialism all were vital problems of the age. Another element in his appeal is the allure of a scholar who was not merely radically disrespectful of disciplinary barriers but who consistently opposed the most influential intellectual currents of the day. His critique of Oxford philosophy was followed by equally disrespectful dismissals of many other ‘churches,’ including Parisian structuralism, Frankfurt neo-Marxism, psychoanalysis (1985b), the ‘Orientalism’ paradigm of Edward Said, and every variety of ‘postmodernism’ (1992). Even his relationship with social anthropology, the discipline in which he came closest to finding a home, was often uneasy. He remained basically loyal to the canons of ‘structural-functionalism’ long after these were abandoned by the profession, yet he combined this with an even more unfashionable sympathy for evolutionist agendas. There were sharp disagreements, for example with Rodney Needham over the ‘concept of kinship’ (1974) and with Clifford Geertz, whom he recognized as the leading contemporary exponent of the erroneous relativist–idealistic concept of culture.

It is not therefore surprising that neither anthropology nor any of the other disciplines to which he contributed has a Gellnerian school. His diffuse impact owes much to his forthright stance in defence of ‘modernist’ enlightenment ideals in a century in which these foundations have seemed increasingly fragile to the great majority of the West’s social scientists. His temperament and essayistic, aphoristic style were well suited to this role, which of course in the eyes of those he attacked became the role of bogeyman. This stubborn defender of Cartesian ideals of rationality, clarity and *truth* delighted in controversy and had a preference for striking images and analogies over more prosaic social science techniques; but these were an integral part of his appeal. He is best seen not as a backward-looking defender of European traditions, but as a scholar who strove to

adapt a European liberal heritage to reach a better understanding of all the problems of the modern world. Caricatures such as 'enlightenment fundamentalist' are wide of the mark (even if he himself regularly suggested such labels). Despite his criticism of Wittgenstein and of the relativist trend in anthropology, he was no more willing than Malinowski to dispense with the concept of culture. He insists on the uniqueness of modern western cognition because it is stronger than any other, but he does not claim that it is *better*. Compared to Popper and Hayek, this Central European is painfully aware that decent human communities require stronger social foundations than those which liberal individualism has been able to provide. It is this balancing act and the suggestiveness of his ideal types which make Ernest Gellner a continuing source of inspiration.

See also: Determinism: Social and Economic; Historical Explanation, Theories of: Philosophical Aspects; Nationalism: Contemporary Issues; Nationalism: General; Nationalism, Historical Aspects of: Arab World; Nationalism, Historical Aspects of: The West

Bibliography

- Gellner E 1959 *Words and Things, A Critical Account of Linguistic Philosophy and a Study in Ideology*. Gollancz, London
- Gellner E 1965 *Thought and Change*. Weidenfeld and Nicolson, London
- Gellner E 1969 *Saints of the Atlas*. Weidenfeld and Nicolson, London
- Gellner E 1974 *Contemporary Thought and Politics*. Routledge and Kegan Paul, London
- Gellner E 1975 *Legitimation of Belief*. Cambridge University Press, Cambridge, UK
- Gellner E 1981 *Muslim Society*. Cambridge University Press, Cambridge, UK
- Gellner E 1983 *Nations and Nationalism*. Blackwell, Oxford, UK
- Gellner E 1985a *Relativism and the Social Sciences*. Cambridge University Press, Cambridge, UK
- Gellner E 1985b *The Psychoanalytic Movement, or The Cunning of Unreason*. Fontana Press, London
- Gellner E 1988b *State and Society in Soviet Thought*. Blackwell, Oxford, UK
- Gellner E 1991 *Plough, Sword and Book*. Paladin Grafton Books, London
- Gellner E 1992 *Post-Modernism, Reason and Religion*. Routledge, London
- Gellner E 1994 *Conditions of Liberty. Civil Society and its Rivals*. Hamish Hamilton, London
- Gellner E 1998 *Language and Solitude. Wittgenstein, Malinowski and the Habsburg Dilemma*. Cambridge University Press, Cambridge
- Hall J A (ed.) 1998 *The State of the Nation: Ernest Gellner and the Theory of Nationalism*. Cambridge University Press, Cambridge, UK

Hall J A, Jarvie I (eds.) 1996 *The Social Philosophy of Ernest Gellner*. Rodopi, Amsterdam

C. M. Hann

Gender and Cancer

Incidence rates for cancer, a group of diseases characterized by the rapid and uncontrolled multiplication and spread of abnormal cells, differ for men and women. Explaining this variation requires consideration of the role of both sex (biologically determined factors) and gender (socioculturally determined factors) in the onset, progression, and treatment of the disease. Although cancer etiology and treatment are complex and multidetermined, there is clear evidence that sex differences in anatomy, physiology, and hormone exposure and gender differences in health-related behavior, environmental exposure, and access to health care play central roles.

1. Cancer Incidence Rates for Women and Men

The neoplastic process can occur in any part of the body, with rates of occurrence differing substantially by anatomic organs, between women and men, and across nations. For example, breast cancer is by far the most frequently diagnosed cancer in women worldwide, due largely to high rates of breast cancer in industrialized countries. Cervical cancer, the site with the second highest incidence rate in women, is diagnosed primarily in women from developing countries, who have much lower rates of breast cancer. Lung cancer is the most common site of cancer in men, whereas it is fifth in frequency among women (Boyle and Maisonneuve 1998). Tables 1 and 2 provide the range in rates of the most common cancers for women and men, indicating where the highest and lowest incidence occurs. The prevalence of cancer in sex organs and the wide international variation within sex suggest that both sex and gender are important.

2. Problems in the Research Literature

It is difficult to ascertain the precise contributions of sex- and gender-related variables to the etiology and treatment of cancer for two sets of reasons. First, these variables have rarely been the focus of careful scientific investigation. Until very recently, subjects in research with both animals and humans have been predominantly male. Many studies included only males and, when both sexes were included, they were typically combined into one sample such that the data from the much larger proportion of males were assumed to generalize to females (Sechzer et al. 1994). The very

Table 1

Highest and lowest incidence rate for common cancer sites for women*

Cancer site	Highest		Lowest	
	Population	Rate	Population	Rate
Breast	US bay area: White	104.2	The Gambia	3.4
Cervical	Peru, Trujillo	54.6	Israel: non-Jews	2.6
Colon	Bermuda: Black	34.4	Algeria, Sétif	0.9
Liver	Thailand, Khon Kaan	38.3	Canada, Prince Edward Island	0.1
Lung	New Zealand: Maori	62.2	India, Madras	1.4
Stomach	Japan, Yamagata	42.9	India, Ahmadabad	1.5
ALL SITES	Canada, British Columbia	345.4	The Gambia	39.6

Source: Parkin et al. 1992

*Rates are per 100,000 and are age-adjusted to the world standard.

Table 2

Highest and lowest incidence rate for common cancer sites for men*

Cancer site	Highest		Lowest	
	Population	Rate	Population	Rate
Colon	US, Hawaii: Japanese	37.2	The Gambia	0.7
Liver	Thailand, Khon Kaan	80.0	Netherlands, Maastricht	0.8
	China, Qidong			
Lung	New Zealand: Maori	119.1	The Gambia	1.0
Prostate	US, Atlanta: Black	102.0	China, Qidong	0.8
Stomach	Japan, Yamagata	93.3	India, Ahmadabad	2.1
ALL SITES	Australia, Tasmania	493.8	The Gambia	59.1

Source: Parkin et al. 1992

*Rates are per 100,000 and are age-adjusted to the world standard.

hormonal and physiological differences that appear to be central to some forms of cancer in women were considered unwanted sources of experimental error. The second set of problems is inherent in the questions of interest. Even with increasing attention being paid to sex, and to a lesser extent to gender, it is difficult to identify clear causal relations because of the complex interactions between and covariation of biological sex and sociocultural gender roles. Animal models cannot adequately explore gender, and there is little opportunity to isolate specific effects in human populations. It is difficult to make blanket statements that are accurate for all cultures and all cancers. Cultural values and constraints regarding gender roles differ substantially across societies, and cancers at different sites vary greatly in terms of incidence and mortality rates, epidemiology, risk factors and causes, and treatments (Harras et al. 1996).

3. Etiology and Prevention

In addition to the obvious relevance of sexual anatomy to several of the most common forms of cancer, other sex-related variables are involved in cancer etiology

(Meyerowitz et al. 2000). Sex hormones often play a central role in the onset and progression of cancers of the reproductive organs and breasts, as well as other sites that have receptors for sex hormones, such as colorectal and laryngeal tissue. Also, the normal physiological variations between women and men in the functioning of virtually all systems of the body have implications for cancer (Legato 1998). For example, the slower colonic transit time and lower excretion of bile acids in women account in part for colon cancer in women, one of the few sex-neutral sites of cancer for which incidence rates for men are not higher than rates for women.

Socioculturally determined factors appear to play an even more important role in cancer etiology than hormones and physiology. It has been estimated that as many as 75 percent of cancer deaths are attributable to behavior, particularly tobacco use and diet (Boyle and Maisonneuve 1998). Women tend to lead healthier lives than men in areas that have been associated with cancer risk. They use less tobacco, drink less alcohol, eat more fruit and vegetables, consume less animal and dairy fat, and apply sunscreen more regularly than men (Meyerowitz et al. 2000). When women's health habits shift to approximate those of men, the gap

between male and female cancer rates shift as well, as seen in the dramatic increase in both smoking and lung cancer rates among non-Hispanic white women in the US over the past few decades.

There are also gender differences in cancer risk due to exposure to environmental carcinogens and, possibly, stress-inducing life circumstances. Occupational segregation in many sectors of the workforce means that women and men are exposed to different toxins and other workplace hazards (Krieger et al. 1993). For example, gender differences in exposure to second-hand smoke in hospitality jobs such as waiting on tables in restaurants or to industrial toxins in factories will vary depending on cultural norms regarding whether these are viewed as primarily jobs for women or for men. Sedentary jobs also have been associated with increased risk of some forms of cancer (Meyerowitz et al. 2000), potentially placing women in clerical jobs at higher risk than men in more physically active lines of work. Additionally, there is growing evidence that stress can influence natural killer cell activity and other aspects of immune functioning that may, in turn, effect cancer onset or progression (Cohen and Herbert 1996). Psychosocial variables that have been associated with immune suppression—inhibition of emotional expression, passive and avoidant coping, lack of social support, and multiple life stressors—can all be influenced by gender-role expectations and cultural values. For example, proscriptions against seeking social support in men and stress associated with filling multiple roles in women have been suggested as important gender differences in the US (Good et al. 2000, Sechzer et al. 1994).

In addition to their individual influences, sex and gender can interact in cancer etiology. The most salient example is cervical cancer, for which a sexually transmitted virus, Human Papilloma Virus (HPV), appears to be a primary causal agent (Boyle and Maisonneuve 1998). Women who have unprotected sex with multiple sex partners or with men who have multiple sexual contacts are at substantially increased risk of contracting the disease. Thus, social norms regarding sexual behavior for women and men and regarding women's rights to demand safe sex practices from their partners will influence cervical cancer rates.

4. Early Diagnosis and Treatment

Once a neoplastic process has begun, survival is dependent upon early diagnosis and effective treatment. Given that women are more likely than men to fall below the poverty line, their access to adequate care can be severely limited in countries where a fee-for-service health care model prevails. Even when finances are not a limiting factor, cultural values about the appropriateness of obtaining care can play a role, as is the case in some cultures where it is up to male relatives to determine whether women should receive medical care. Data regarding gender differences in the

extent to which individuals who do have access to care adhere to medical recommendations for cancer screening and treatment are mixed, although most studies seem to indicate lower adherence rates for men (Good et al. 2000). Rigid adherence to strictly traditional gender roles for either masculinity, such as an unwillingness to accept physical vulnerability and dependence, or femininity, such as passivity in the face of medical authority or pressures to maintain a caregiver role, can impede optimal care (Good et al. 2000, Sechzer et al. 1994). Even when the individual does not hold these values himself or herself, gender roles can have an influence either through policy decision or implicit biases. For example, male physicians have been found to be less likely than female physicians to recommend breast and cervical cancer screening, whereas they are equally likely to recommend sex-neutral screenings such as blood pressure tests (see *Cancer Screening*).

In addition to the influence of gender roles on diagnosis and treatment, there is some evidence to suggest sex differences. Rates of cancer survival differ between men and women, with women having the survival advantage for most cancer sites (Legato 1998). There also is growing evidence that the pharmacokinetics and drug toxicities of some chemotherapeutic agents differ between women and men, indicating the need for sex-specific prescribing standards (Legato 1998).

5. Quality of Life

Psychosocial adjustment to the diagnosis and treatment of cancer is another area in which gender issues are likely to play a role in cancer outcomes. Until recently, almost all of the relevant literature has assessed quality of life among women, with the few studies with men focusing on physical symptoms and functional abilities rather than the emotional or social aspects of adjustment (Meyerowitz and Hart 1995). Therefore, it is not possible to draw clear conclusions regarding the role of gender in adjustment, and further research is required (see *Chronic Illness: Quality of Life*).

6. Conclusions

Clearly, values and standards regarding the appropriate behaviors and roles for men and women, both at the level of the individual and the larger society, influence cancer onset and progression. The precise nature of the influence is likely to vary by cancer site and by culture.

Relatively little research has considered the biological, behavioral, and healthcare pathways through which gender plays a role. Methodological and conceptual refinements are needed in order to allow future research to expand beyond simple description to

consider multiple causal pathways within a multidisciplinary biopsychosocial model.

See also: Biopsychology and Health; Cancer-prone Personality, Type C; Cancer: Psychosocial Aspects; Gender and Cardiovascular Health; Gender and Health Care; Gender and Physical Health; Men's Health; Women's Health

Bibliography

- Boyle P, Maisonneuve P 1998 Cancer prevention in women: Clues from epidemiology. In: Kavanagh J J, Singletary S E, Einhorn N, DePetrillo A D (eds.) *Cancer in Women*. Blackwell, Malden, MA
- Cohen S, Herbert T B 1996 Health psychology: Psychological factors and physical disease from the perspective of human psychoneuroimmunology. *Annual Review of Psychology* **47**: 113-42
- Good G E, Sherrod N B, Dillon M G 2000 Masculine gender role stressors and men's health. In: Eisler R M, Hersen M (eds.) *Handbook of Gender, Culture, and Health*. Erlbaum, Mahwah, NJ
- Harras A, Edwards B K, Blot W J, Ries L A G 1996 *Cancer Rates and Risks* (NIH Publication No. 96-961). US Department of Health and Human Services, Public Health Service, and National Institutes of Health, National Cancer Institute, Bethesda, MD
- Krieger N, Avery B, Rowley D L, Phillips M T, Herman A A 1993 Racism, sexism, and social class: Implications for studies of health, disease, and well-being. *American Journal of Preventative Medicine* **9** (6 Suppl.): 82-122
- Legato M J 1998 Women's health: Not for women only. *International Journal of Fertility* **43**(2): 65-72
- Meyerowitz B E, Bull A A, Perez M A 2000 Cancers common in women. In: Eisler R M, Hersen M (eds.) *Handbook of Gender, Culture, and Health*. Erlbaum, Mahwah, NJ
- Meyerowitz B E, Hart S 1995 Women and cancer: Have assumptions about women limited our research agenda? In: Stanton A L, Gallant S J (eds.) *The Psychology of Women's Health: Progress and Challenges in Research and Application*. American Psychological Association, Washington, DC
- Parkin D M, Muir C S, Whelan S L, Gao Y T, Ferlay J, Powell J 1992 *Cancer Incidence in Five Continents*, Vol. VI. IARC Scientific Publication No. 120. World Health Organization, International Agency for Research on Cancer, Lyon, France
- Sechzer J A, Rabinowitz V C, Denmark F L, McGinn M F, Weeks B M, Wilkens C L 1994 Sex and gender bias in animal research and in clinical studies of cancer, cardiovascular disease, and depression. *Annals of the New York Academy of Sciences* **736**: 21-48

B. E. Meyerowitz

Gender and Cardiovascular Health

Despite the decline in heart disease in most Western industrialized countries, the disease remains a major health problem and has reached epidemic proportions

in many countries of Eastern Europe, particularly among men (Bobak and Marmot 1996, Weidner et al., in press). Typically, the category heart disease includes nonrheumatic heart disease, hypertension, and coronary heart disease (CHD). The latter comprises the major subgroup (at least 80 percent of heart disease). The focus of this article will be on this major subgroup of heart disease (see *Coronary Heart Disease (CHD): Psychosocial Aspects*).

1. The Gender Gap in CHD

Data from the Framingham Heart Study indicate that the lifetime risk of developing CHD at age 40 is one in two for men and one in three for women (Lloyd-Jones et al. 1999). Recent international CHD mortality data confirm that CHD is lower in women than in men, yielding female:male ratios ranging from 0.1 to 0.6, depending on country (Jackson et al. 1997). At first glance, this universal female advantage in countries differing in heart disease rates and lifestyles points to protective factors of a biological nature (i.e., estrogen) (see *Gender and Physical Health*). However, several studies have cast doubt on the role of sex hormones in the protection from CHD in women (Barrett-Conner 1997). Furthermore, the differences between countries are larger than the difference between the sexes, suggesting that 'male anatomy is not destiny' with regard to CHD. Also, male excess mortality from cardiovascular disease has not always been present throughout history (Nikiforov and Mamaev 1998). Finally, the recent epidemic of cardiovascular disease among young and middle-aged Eastern European men has widened the gender gap over a very brief time span. All these phenomena indicate that nongenetic factors influence the gender gap in heart disease (Weidner 1998).

2. Why Do Men Get More Heart Disease?

2.1 The Role of Traditional Risk Factors

Considerable progress has been made in the identification of physiological and behavioral risk factors for CHD. The major risks are elevated plasma cholesterol, low HDL-cholesterol, hypertension, diabetes, cigarette smoking, and abdominal obesity. Generally, men's levels of these risk factors are less favorable than those of women's (Barrett-Conner 1997). This raises the question of whether gender differences in traditional risk factors can explain the observed gender difference in CHD mortality. Analyses of data from the World Health Organization coordinated MONICA Project and the US Atherosclerosis Risk in Communities Study indicate that only 40 percent of the variation in the gender ratios of CHD mortality in 24 countries could be explained by gender differences

in five major CHD risk factors (i.e., high plasma cholesterol, low HDL-cholesterol, high blood pressure, cigarette smoking, and general obesity; Jackson et al. 1997).

In this regard, it is of interest to note that the traditional risk factors are not good candidates to account for the cardiovascular disease epidemic in Eastern European men either. For example, Kristenson et al. (1998) compared traditional and psychosocial risk factors of 50-year-old Swedish men with age-matched Lithuanian men, who are four times more likely to die from CHD than their Swedish counterparts. There were hardly any group differences in traditional risk factors. However, Lithuanian men were more likely to report greater psychosocial strain (e.g., depression, low social support, job strain) than Swedish men. This raises the possibility that psychosocial coronary risk factors do not only contribute to the recent increase in CHD among men in Eastern Europe, but may also account for some of the variance in the gender differential in heart disease.

2.2 The Role of Psychosocial Risk Factors

The impact of psychosocial factors on the etiology and prognosis of coronary heart disease has been summarized recently (Hemingway and Marmot 1999, Rozanski et al. 1999). These factors include coronary-prone personality attributes such as Type A and hostility, job strain, lack of social support, negative emotions (particularly depression), and low socioeconomic status. Gender-specific associations with CHD for these variables have been summarized previously (Weidner 1995, Weidner and Mueller 2000, Orth-Gomer and Chesney 1997, respectively).

Overall, men appear to be at a disadvantage when considering psychosocial risk factors. For example, they score higher on Type A behaviors and hostility than women (see *Coronary-prone Behavior, Type A*). Both of these attributes are characteristics of the male ('macho') gender role, which has been linked to behavioral risk factors, such as smoking and excessive alcohol consumption (Waldron 1997) as well as other health risks (Courtenay 2000).

Men not only report less social support than women, but also appear to have fewer sources of social support, thus increasing their dependency on a single source (see *Social Support and Stress*). The Male Aging and Women's Health Studies in Massachusetts examined sources of social support among mid-aged women and men, including support from spouse, relative, friend, and co-worker. Men were more than twice as likely as women to name their spouse or their partner as their primary provider of social support (65.5 percent vs. 26.4 percent). Furthermore, 24.2 percent of men (but only 6.1 percent of women) said this was their only source of support! These data may, in part, explain why men's health is more seriously affected by partner

loss through separation, divorce, or widowhood. Finally, men were twice as likely as women to report no supporters (9.2 percent vs. 4.3 percent; New England Research Institutes 1997).

At first glance, gender differences in negative emotions appear to favor men (see *Emotions and Health*). In most studies, women report more negative emotions such as depression than men (although this is not consistently found in populations where women and men have similar roles, such as among college students). Although women may report more depression, they appear to be coping more effectively than men. Generally, men are more likely to use avoidant coping strategies, such as denial and distraction, whereas women are more likely to employ vigilant coping strategies, paying attention to the stressor and its psychological and somatic consequences (Weidner and Collins 1993). Which style is more adaptive depends largely upon the situation. Most stressful experiences consist of uncontrollable daily hassles, which are short-lived and typically of no great consequence. Here, avoidant strategies would be more adaptive ('what I cannot control and what can't hurt me, is best to be ignored'). Thus, men's strategies are likely to pay off for these types of events, contributing to their lesser experience (or report) of emotional discomfort or distress. But what if disaster hits? Avoidance of uncontrollable events—such as divorce, loss of a job or a loved one, sudden financial crisis and economic decline—may not be adaptive in the long run. Here it may be women's greater vigilance that is more adaptive: preparing for the crisis, seeking help, advice, etc. Consistent with this reasoning are recent data from the Hungarian population showing that women tend to accept their negative mood as a disorder to be treated, whereas men are more likely to engage in self-destructive behavior, such as excessive alcohol consumption (Rethelyi et al., in press). Similarly, research on how people cope with disasters (e.g., hurricanes and tornadoes) supports the notion of men's maladaptive coping: increases in alcohol consumption and depression were related to personal disaster exposure among men, whereas no such direct relationship was evident among women (Solomon et al. 1987). Thus, in addition to an unfavorable traditional coronary risk factor profile, men's psychosocial coronary risk factors seem to further enhance their susceptibility to CHD.

2.3 Gender-specific Associations Between Risk Factors

Another major question concerns the interrelationship among traditional, behavioral, and psychosocial risk factors. It is conceivable that gender-specific associations among risk factors may be stronger among men than among women, that is, an increase in one factor may elevate another factor in a gender-specific fashion.

There is some indication that this indeed may be the case. For example, job strain appears to be associated with increases in health-damaging behaviors (e.g., cigarette smoking, excessive alcohol and coffee consumption, lack of exercise) among men, but not among women (Weidner et al. 1997). Similarly, men respond with exaggerated cardiovascular reactions to a wider range of environmental stressors than women. A male ('macho') sex role orientation is associated with decreased motivation to learn stress management skills in men, but not in women (Sieverding 1997). Socio-economic deprivation and depression are more closely related to mortality in men than in women (Kopp et al., in press).

In contrast, men appear to benefit more from social support (i.e., decreased cortisol response to stress) provided by their partner than do women (Kirschbaum et al. 1995, also see Orth-Gomer and Chesney 1997). Similarly, the association between physical activity and CHD appears to be stronger for men than for women. This latter finding, however, may be due to the use of questionnaires designed for men, which focus on sports and neglect physical activities associated with house work (Barrett-Connor 1997).

3. Conclusions

Women score better on traditional and psychosocial coronary risk factors when compared with men (see *Men's Health; Women's Health*). There is also some indication that associations between risk factors may be stronger among men than among women. While gender differences in traditional risk factors appear to contribute substantially to the gender gap in heart disease, we currently do not know to what extent psychosocial risk factors contribute to this gap. Most population studies do not include detailed assessments of psychosocial coronary risk factors, which make analyses of the sort conducted by Jackson and his colleagues impossible. It is likely that gender differences in psychosocial risk factors will explain additional variance in the gender ratios of CHD mortality. Furthermore, currently ongoing research on 'new' coronary risk factors, such as abdominal obesity, cardiovascular stress reactivity, and mechanisms related to antioxidant state in women and men promises to further illuminate factors contributing to the gender gap in heart disease (Weidner et al., in press).

See also: Coronary Heart Disease (CHD): Psychosocial Aspects; Coronary-prone Behavior, Type A; Emotions and Health; Gender and Health Care; Gender and Physical Health; Gender Role Stress and Health; Men's Health; Social Support and Health; Social Support and Stress; Stress and Coping Theories; Women's Health

Bibliography

- Barrett-Connor E 1997 Sex differences in coronary heart disease: Why are women so superior? The 1995 Ancel Keys Lecture. *Circulation* **95**: 252–64
- Bobak M, Marmot M 1996 East–West mortality divide and its potential explanations: Proposed research agenda. *British Medical Journal* **312**: 421–25
- Courtenay W 2000 Constructions of masculinity and their influence on men's well-being: A theory of gender and health. *Social Science & Medicine* **50**: 1385–1401
- Hemingway H, Marmot M 1999 Psychosocial factors in the aetiology and prognosis of coronary heart disease: Systematic review of prospective cohort studies. *British Medical Journal* **318**: 1460–7
- Jackson R, Chambless L, Higgins M, Kuulasmaa K, Wijnberg L, Williams D 1997 Gender differences in ischemic heart disease and risk factors in 46 communities: An ecologic analysis. *Cardiovascular Risk Factors* **7**: 43–54
- Kirschbaum C, Klauer T, Filipp S-H, Hellhammer D H 1995 Sex-specific effects of social support on cortisol and subjective responses to acute psychological stress. *Psychosomatic Medicine* **57**: 23–31
- Kopp M S, Skrabski A, Szekely A 2001 Risk factors and, inequality in relation to morbidity and mortality in a changing society. In: Weidner G, Kopp M S, Kristenson M (eds.) *Heart Disease: Environment, Stress, and Gender*. IOS Press, Amsterdam
- Kristenson M, Kucinskiene Z, Bergdahl B, Calkauskas H, Urmonas V, Orth-Gomer K 1998 Increased psychosocial strain in Lithuanian versus Swedish men: The LiViciordia Study. *Psychosomatic Medicine* **60**: 277–82
- Lloyd-Jones D M, Larson M G, Beiser A, Levy D 1999 Lifetime risk of developing coronary heart disease. *The Lancet* **353**: 89–92
- New England Research Institutes 1997 Gender differences in social supports: Data from the Massachusetts Male Aging Study and the Massachusetts Women's Health Study. *Network Spring/Summer*: 12
- Nikiforov S V, Mamaev V B 1998 The development of sex differences in cardiovascular disease mortality: A historical perspective. *American Journal of Public Health* **88**: 1348–53
- Orth-Gomer K, Chesney M A 1997 Social stress/strain and heart disease in women. In: Julian D G, Wenger N K (eds.) *Women and Heart Disease*. Martin Dunitz, London pp. 407–20
- Rethelyi J M, Purebl G, Kopp M S (in press) Sociodemographic and behavioral correlates of depression in Hungarian women and men. In: Weidner G, Kopp M S, Kristenson M (eds.) *Heart Disease: Environment, Stress, and Gender*. IOS Press, Amsterdam
- Rozanski A, Blumenthal J A, Kaplan J 1999 Impact of psychological factors on the pathogenesis of cardiovascular disease and implications for therapy. *Circulation* **99**: 2192–2217
- Sieverding M 1997 Die bedeutung von prototype-matching für präventives verhalten: ist die teilnahme an stressbewältigungskursen, unmännlich? ('The significance of prototype matching for preventive behavior: Is participation in stress-coping courses, unmanly?') *Zeitschrift für Gesundheitspsychologie* **5**: 272–89
- Solomon S D, Smith E M, Robins L N, Fishbach R L 1987 Social involvement as a mediator of disaster-induced stress. *Journal of Applied Social Psychology* **17**: 1092–1112
- Waldron I 1997 Changing gender roles and gender differences in health behavior. In: Gochman D S (ed.) *Handbook of Health*

- Research I: Personal and Social Determinants*. Plenum, New York
- Weidner G 1995 Personality and heart disease in women: Past research and future directions. *Zeitschrift für Gesundheitspsychologie* 1: 4–23
- Weidner G 1998 Gender gap in health decline in East Europe. *Nature* 395: 835
- Weidner G, Boughal T, Connor S L, Pieper C, Mendell N R 1997 The relationship of job strain to standard coronary risk factors in women and men of the Family Heart Study. *Health Psychology* 16: 239–47
- Weidner G, Collins R L 1993 Gender, coping, and health. In: Krohne H W (ed.) *Attention and Avoidance*. Hogrefe & Huber, Seattle, WA pp. 241–65
- Weidner G, Kopp M S, Kristenson M (eds.) 2001 *Heart Disease: Environment, Stress, and Gender*. IOS Press, Amsterdam
- Weidner G, Mueller H 2000 Emotions and heart disease. In: Goldman M B, Hatch M C (eds.) *Women and Health*. Academic Press, San Diego, CA pp. 789–96

G. Weidner

Gender and Environment

The environment is very much a product of the interplay between societies and the natural world they inhabit. Simply by existing, human beings affect and modify this natural world and render it useful (or otherwise) to society. However, society is notoriously segmented and it follows that different sectors of society will experience the environment differently and will have different degrees of control and influence over it. Entries on environment and development, environmental justice, sustainable development, and environmental social movements explore how differences in race, ethnicity, class, and state of development shape individuals' and communities' relationships with the environment. This section extends this analysis to focus on how gender influences that relationship.

Society is powerfully structured by gender relations that conspire to render women less powerful than men. (The entry on gender defines what the term means and the importance of recognizing its social construction.) These social relations, which are explained variously by reference to class and to patriarchy, result in different, and hierarchical roles for, and opportunities available to, women and men, and structure their relationship with the environment, as well as with each other. Whilst these relationships have persisted, little has changed in structure, if not magnitude, through time. They have only been made explicit in the latter part of the twentieth century, largely due to intersecting debates on development, ecology, and feminism since the 1970s. This article will explore these intersections to show how gender affects the relationships between individuals and communities and the environment, at a variety of scales.

1. Feminist Research

Feminist scholarship has exposed inequalities between men and women which result in an uneven distribution of resources. This relationship has consigned women to the domestic (unpaid) sphere, where they enjoy little economic and political independence. This sphere is notoriously undervalued, with no recognition given to the economic and social value of unpaid work in national accounting procedures (Waring 1988). However, while women are increasingly entering the paid labor force, they do so in a multiply disadvantaged position, which has implications for their relationship with the environment. Most women are paid less, pro rata, than men, through a system of horizontal (through the gendering of different sectors, such as mining or teaching) and vertical (through the gendering of jobs within sectors, such as clerical or managers) occupational segregation, as well as through promotion inequalities. In addition, women engage in this work (full-time and part-time) in addition to their domestic responsibilities, which men have been less willing to share.

Class is often used to explain economic and occupational inequality, and Marxist analysis suggests that women have formed a reserve army of labor which is maneuvered in and out of the labor market in tune with demand. However, this explanation does not sufficiently explain inequalities in the home, nor do they explain systematic male violence towards, and control over women; factors which might better be accounted for by patriarchy. These two systems of domination (class and patriarchy), however, are increasingly being seen as interacting and mutually reinforcing (McDowell 1986).

1.1 Implications for Gender–Environment Relations

This 'double bind' of relatively low paid work outside the home, and responsibility for most domestic tasks within it, exposes women to a number of environmental problems. People on low incomes and in relative poverty are more likely to be exposed to high levels of pollution and other negative environmental effects (see *Environmental Justice*), and since women comprise a majority of adults in poverty, they are more likely to experience environmental problems. For example, DDT, banned in the West, is still used in 2001 as a pesticide in the developing world where women undertake the majority of agricultural work (80 percent worldwide—although women hold title to only 1 percent of the world's land). This chemical accumulates in body fat and is passed from mother to her breastfed children causing, amongst other things, breast cancer, endometriosis and reduced fertility in female offspring. Research in parts of Central Asia found that the Republic of Karahalpakistan (close to the heavily polluted Aral Sea) experienced the highest levels of infant mortality in the former Soviet Union.

Poor diet and hygiene, concomitant with poverty, were found to be contributory factors to mortality and infant anemia (Gabizon 1998).

DDT and other hormone disrupters provide a disturbingly clear example of how environmental manipulation can have a profound effect on the nature of sex itself. Estrogen (produced naturally, by the excretion of estrogen-based contraceptives, but most intensely by PCBs, surfactants used in cleaning materials, and phthalates used in the manufacture of flexible plastics) is being associated with a major drop in fertility and an increase in reproductive disorders. Transsexuality and hermaphroditism in fish is being considered as a direct result of estrogen pollution.

The task of social reproduction accorded to women (unlike the paid tasks of production) are given no value or recognition by national economies (for example, in GNP or GDP) or by the United Nations, which specifies components of national accounting systems (Waring 1988). (Gross national product (GNP) comprises gross domestic product (GDP) plus net income from abroad. GDP is an agdmonetary measure of the value, at market prices, of goods and services produced within a national economy over a given period of time, usually one year (Johnston et al. 2000)). Yet such tasks are made increasingly difficult by environmental exploitation. In the 'Global South' or Third World, in addition to undertaking most of the agricultural work, women are the main water collectors, which can involve journeys on foot of up to four hours each day. Practices that deplete water sources (such as cash crop irrigation) increase this time and energy burden on women, as does deforestation (for commercial farming, mineral extraction, and infrastructure development) for women who can spend four to five hours each day collecting fuel wood when once they would have done so every four or five days (Dankleman and Davidson 1988). In addition, women are carrying increasingly heavy loads of up to 35 kilograms, which is up to 15 kilograms over the acceptable limit laid down by the International Labor Organization Maximum Weight Convention.

In the South, women are exclusively responsible for food preparation, on which they can spend four hours each day. Cooking with firewood (the most common cooking fuel in rural areas) causes inhalation of chemicals equivalent to inhaling the smoke of twenty packs of cigarettes a day. Pollution is even greater when cooks are forced to rely on firewood substitutes, such as husks, straw, and dung, because of the commercial degradation of forests.

In the West, while women do not have the burden of collecting fuel sources and water, they are still primarily responsible for food provision and preparation and, therefore, are more directly concerned with food safety issues such as genetically modified ingredients, BSE, and pesticide use. This is illustrated by the proportion of local environmental and health

campaigners who are women. (Joni Seager (1993) estimates that 60–80 percent of grass-roots environmental activists are women, although this proportion drops drastically for paid/professional appointments in environmental nongovernmental organizations [NGOs].)

In addition to social roles, cultural practices create an additional burden on women in societies in which sex can determine survival. The Chinese practice of limiting families to one child in an attempt to keep within the state's environmental carrying capacity, has resulted in a high female fetus abortion rate and the abandonment of girl children (Bandarage 1997). Likewise, in food deprivation situations Dreze and Sen (1989) have found that women are most likely to go without food. Other research has noted that women are also least likely to receive warm clothing and medical assistance and has confirmed that even pregnant and lactating mothers are last in food priorities (Bandarage 1997).

Because women have largely been confined to the domestic sphere, they have had little formal input into the way in which societies' relationship with the environment has been structured. The practice of science has been particularly masculine, both in that it is undertaken by men, and in its underlying principles aiming to control nature, although this approach is increasingly challenged with science. Even at the start of the twenty-first century, the majority of scientists and science students are men and, it could be argued, this continues to shape the subjects studied (for example, military applications) and the methodologies used (Harding 1986, Haraway 1988).

Other environmental professions, such as planning, civil engineering and architecture, political decision making fora, and business, continue to be male-dominated and masculinist and so women have few opportunities for challenging or contributing to decisions made which have profound implications for the environment itself, and for gender–environment relations, which are then perpetuated.

2. *Ecofeminism*

Feminist politics and scholarship have widely aired aspects of this debate, particularly around inequalities of income and opportunity, the uneven division of labor, and the lack of regard for unpaid work. Linking this feminist debate with environmental issues, and their differential impacts, is the ecofeminist movement that emerged in the 1970s out of the parallel social movements of feminism, ecology, and peace/nuclear disarmament. Its initiators were concerned that economic and political gender equality, without due regard for ecological concerns, would result in doubling the burden the environment would have to bear as women emulated men's behavior. The incorporation of women into 'a man's world' would be unlikely to

challenge research priorities, management practices, social values, or consumption patterns. Instead, ecofeminists argued, society needs to reassess its practices and values and particularly to value the more feminine principles of cooperation and care.

Although the term 'ecofeminism' was coined by a Frenchwoman, Francoise d'Eubonne, the thinking behind it has developed across Europe, North America, and Australasia as both an analysis of socio-economic relations and a prescription as to how these relations can be transformed (Mary Mellor (1992) Carolyn Merchant (1996) and Val Plumwood (1993) respectively). There are two broad strands of ecofeminism: cultural or essentialist (which tend to be more enthusiastically pursued in North America) and social or constructivist (which dominates European thinking).

Cultural ecofeminism celebrates a powerful and positive link between women and nature, arguing that women are more closely in touch with nature through their reproductive potential or capacity. This connection is then thought to better place women (as opposed to men) as advocates for nature. Social ecofeminists do not believe that the essence of what is female (if such an essence actually exists) makes women inherently closer to nature than men. However, they argue that because women are subjugated by a society dominated by men (which also dominates nature), this experience of domination, and consignment to roles and jobs which are less abstracted from nature, and by their dominated status which exposes them more to environmental hazards, places them closer to nature than men who, conversely, are more likely to take on roles and jobs which distance themselves from nature.

These positions, both broadly underpinned by conceptual frameworks based on patriarchy, have led to a debate about the legitimacy of essentialism (Biehl 1991), the need to transcend dualisms of maleness versus femaleness to create a partnership ethic with nature (Merchant 1996, Plumwood 1993), the need to distinguish between different groups of women, and to challenge the unity of their experience (Jackson 1993).

3. Gender and Development

A third key influence on the development of gender and environment as an important relationship has emerged from the Development literature and experience, some of which has already been referred to. This is gradually replacing the arguably tokenistic 'women in development' approach to development project planning. The 'women in development' approach tended to 'bolt' women's concerns on to existing policies rather than questioning these policies and the complex social interactions between women and men which underpinned them. This inevitably affects environment/development projects. Taking these social interactions into account has resulted in a

number of projects and initiatives, such as the Grameen Bank and other microcredit schemes primarily for women in poverty, which also have the potential to improve society-environment relations. Such microcredit schemes are being identified, not only as economically empowering for women, but as giving women more control over their lives in other ways, such as fertility control. Such experiences are slowly penetrating organizations which have the power to impact on society/gender relations, such as The World Bank and the United Nations, and are beginning to affect their lending policies and practices.

4. The United Nations and Sustainable Development

The United Nations, particularly through the *United Nations Conference on Environment and Development* in 1992 and the *World Conference on Women* in 1995, argues that there is a close relationship between the status of women and the state of the environment. It has signaled that the education of women, and the eradication of poverty (targeted at women) are both critical for improving women's status and to reduce societies' impact on the environment. *Agenda 21*, agreed at UNCED 1992, dedicated a chapter (Chap. 24) to the need to bring women into environmental decision-making. This is because women are disproportionately affected by negative environmental impacts due to their social and domestic roles and their greater tendency to be in poverty. UNCED acknowledge that these roles lead to a closer relationship with, and knowledge of the environment, whether as farmers, food and meal providers, or primary health carers. There is, however, little evidence as yet that there are any substantial changes in decision-making locally as a result of this global agreement.

5. Conclusion

Any discussion on gender and environment demands an understanding of the ways in which society is structured along gendered lines. In turn, this requires an understanding of how society creates, reproduces, and mobilizes gender differences. The 'gender and environment' relationship, therefore, is a contested concept which turns on various understandings of gender and gender relations and brings into question the ability to totalize the concept of gender. Whilst this entry has focused on the modernist explanatory concepts of class, and particularly, patriarchy as underpinning gender-environment relations, these relations can be further problematized by recognizing the difficulties in defining gender as a single, undifferentiated, entity. Some critics of ecofeminism draw on the literature of post-modernism to question the validity of treating all men and all women similarly, arguing that class, race, ethnicity, age, disability,

sexuality, and parenthood all cut across gender to render it inadequate as a single explanatory tool. It is important to remember the differences within as well as between genders. Research consistently shows that, while there are clearly differences in income, occupation, and opportunity depending on other factors, the relative differences between most men and most women in many societies are still significant enough to warrant concern about gender inequality and therefore the gendered impact of, and on, environmental issues.

See also: Development: Sustainable Agriculture; Ecology, Cultural; Economic Development and Women; Environment and Common Property Institutions; Environment and Development; Environmental and Resource Management; Environmental Cognition, Perception, and Attitudes; Environmental Justice; Environmental Planning; Environmental Psychology: Overview; Feminist Theory: Ecofeminist and Cultural Feminist; Gender and Place; Human–Environment Relationships; Land Rights and Gender; Social Movements: Environmental Movements; Social Stratification; Sustainable Development

Bibliography

- Bandarage A 1997 *Women, Population and Global Crisis*. Zed Books, London
- Biehl J 1991 *Rethinking Ecofeminist Politics*. South End Press, Boston, USA
- Buckingham-Hatfield S 2000 *Gender and Environment*. Routledge, London
- Dankleman I, Davidson M 1988 *Women and the Environment in the Third World, Alliance for the Future*. Earthscan in association with IUCN, London
- Dreze J, Sen A 1989 *Hunger and Public Action*. Oxford University Press, Oxford, UK
- Gabizon S 1998 A dying sea and a dying people. In: *Women, Health and Environment*. Women in Europe for a Common Future. Utrecht, The Netherlands
- Haraway D 1988 Situated knowledges: the science question in feminism and the privilege of partial perspective. In: Keller E F, Longino H E (eds.) 1996 *Feminism and Science*. Oxford, University Press, Oxford, UK
- Harding S D 1986 *The Science Question in Feminism*. Cornell University Press, Ithaca, NY
- Jackson C 1993 Women/nature or gender/history? A critique of ecofeminist 'Development', *Journal of Peasant Studies* 20(3): 389–419
- Johnston R J, Gregory D, Pratt G, Watts M 2000 *The Dictionary of Human Geography*. Blackwell, Oxford, UK
- McDowell L 1986 Beyond patriarchy: a class-based explanation of women's subordination. *Antipode* 18: 311
- Mellor M 1992 *Breaking the Boundaries, Towards a Feminist Green Socialism*. Virago Press, London
- Merchant C 1996 *Earthcare. Women and the Environment*. Routledge, London
- Plumwood V 1993 *Feminism and the Mastery of Nature*. Routledge, London
- Seager J 1993 *Earth Follies: Feminism, Politics and the Environment*. Earthscan, London

United Nations Commission on Environment and Development 1992 *Agenda 21*. UNCED, Geneva, Switzerland

Waring M 1988 *Counting for Nothing, What Men Value and What Women are Worth*. Allen & Unwin, Wellington, New Zealand

S. Buckingham-Hatfield

Gender and Feminist Studies

'Gender,' as the term is used in the social and behavioral sciences, refers to all of the ways in which being a male or female, or being socially classified as such, affects one's life. The focus is on the social, political, or economic arrangements that relate to gender. Some gender differences stem from biological differences, and are studied by biologists or others focusing on evolutionary theories. Many facets of gender inequality flow from social norms, beliefs, laws, and institutional practices. Even gender differences rooted in biology will have different consequences depending on the social environment. For example, the fact that children are born from women's bodies is a biological reality, but societies and time periods vary greatly in the consequences of this fact. The consequences depend upon whether birth control is widely available, whether women can use contraception without men's permission, how much fathers participate in the care of children, how much the state collectivizes the costs of rearing children, the extent to which jobs accommodate the demands of parenting, and other factors.

The topic of gender was largely ignored in the social sciences until the 1970s. When relationships and inequalities between men and women were studied in earlier decades, many scholars 'naturalized' existing social patterns. This was true even among social scientists who advocated social explanations for most other forms of social differentiation and inequality. Since 1970 there has been a flowering of scholarship on gender. The increased interest in the topic flows in part from the increased representation of women among social scientists; although there are both women and men among prominent scholars of gender, proportionately more women than men have dedicated research careers to the topic of gender.

The term 'feminist studies' has many meanings in the social sciences and humanities. Some proponents (as well as some opponents) of feminist studies see the term to imply adopting women's point of view. Opponents believe this implies that feminist scholarship is more biased against finding the truth than conventional scholarship. In contrast, most feminist scholars believe that scholarship on gender, men, and women has been, and often still is, distorted by androcentric (male-centered) bias. Some people see the appropriate corrective as an adoption of women's

standpoint; those who view it this way doubt that there is any way the scientific method can insure objectivity. Others see feminist studies as correcting androcentric bias, not by adopting women's point of view, but by taking a less biased and broader view than conventional scholarship.

Feminist theory is a similarly broad and contested term. (See *Feminist Theory*) It generally refers to theories of women's experience and of the subordination of women by men. It is commonly taken to include both what philosophers call 'positive' and 'normative' claims. The positive claims are descriptions and explanations of how gender systems work; they are the conceptual and theoretical part of empirical gender studies. Normative claims answer philosophical questions about how gender arrangements ought to be; they deal with questions of ethics and social justice. Philosophers (including political theorists) wrote much of the normative feminist theory of recent decades.

This essay discusses three prominent themes in gender and feminist studies. Scholars have contributed positive and normative theorizing and empirical scientific research on each. They are illustrative of work in the area rather than being a summary of the field. The three themes are the exclusion of women from well-rewarded positions, the devaluation of activities and qualities associated with women, and the intersection of gender with inequalities of class, race, and nation.

1. *Women's Exclusion from Positions of Power and Resources*

The concentration of men in positions that bring them power and resources is an important theme in gender studies. Women and men typically hold different positions. The fact that relatively few men engage in typically female activities and the fact that relatively few women engage in typically male activities are both constitutive of gender differentiation. However, research has focused on the latter. In most societies, men have monopolized positions that give their incumbents the highest status, power, and rewards. Understanding the forces that keep women out of these positions is crucial for understanding sources of gender inequality. Social scientists from several disciplines have studied determinants of women's entrance into 'male' spheres. Anthropologists studying pre-industrial societies have concluded that cultures that are more competitive and combative are more likely to exclude women from positions of power (Sanday 1981). Economists contend that rising wage rates during the twentieth century increased female employment and lowered fertility, because higher potential wages made employment more lucrative for women (Bergmann 1986). When higher wages are available to women, their incentive to be employed is greater; the opportunity cost of staying at home is higher.

Although women's participation in paid work has increased in all industrial societies; jobs have remained somewhat sex segregated. (See *Sex Segregation at Work*.) Some economists explain the exclusion of women from male jobs in terms of 'statistical discrimination.' This term refers to using an easily observable characteristic such as sex to make inferences about harder-to-measure qualifications. The idea is that employers have some knowledge of average gender differences in the population, and, based on this information, treat all women as if they were like the average woman. For example, if the average woman doesn't have the skills to be an auto mechanic, employers may assume that this is true of every woman, and be unwilling to consider even qualified women. Social psychologists add to this that perceptions may be systematically biased. Most people exaggerate the size of gender differences. Sociologists have focused on other discriminatory barriers to women's entry into jobs typically held by men. These include harassment of women entering male jobs, as well as political maneuvering to ensure policies that will keep women out of their fields. Sociologists, anthropologists, and psychologists point out that cultural beliefs and socialization throughout the life cycle encourage both men and women to enter stereotypic jobs for their own sex. In some periods and places, laws have prohibited women's entry into many positions of power.

Liberal feminist theory focuses on this exclusion of women from well-rewarded positions. (See *Feminist Theory: Liberal*.) This focus follows from the classical liberal value placed on equality of opportunity. The exclusion of women from male spheres is a clear violation of equality of opportunity. In Western thought, philosophers had traditionally justified the exclusion of women from powerful, esteemed positions in the economy and polity by arguing that women are innately inferior to men. Liberal thought locates the basis of rights to equal treatment in rationality and autonomy, seen as the most valuable human qualities. The traditional view saw women as less rational and autonomous than men. Liberal feminists, then, argued for women's equal rights by arguing that women were as capable of rationality and autonomy as men were if given the right socialization and opportunities for education.

A key tenet of liberal thought has been a dichotomy between the public and private spheres, and a belief that, due to individual rights, the state should not intervene in what is private. In this view, justice only applies in the public sphere. But the location of the boundary between public and private is continually under dispute in both theory and real politics (Jaggar 1983). Gender relations in the family were excluded from the public sphere by the early liberal writers (Benhabib 1987). With a few exceptions, early liberal feminists did not dispute this. Rather, they focused on women's equal rights in public matters, such as

education, politics, and careers. Exceptions to this include instances of liberal feminist support for laws against rape or violence by men against their wives, or support for laws mandating that employers give parents the job flexibility they need to combine being parents with employment (Okin 1989).

While part of the liberal tradition is *laissez-faire* and opposes state intervention in the economy, another part emphasizes the need for state action to achieve equality of opportunity by class, race, or gender. It is this second camp that is now called 'liberal' in popular discourse, but both are descendants of classical liberalism. The emphasis in egalitarian liberalism has been on equality under the law, equality of opportunity, or equality of treatment, not on equality of outcomes. The application of the egalitarian strand of liberalism to gender has led to advocacy of laws against discrimination in employment, credit, and educational opportunities. Encouraging more women in elected political positions and in the military has been another priority.

In both politics and research, the emphasis has been on what will get women into male-dominated enclaves. Less consideration has been given to what would increase men's participation in traditionally female spheres such as child rearing, paid caring labor, and other female occupations. Researchers, like the general public, think of egalitarian gender role-change largely in terms of women's movement into male spheres. But this overlooks the fact that women are unlikely to achieve equality in jobs and politics unless men share the burden of household and child-rearing work. But men seldom do; sociological research has shown that even in couples where the woman has a demanding career and earns about as much as the man, men typically perform much less household work than women. (See *Time-use and Gender*) Similarly, occupations cannot desegregate without some men taking up what have been women's jobs as well as more women moving into male fields. The failure to see that gender equality is unlikely unless gender-related change is more of a two-way street may result from an unrecognized devaluation of the traditionally female sphere. Precisely because work and qualities traditionally associated with women have been deprecated, men have had little incentive to enter them. This devaluation of everything associated with women is now a theme in its own right in feminist theorizing and in research on gender.

2. *The Devaluation of Roles and Qualities Associated with Women*

Much of feminist theory contests the claim that women are inferior to men. The liberal feminist tradition, discussed in Sect. 1, contests the claim that women are innately unqualified to fill well-rewarded public roles such as politician, soldier, scholar, manager, or wor-

ker. Liberal feminists point out that many women meet the qualifications for these positions, and many more could if provided with the same education and informal socialization men receive. Some thinkers, called radical or cultural feminists, make a different critique. (See *Feminist Theory: Ecofeminist and Cultural Feminist*; *Feminist Theory: Radical Lesbian*) In their view, the sexism in our society is not just a failure to give women the preparation to live up to the proper standards, or to judge fairly whether they meet these standards. Rather, they argue that the standards themselves are androcentric. They see sexist bias in our notions of what qualities people need to be in important positions, and the values determining which positions are rewarded most. In this view, what is seen as valuable has been socially constructed around what is seen as masculine. Examples of traditionally male characteristics highly valued in our culture include independence, abstract rationality, and repression of emotion. Contesting the valuation of these characteristics over traditionally female characteristics, these feminists ask why patriarchal and liberal views have elevated the spiritual or rational (which they associated with men) over the physical or emotional (which they associate with women)? (See *Rationality and Feminist Thought*) Where patriarchal views revere the bravery of risking one's life in hunting, sport, or war, the cultural feminist reveres nurturing for its preservation of life. Whereas patriarchal thinkers, classical liberals, and some Marxists have revered humans for their domination over nature, cultural feminism reveres harmony with nature.

The traditionally female characteristics that cultural feminists believe our culture has undervalued include nurturing, nonviolence, emotional sensitivity to the feelings of others, unselfishness, kinship with rather than domination of nature, acceptance of our physical bodies, humility, flexibility rather than rigid adherence to abstract principles, and intuition of wholes. People have always benefited from women's practice of these skills and values, but this benefit is seldom acknowledged in patriarchal societies. Rather, these virtues have been seen as weakness, lack of proper individuation, or lack of rationality; and social roles that center on these virtues (such as mothering, paid child care, nursing, teaching) are not allocated high rewards. (Examples of feminist works with this theme are Held 1993, Keller 1986, Nelson 1996, Starhawk 1987, Tuana 1993.)

When feminists argue for a valorization of women's traditional virtues and roles, this is often confused with advocating a return to traditional roles. If we see the value of the work that women have traditionally done as nurturers, this does not imply that women should stay at home and be full-time mothers. It implies that we should respect whoever does nurturing, men or women, encourage the development of these skills in all humans, and esteem and reward those who do the work of caring. Respecting traditionally female work

is entirely consistent with this work being shared by men. If the culture values and rewards traditionally female characteristics, why would it make sense to teach these things only to girls or save these roles for women? The radical and cultural feminists argue that traditionally male values are inappropriate for human behavior by either men or women in either domestic or public roles, at least without greater balance from traditionally feminine values than currently obtains.

Emotional connection and care are often involved in the activities traditionally assigned to women. By contrast, western thought has glorified a 'separative self' that is autonomous. Benhabib (197) traces the ideal of separative autonomy through the liberal political philosophy of Hobbes, Locke, Rousseau, and Kant. All presumed men to be separative and autonomous and presumed women would continue to rear children and provide emotional comfort and sexual satisfaction for men. But the value of women's work went unrecognized, as did the extent to which men were dependent upon women's work rather than autonomous. Such work was seen as 'part of nature' in a metaphysic that denigrated nature. These activities did not count as moral since only exercising autonomy in the public spheres counted as moral. This same bias in favor of individuation is seen in developmental psychology, according to Carol Gilligan (1982). She argued that Freud, Jung, Erikson, Piaget, and Kohlberg all viewed individuation as synonymous with maturation but viewed connection to others as developmentally regressive. They assumed women would do the emotional work of child rearing yet did not acknowledge learning the capacity for intimacy and nurture as part of maturation. This atomistic bias has also distorted neoclassical economics, according to many feminist thinkers (Ferber and Nelson 1993).

Paralleling the theoretical work showing a bias in western thought against qualities and activities associated with women, a body of empirical gender research has documented the low value and reward given to activities associated with women. One example is research associated with the policy issue known as comparable worth. (See *Comparable Worth in Gender Studies*.) Research in sociology, industrial psychology, and economics suggests that the sex composition of jobs exerts an effect on how much employers are willing to pay. If a job is filled largely by women, employers offer a lower wage. That is, after statistically controlling for job characteristics that affect pay, such as requirements for skill or difficult working conditions, researchers find that predominantly female jobs pay less than male jobs (England 1992). Thus, sexism affects earnings not only through keeping women out of high-paying jobs, but also through devaluing the jobs in which women are concentrated.

Sociologists and economists have recently begun to study the part of the economy that involves caring for people. Jobs that entail face-to-face care giving are

especially badly paid for their educational levels. Such jobs include childcare, teaching, counseling, and social work. When children are reared well, there are broad social benefits, including a lower crime rate, a more productive economy, better informed citizens, and better care for the next generation. Many of those who receive these benefits pay nothing to those who did the paid and unpaid work of raising children. This means that the private return that markets will provide for such work will be lower than what is consistent with the broad social return. This is one rationale for government to reward the work more generously, to ensure that enough of these services are supplied, and as a matter of equity. Without this state involvement, and the sort of cultural change that cultural feminists advocate, women are likely to leave care work as other opportunities open up, or those who remain in it will pay a high price (in foregone wages) for doing this work. Whether as mothers or as workers in care work, women bear a disproportionate share of the costs of rearing children (Folbre 1994).

Many modern nations collectivize some of the costs of rearing children, through public financing of education and health care. Most European nations also provide cash family allowances to all families with children. The motivation for these programs was typically pronatalist rather than feminist, but they reduce the extent to which women bear the costs of privately rearing children, and in this sense are in women's interest. However, these subsidies to those caring for children are a relatively small part of the modern welfare state. Scholars studying gender and the welfare state have pointed out that many governments have developed two tiers to their benefits. A 'masculine' set of programs is geared to individuals who make claims as earned rights based on past employment. The 'feminine' programs support single mothers when a male breadwinner isn't present; the programs were seen to compensate for family failures. Underlying the two-tiered benefit system are gender norms that define the home as a female sphere and outside work as a male sphere. Men rely on social insurance schemes for retirement income, disability, and unemployment insurance, and women can make claims on these programs too when they are employed regularly or as wives or widows of earners. But single women with children are segregated into 'female' income-tested benefits, and these programs are more stigmatized and pay much less. (See *Social Welfare Policies and Gender*.) Women are expected to care for children, but this work does not provide as respected or rewarded an entitlement to state help in a time of need as a history of employment. More generally, feminists studying the state have argued that the seemingly gender-neutral notion of citizenship is in fact gendered; it has valorized men's typical activities such as worker and soldier, while failing to give rights and privileges for the parenting work that mothers do.

3. *Intersections of Gender with Class, Race, and Nation*

In the 1970s, gender scholars often wrote as if the subordination of women to men was universal. Depending on which theory one chose, patriarchy was caused by the public/private split, men's ownership of property, employment discrimination, or a combative culture. The idea was that once we found the correct explanation of patriarchy, this would reveal how women could be liberated from patriarchy. So it seemed to many of the upper-middle class white feminists in affluent nations, which were the groups predominating in the women's movement and in academic gender and feminist studies. A challenge to this view emerged from women of color in Western nations and in poorer Southern nations. (See *Feminist Theory and Women of Color; Race and Gender Intersections*.)

These critics argued that the analyses of gender proposed by affluent, white, Western women distorted the way gender affected the lives of many women in the world. The argument was not merely that, for example, the life of an affluent, married white professional woman in America is very different from that of an African-American welfare mother, or a Mexican immigrant to the US, or a middle-class Pakistani woman. That is obvious, since the latter three groups lack privilege on several fronts other than their status as women. The argument was deeper—that 'gender' operates differently for these different groups. The claim was that gender interacts with other axes of subordination. (See, for example, Anzaldúa 1990, Collins 1990, Hooks 1989, Lorde 1984, Stoler 1996.)

For example, ideas constructed by men of a dominant racial group about womanhood in fact often create one conception of woman for members of their racial group and another for the subordinate group. In the American South, during and after slavery, white women were seen as chaste and in need of protection (especially from black men). Yet white men did not see black women as needing or deserving this chivalrous protection. They were seen as naturally sexual. Such a construction served to justify white men's interest in having a group of women who had little power to resist sexual coercion and few claims to white men's help in supporting any resulting children. Here clearly gender and race interacted. Gender oppression differed by race. White middle-class women found themselves limited to the role of asexual, married mother, while women of color were unprotected from sexual exploitation by white men. Similar patterns emerged with the European colonial expansion to North and South American, Africa, and Asia.

Hegemonic notions of manhood often impact men and women very differently depending on their class and race, even when there is agreement across class and race groups on what it means to be a man. When someone says 'Be a man!' the implied notion of

masculinity revolves around power. This notion, while certainly not universal, is shared by men in most class and ethnic groups in modern societies. But men in the dominant ethnic groups have easier access to socially accepted ways of getting power, such as success in school and the labor market. This combination of cultural notions of masculinity as power with class-based or race-based lack of access to 'legitimate' sources of power creates a systemic strain making it likely that poor men will seek power through risky and illegal behavior. The gender-related nature of class and race privilege is illustrated by the fact that what men of less privileged class or race groups may envy about privileged men is that the latter often have power over women of all race and class groups. Women in subordinate class or racial groups may sympathize with the sense of powerlessness that men of their class or race feel. Yet they may also see how disadvantaged they are by the fact that they are the only actors in the social system less powerful than men of their race or class are.

Interactions of gender and nation can lead to complex patterns of identity and collective action. Consider, for example, the Islamic practice of women wearing a veil or the practice of female genital mutilation practiced in certain parts of the world. Clearly these practices are linked to male dominance. Yet, in some settings where indigenous people resent Western imperialism, which has harmed both women and men, women as well as men may see these practices to symbolize nationalist solidarity against what is seen as the decadence of Western individualism and greed. (See *Female Genital Mutilation; Islam and Gender*.)

The lesson from scholars of these intersections is that to understand many facets of the gender system, it needs to be studied in combination with axes such as race, ethnicity, class, sexual orientation, and nation. Yet this can sometimes lead to a rich mosaic of particulars that threaten to make gender and feminist studies descriptive rather than theoretical and explanatory. The challenge is to study gender in a way that does not falsely universalize, while looking beyond the unique aspects of each situation to search for broad patterns that help us understand what explains continuity and change in gender.

See also: Androgyny; Beauvoir, Simone de (1908–86); Bernard, Jessie (1903–96); Boserup, Ester (1910–99); Economic Development and Women; Family and Gender; Feminist Economics; Feminist Epistemology; Feminist Ethics; Feminist Legal Theory; Feminist Movements; Feminist Political Ecology; Feminist Theology; Feminist Theory; Feminist Theory and Women of Color; Feminist Theory: Liberal; Feminist Theory: Marxist and Socialist; Feminist Theory: Postmodern; Feminist Theory: Psychoanalytic; Feminist Theory: Radical Lesbian; Gender and Feminist Studies in Anthropology; Gender and Feminist Stu-

dies in Economics; Gender and Feminist Studies in Geography; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Gender and Place; Gender, Economics of; Gender, Feminism, and Sexuality in Archaeological Studies; Gender History; Gender Ideology: Cross-cultural Aspects; Klein, Melanie (1882–1960); Male Dominance; Masculinities and Femininities; Mead, Margaret (1901–78); Multicultural Feminism: Cultural Concerns; Population Dynamics: Two-sex Demographic Models; Poverty and Gender in Affluent Nations; Poverty and Gender in Developing Nations; Race and Gender Intersections; Sex Differences in Pay; Sexuality and Gender

Bibliography

- Anzaldúa G 1990 *Making Face, Making Soul*. Aunt Lute Foundation Books, San Francisco, CA
- Benhabib S 1987 The generalized and the concrete other: The Kohlberg-Gilligan controversy and feminist theory. In: Benhabib S, Cornell D (eds.) *Feminism as Critique: On the Politics of Gender*. University of Minnesota Press, Minneapolis, MN, pp. 77–95
- Bergmann B 1986 *The Economic Emergence of Women*. Basic Books, New York
- Chodorow N 1978 *The Reproduction of Mothering*. University of California Press, Berkeley, CA
- Collins PH 1990 *Black Feminist Thought: Knowledge, Consciousness, and Empowerment*. Unwin Hyman, Boston, MA
- England P 1992 *Comparable Worth: Theories and Evidence*. Aldine de Gruyter, New York
- Ferber M, Nelson J A 1993 *Beyond Economic Man: Feminist Theory and Economics*. University of Chicago Press, Chicago, IL
- Folbre N 1994 *Who Pays for the Kids: Gender and the Structures of Constraint*. Routledge, New York
- Gilligan C 1982 *In a Different Voice: Psychological Theory and Women's Development*. Harvard University Press, Cambridge, MA
- Held V 1993 *Feminist Morality: Transforming Culture, Society, and Politics*. University of Chicago, Chicago, IL
- Hooks B 1989 *Talking Back: Thinking Feminist, Thinking Black*. South End Press, Boston, MA
- Jaggar A 1983 *Feminist Politics and Human Nature*. Rowman & Allanheld, Totowa, NJ
- Keller C 1986 *From a Broken Web: Separation, Sexism, and Self*. Beacon Press, Boston, MA
- Lorde A 1984 *Sister Outsider: Essays and Speeches*. Crossings Press, Trumansburg, NY
- Nelson, J 1996 *Feminism, Objectivity, and Economics*. Routledge, London
- Okin SM 1989 *Justice, Gender, and the Family*. Basic Books, New York
- Sanday P R 1981 *Female Power and Male Dominance*. Cambridge University Press, Cambridge, UK
- Starhawk 1987 *Truth or Dare: Encounters with Power, Authority, and Mystery*. Harper & Row, San Francisco, CA

Stoler A 1996 Carnal knowledge and imperial power: Gender, race and morality in colonial Asia. In: Scott JW (ed.) *Feminism and History*. Oxford University Press, New York, pp. 209–66

Tuana N 1993 *The Less Noble Sex: Scientific, Religious, and Philosophical Conceptions of Woman's Nature*. Indiana University Press, Bloomington, IN

P. England

Gender and Feminist Studies in Anthropology

The development of studies of gender in anthropology received its impetus from the emergence during the late 1960s and early 1970s in Europe and the USA of what became known as the second wave feminism. This article primarily will review theoretical challenges and critical perspectives contributed by feminist anthropology to both the field of anthropology and feminist studies. The article will address main areas of feminist research and scholarly production such as reproduction, colonialism and postcoloniality, labor and transnational capital, Third World feminisms, and feminist ethnographic writing. Finally, the article will address recent work on masculinity.

1. In Search of Theoretical Frameworks

In anthropology, gender difference had played a crucial role in the theoretical development of the field providing especially during its early history a powerful metaphor for thinking about questions of cultural difference. Many characteristics attributed to 'primitive' societies in the armchair anthropology of the late nineteenth century resembled dominant definitions of femininity in Victorian society, such as nonlogical thinking and excess of emotionality. So too gender difference provided Victorian anthropology with an unquestioned 'natural' model for conceptualizing the notion of race and racial difference which was becoming increasingly central to the discipline. As feminist anthropologists came to argue, the position occupied by the notion of gender in the field of anthropology cannot be separated from the politics of colonial expansion and the development of a theory of cultural difference formulated around the notion of race. This linkage between gender, race, and colonial politics has become an important arena of contestation within contemporary feminist anthropology as this article will show. Not only as a metaphor but also as a theoretical premise, gender difference has been a cornerstone of anthropological knowledge as in, for example, Levi-Strauss' structuralist theory of culture.

While gender difference was crucial to the production of anthropological theory, gender was a

category of analysis filled with uncritical assumptions about the naturalness of inequality between the sexes. Gender was that piece of anthropological theorizing that was unproblematic, because it was conceptualized as natural and as such could provide the support for the study of other cultural variables. Such naturalization of gender difference was challenged strongly in the 1970s by a group of young feminist scholars in a series of foundational texts for contemporary feminist anthropology. Gathered predominantly in two volumes (Rosaldo and Lamphere 1974, Reiter 1975) these texts questioned the naturalness of female subordination. They departed from the premise that gender difference and gender inequality were not the product of biology but were culturally constructed, and attempted to formulate theoretical frameworks that could explain 'the universal asymmetry in cultural evaluation of the sexes' (Rosaldo 1974). These first attempts of feminist anthropology at thinking gender difference and sexual inequality argued that women's subordination was not the product of nature but the product of a number of social factors. Most important among these factors were a social organization dividing private from public domains and the cultural ideology that assigns importance to the public over the private and associates women with the latter and men with the former (Ortner 1974, Rosaldo 1974).

The cultural ideology that sustained this division of private/public social domains was aided by an economic division of labor that undervalues women's work in the reproduction of the species and labor force (Sacks 1974). It was also aided by a family structure that encouraged independence in boys and responsibility for care giving in girls (Chodorow 1974). In their search for theoretical frameworks to study gender inequality, the feminist anthropologists of the 1970s engaged the major social theories of twentieth century—structuralism, Marxism, and psychoanalysis—challenging their gender presuppositions and opening new and exciting avenues of analysis (Rubin 1975). A number of important feminist ethnographies challenged male centered accounts of culture and through the detail of ethnographic cases they showed women as social actors deploying a variety of strategies in their engagements with power, prestige, economic exchange, and other social fields.

2. Beyond Dichotomies

The emergent and vibrant feminist anthropology of the 1970s rested on some assumptions. Much of its theorizing was predicated on crucial dichotomies such as private/public, nature/culture, male/female. Gender was understood primarily in unitary terms as an identity all women and all men shared. At issue was the power difference between male and female. A universalist *Woman* rather than women was the object

of analysis. At the turn of the 1980s the theoretical dualities characterizing feminist theorizing of gender began to be questioned and with them an essentialist conception of *Woman*. This challenge was characterized by an emphasis on history, meaning, and power in the feminist studies of the early 1980s (Ortner and Whitehead 1981, MacCormack and Strathern 1980, Bourque and Warren 1981). Some scholars studied the change in meaning of seemingly unchanging categories such as nature and culture so instrumental in the construction of gender, and argued that gender as a category needed to be examined historically. Another strand of theorizing called attention to power difference among women. New research on the impact of colonialism on women argued that the oppression of women was complicated by their position vis a vis colonial power (Leacock 1980). Thus, while European women could be oppressed as women within a gender hierarchy, they were the agents of colonial domination of native men and women. Feminist studies of colonialism also emphasized the different forms in which colonialism as system of oppression changed the realities of women in ways that were different from those experienced by men. Gender as a system of power relations was inseparable from other relations of power. To complicate things more, another strand of research noted the versatility of gender as sign and symbol (Scott 1988). These researchers affirmed that gender had been historically used to express, conceptualize, or advance other kinds of struggles, whether these were about citizenship, nationalism, or race. The point emphasized by this corpus of research was that gender could not be taken for granted as self-evident reality but had to be examined within a multiplicity of social, political, and ideological contexts as well as in historical perspective. In a path breaking essay greatly influential in the anthropology of gender Joan Scott offered a definition of gender that was both comprehensive and complex, bringing together the various foci of research. Scott (1988, pp. 42–7) defined gender as:

(a) A constitutive element of social relations based on perceived differences between the sexes which involved four interrelated elements:

(i) Culturally available symbols that evoke multiple (often contradictory) representations.

(ii) Normative concepts used for interpreting the meanings of symbols.

(iii) A notion of politics and reference to social institutions and organizations necessary to understand how the normative concepts are used.

(iv) Subjective identity which entail the question of how is gender identity constructed in relation to a range of activities, social organizations and historically specific cultural representations.

(b) Gender is a primary way of signifying relations of power. In other words, gender is a primary field within which or by means of which power is articulated. This understanding of gender as signifier of power

was very influential in feminist ethnographies of colonialism, nationalism and violence. For Scott, concepts of gender structure, not only perception of social life, but its concrete and symbolic organization. Sex related differences between bodies are summoned as legitimization of social relations and phenomena that have nothing to do with sexuality.

In anthropology the study of gender challenged the more canonic fields of social structure, power relations, and symbolic constructions. Gender was not one area of analysis but one that forced a rethinking of the thematic pillars of the discipline.

3. New Explorations in Contentious Domains

During the 1980s there was an explosion of very creative feminist ethnography. This scholarly production was less interested in filling the gap of knowledge about women in the discipline than in the study of women as complex social actors engaged in economic, social, and political fields. An important area of feminist production was kinship. Kinship and reproduction, which had long been a privileged area of research in anthropology, became an important domain of feminist ethnography. Feminist ethnographers challenged male centered analysis of kinship by focusing on the role of women in maintaining or transforming kinship ideologies and structures. They also looked at the ways in which kinship structures and ideologies placed women in disadvantageous positions. Some researchers argued that not only gender but also social class, ethnicity, and sexual orientation had an effect in the way kinship and family were perceived and structured. Studies of kinship among poor African-Americans questioned the theory that the high number of women headed families in this social group was a result of a culture of poverty. Instead, these studies argued that complex kinship networks, articulated around women, played a crucial role in ensuring economic and social survival of large numbers of people in a situation of chronic economic disadvantage (Stack 1974). Research on gay and lesbian families in the Bay area of San Francisco challenged the dominant thesis of an American kinship system based on the ideology of blood connections by mapping the ways in which gay and lesbian people constructed family networks and ideologies not based on blood connection (Weston 1991). In England, feminist analysis of kinship was linked to the emergence of new reproduction technologies and their effect on nationalist ideologies (Strathern 1992). These studies argued that new reproductive technologies were changing not only ideas about women and families, but also ideas about what constituted nature and culture. The association of women with nature that had its support on women's reproductive capacity was challenged with the development of new technologies of reproduction that made reproduction a

highly technological affair but also allowed for a complex array of family forms not based on biological connection. Feminist anthropologists studying reproduction also focused on the ways in which women's bodies were represented in the discourse of biomedicine and the ways in which this discourse excluded women from the control of their own bodies. Approaching science and bio-medicine as cultural discourse some researchers inquired about the gender and race assumptions underlying medical discourses.

Another important area of research during the 1980s was the study of the effects of transnational capital on women. One strand of research focused on the transformations of the role of women in the family and community triggered by the introduction of transnational manufacture in traditional societies of the Third World. Some researchers studying the new international division of labor triggered by transnational capital, argued that multinationals was producing a class of poor, working-class women worldwide. Other researchers focused on the use of traditional beliefs and rituals such as spirit possession to manage the stress and alienation of factory production and capitalist discipline (Ong 1987).

The anthropological research on transnational capital brought to the anthropological agenda new issues about the interrelation between gender and social class, the transformation of traditional societies, and the linkages between new forms of sexual violence and capitalist expansion in the Third World. During the 1990s, the growing industry of sexual tourism has become a new important area of research.

A renewed interest during these years on history and change within anthropology, as well as on the labor politics of transnational capital fueled research on the neglected areas of colonialism and nationalism. Some researchers turned to history to study the cultural assumptions about gender, sexuality, and race that colonial officers and soldiers brought with them to the colonial context (Stoler 1991). A central preoccupation of these studies has been the study of the role of tradition in colonial societies. Tradition is important because it has often been associated to gender identity and used for a variety of purposes. For example, feminist studies of the practice of Sati (the ritual burning of widows in India) show that the practice was subject to a number of contingencies and hardly generalized in precolonial India. The English, however, took this practice as an example of a barbarian tradition that had to be abolished to bring the country to the English level of civilization. In this case an argument concerning violence against women was furnished to legitimize a colonial power relation. In so doing a practice which was not tradition became formalized as such. Once Sati was understood and treated by the British as Indian tradition it began to be associated by some nationalist strands with Indian national identity as well as with a feature of gender relations (Mani 1990). Similar processes concerning

tradition and ethnic identity were at play in colonial Africa in relation to practices of female circumcision. Traditions concerning gender relations produced by colonial scholarship and colonial policy became later on an emblem of national identity. The consequence, almost everywhere, has been a greater control of women and an increase on violence against women. This charged area surrounding the definition of cultural tradition and its relation to ethnic identity has become a central part of contemporary discussions about immigration, race, and citizenship in European countries and the USA. These discussions have been dramatized in the courtrooms and mass media, triggered by charges brought against immigrants performing various forms of genital mutilation on girls. In France, for example, this discussion has confronted feminist activists against cultural specialists. Feminists have argued that female mutilation is a form of violence that violates the right of girls and women to physical integrity and, as an illegal practice, should be prosecuted by the law. Cultural specialists, in turn, have argued that genital mutilation is part of the traditional culture that constitutes the identity of those who practice it. A variety of associated meanings and situations complicate this discussion. At issue is, for example, the fact that those brought to trial are poor immigrants subjected to institutionalized forms of racism in their everyday life. At issue is, too, the forms in which the discourse of human rights has become part of French national identity. The point is that neither 'tradition' nor 'culture' are neutral or self-evident terms, but are terms used to define regulatory practices between men and women and between the state and its people. These practices of power are articulated and symbolized on the bodies of women and particularly Third World women and women of color.

A growing body of feminist anthropology has become preoccupied with the question of how women figure in relation to nationalist identity, nationalist ideologies, and postcolonial nationalist struggles. This line of research has investigated the way nationalist ideologies constrain the lives of women by making them either the repository of national tradition or the passive allegory of the nation. Part of the problem is that colonial discourse often defined colonial subjects in gender codes, attributing them a feminine incapacity for government. To counteract this definition, nationalist movements have often resorted to a reverse gender language that represents the people of the nation as masculine while representing the nation as a female object. When certain women's roles such as motherhood are defined as essential to national identity, such roles are often enforced through legislation, restricting the choice and development of women in other social areas. In postcolonial nations the pressures on women to conform to nationalist demands become complicated by the fact that the new nation represents freedom from the bonds of coloni-

alism. Thus, to rebel against its constraints can be perceived and experienced as national disloyalty. These positions create double bind-like dilemmas for women whose aspirations go against the grain of the national project (Aretxaga 1997).

Notwithstanding its constraints, for women engaged in struggles of liberation such as Palestinian women, nationalism can open a space of social and personal liberation. Research on women in nationalist movements shows that political activism has the paradoxical effect of constraining and liberating women. Women often maneuver between contradictory demands and divided loyalties. At the same time, women's activism in struggles of liberation often leads to reformulation of gender relations but research shows that such reformulation is always precarious and shifting (Peteet 1991). The anthropological research of gender and nationalism has demonstrated that gender identity is never defined in isolation from other forms of identity such as class, ethnicity, or nationality. This research has also brought to light the differences among women showing that the very category of women even within an ethnic group is internally divided. The research on women and nationalism has counteracted a dominant ideology that has represented women in the Third World and women of color as passive victims of authoritarian nationalism or political violence.

4. Third World Women and Women of Color

During the late 1980s and early 1990s women of color in the USA and Europe claimed a specific epistemological and political position within the feminist movement. They criticized the feminist mainstream for silencing the voice of minority women and women of color and contended that the emphasis on women's solidarity had obscured relations of power between and among women of different class, ethnicity, and color. This critique was followed by a critique advanced by Third World women who claimed to be in a subordinated position vis-à-vis western women. Both the notion of 'Woman' and 'Western' were the object of severe criticism. If this moment had a tremendous political impact on the feminist movement worldwide, it had an impact on feminist anthropology as well. Most research on gender and women in anthropology had been produced by Western women studying women in Third World countries or minority women in Europe or the USA. The critique of imperialism launched against white mainstream feminism by Third World feminists cut right across feminist anthropology. The result among feminist anthropologists was a self-critical reflection that emphasized the awareness of the position of power the anthropologist held vis-à-vis her subject and how such position influenced the production of ethnographic knowledge. Feminist knowledge in anthropology became 'situated

knowledge' rather than unmarked universal knowledge. There was a critique of representation and an emphasis on the production of 'subjugated knowledge' (Haraway 1991). Feminist anthropologists turned their attention to questions of experience and self constitution as generating ethnographic knowledge. There was a new emphasis on the role of subjectivity in scholarship which characterized the 1990s. Theoretical production about hybridity, diversity, and partial knowledge rather than universal generalization was at the center of the feminist anthropology of the 1990s.

5. *Writing Gender*

The attempt at presenting rather than representing the voice of minority, subjugated women led feminist anthropologists to experiment with the form of ethnographic writing. Some anthropologists reflected on the difficulty of capturing the subjectivity of women in contexts where women only spoke through the conventions of male oratory. Others found that conventions of ethnographic representation were fundamentally male oriented and inadequate to narrate women's experience. Feminist ethnographers experimented with literary genres and strategies to assert a new form of writing culture suitable to convey hitherto invisible women's experience. Fiction, journal writing, first person narrative, and autobiographical narrative became part of feminist ethnographic writing during the 1990s. Some ethnographers experimented with transcribing women's narratives with minimal interference of analytic voice (Behar and Gordon 1995).

Others juxtaposed the biographical history of the ethnographer to the autobiography of the women they wrote into books, yet others resorted to mixed fictional forms to express the dilemmas of feminist ethnography in a postmodern, fragmented world divided by relations of power and increasingly hybrid. Feminist ethnographies reflected not only on the position of the women they wrote about but on the hybridity of the authors, themselves positioned oddly between the West and the Third World.

For these feminist anthropologists, ethnography in the contemporary world becomes a kind of fiction writing. Their ethnographic stories are composed of stories told in the field by informants, often informally. They freeze a fragment of social life and a fragment of history out of the ongoing unfolding of events. In ethnography, like in fiction, closure is always arbitrary and the goal for feminist anthropology is to capture the dilemmas, complexities, and uncertainties of women's experience. The kinds of questions that traverse experimental feminist ethnography gravitate towards the ambiguities of identity, community, the meaning of home, and fieldwork. In writing about other women, feminist ethnographers reflect on their problematic identity and the kind of home and community that made them feminist ethnographers.

6. *The Anthropology of Masculinity*

Although anthropology had focused traditionally on men, these were represented as universal subjects rather than men whose masculinities were culturally constituted in a variety of ways. During the late 1980s and 1990s, a growing number of ethnographic works have focused on the cultural construction of masculinity. Ethnographies of masculinity have broadened the anthropology of gender often associated with women's studies. However, as Guttman (1997) notes there has been an awkward avoidance of feminist theory by those anthropologists studying manhood. The early anthropology of masculinity focused on how men performed their manhood in a variety of cultural contexts. More recent work analyzes masculinity in relation to a broader number of themes. Guttman (1997) notes several topics of interest in relation to which masculinity has been discussed by anthropologists, such as nationalism, divisions of labor, family, friendship, sexuality, and struggles of power, yet there are no general theories of masculinity produced from anthropology. Part of the problem is the difficulty in defining masculinity which has been taken so much as the norm that it remains conceptually elusive in its specificity. While a number of works have explored performances of manhood in relation to other men, women, and the national identity (Herzfeld 1985), an important contribution has come from the discussion of men's subjective perceptions and feelings about other men.

7. *Conclusion*

Since the 1970s, the anthropology of gender has grown into a complex, diverse, and theoretically sophisticated field. It has gone from a focus on women's studies to include increasing study of masculinity across cultures. Much of the anthropology of gender falls into the rubric of feminist studies and this has posed political as well as epistemological challenges that have resulted in a sophisticated, mature, and rich body of scholarship.

See also: Family and Gender; Family Theory; Feminist-Economist Critique; Female Genital Mutilation; Feminist Theory; Gender and Feminist Studies; Gender and Feminist Studies in Sociology; Gender Differences in Personality and Social Behavior; Gender Ideology: Cross-cultural Aspects; Male Dominance; Masculinities and Femininities; Nation-states, Nationalism, and Gender

Bibliography

Aretxaga B 1997 *Shattering Silence: Women, Nationalism and Political Subjectivity in Northern Ireland*. Princeton University Press, Princeton, NJ

- Behar R, Gordon D A (eds.) 1995 *Women Writing Culture*. University of California Press, Berkeley, CA
- Etienne M, Leacock E (eds.) 1980 *Women and Colonization: Anthropological Perspectives*. Praeger, New York
- Guttman M C 1997 Trafficking in men: The anthropology of masculinity. *Annual Reviews of Anthropology* 26: 385–409
- Haraway D J 1991 *Simians, Cyborgs, and Women: The Reinvention of Nature*. Routledge, New York
- Herzfeld M 1985 *The Poetics of Manhood: Contest and Identity in a Cretan Mountain Village*. Princeton University Press, Princeton, NJ
- MacCormack C P, Strathern M (eds.) 1980 *Nature, Culture and Gender*. Cambridge University Press, New York
- Ong A 1987 *Spirits of Resistance and Capitalist Discipline: Factory Women in Malaysia*. State University of New York Press, Albany, NY
- Ortner S B 1974 Is female to male as nature is to culture. In: Rosaldo M Z, Lamphere L (eds.) *Woman, Culture and Society*. Stanford University Press, Stanford, CA
- Ortner S B, Whitehead H 1981 *Sexual Meanings: The Cultural Construction of Gender and Sexuality*. Cambridge University Press, New York
- Peteet J M 1991 *Gender in Crisis: Women and the Palestinian Resistance Movement*. Columbia University Press, New York
- Reiter R R (ed.) 1975 *Toward an Anthropology of Women*. Monthly Review Press, New York
- Rosaldo M Z 1974 Woman, culture, and society: A theoretical overview. In: Rosaldo M Z, Lamphere L (eds.) *Woman, Culture and Society*. Stanford University Press, Stanford, CA
- Rosaldo M Z, Lamphere L (eds.) 1974 *Woman, Culture and Society*. Stanford University Press, Stanford, CA
- Rubin G 1975 The traffic in women: Notes on the 'Political Economy' of sex. In: Reiter R (ed.) *Toward an Anthropology of Women*. Monthly Review Press, New York
- Sacks K 1974 Engels revisited: Women, the organization of production and private property. In: Rosaldo M Z, Lamphere L (eds.) *Woman, Culture and Society*. Stanford University Press, Stanford, CA
- Scott J W 1988 *Gender and the Politics of History*. Columbia University Press, New York
- Stack C B 1974 *All our Kin*. Harper & Row, New York
- Stoler A L 1991 Carnal knowledge and imperial power: Gender, race and morality in colonial Asia. In: di Leonardo M (ed.) *Gender at the Crossroads of Knowledge: Feminist Anthropology in the Postmodern Era*. University of California Press, Berkeley, CA
- Strathern M 1992 *Reproducing the Future: Essays on Anthropology, Kinship and the New Reproductive Technologies*. Manchester University Press, Manchester, UK
- Warren K B, Bourque S C 1981 *Women of the Andes: Patriarchy and Social Change in Two Peruvian Towns*. Michigan University Press, Ann Arbor, MI
- Weston K 1991 *Families We Choose: Lesbians, Gays, Kinship*. Columbia University Press, New York

B. Aretxaga

Gender and Feminist Studies in Economics

A feminist approach to economics recognizes the importance of gender in structuring human experi-

ence, including economic life, and seeks economic knowledge for the purpose of combating gender-based oppression. While precursors of such feminist studies go back a century or more, feminist economics did not emerge as a body of thought among academic economists until the early 1990s. The ensuing decade brought an expansion of feminist scholarship not only about women's (and men's) economic status, but also about the gendered structure of economics as a discipline, itself. Can a discipline that seeks for reliable knowledge about economic life *not* be feminist?

1. Defining the Terms

For a branch of scholarship to be considered 'feminist,' it should recognize the importance of gender in structuring human experience, and should be part of a liberatory project to eliminate sexist structures that keep women in inferior and submissive positions relative to men. Work that is simply about women, or by women, but that does not challenge structures of domination, is not feminist in this sense (and hence will not be discussed here). Feminist theories include liberal, Marxist, radical, and other variants (see *Feminist Theory: Ecofeminist and Cultural Feminist; Feminist Theory; Feminist Theory: Marxist and Socialist; Feminist Theory: Psychoanalytic; Feminist Theory: Liberal; Feminist Theory: Postmodern; Feminist Theory: Radical Lesbian*).

Academic economics includes schools of neoclassical, Marxist, institutionalist, post-Keynesian, and Austrian thought. The neoclassical school, however, is clearly the 'mainstream,' as evidenced by its absolute hold on the highest-status institutions and their graduate and research programs, as well as its ubiquitous presence elsewhere. The neoclassical school is paradigmatically characterized by centrality of the question of how self-interested individual agents make rational choices when confronted with scarcity, by a strong belief in the allocative efficiency of markets, by a near-sole reliance on mathematical reasoning as the means of creating knowledge, and by the application of statistical testing to large-scale, often government-collected, data. This neoclassical definition of 'economics' is itself at issue in feminist studies, as will be discussed below. Because of the hegemony of this one school, this article will emphasize work that grows out of and/or reacts against it, although feminist economic studies can also be found within the institutionally weaker heterodox schools, the other social sciences, and popular, nonacademic, economic literature.

Some may doubt that feminist studies have a legitimate place with economics, since they may seem to mix a political agenda into an otherwise value-free, neutral scientific enterprise. Much of the accomplishment of the last decade, however, has been precisely to reveal the value-laden, masculine-biased, and evi-

dence-ignoring aspects of mainstream economic thought, and to begin to envision less biased and more satisfactory means to warranted knowledge.

2. *The Rise of Feminist Studies in Economics*

Feminist studies in economics, as a body of thought, formed in the early 1990s from a convergence of two streams: first, concerns about women's economic status, and second, feminist scholarship on the role of gender in the structuring of science.

Increasingly, in the 1970s and 1980s, concerned women had pointed out the neglect and/or distortion of women's experience in several areas. Marianne Ferber and Bonnie Birnbaum, for example, debated the proposition (put forth by the University of Chicago-inspired 'New Home Economics') that women's exclusive specialization in household work could be explained as their free and advantageous choice (Ferber and Birnbaum 1977). Paula England cast doubt on the empirical validity of the 'human capital' explanation of male–female wage differentials, in which women's lower pay was portrayed as the result of their own utility-maximizing choices of training and occupation (England 1982). Barbara Bergmann highlighted the importance of, and difficulties facing, the movement of women into the labor force (Bergmann 1986). Marilyn Waring pointed out that the United Nations systems of national income and product accounts excluded any consideration of unpaid work by women (Waring 1988). While some scholars took what Sandra Harding has called a 'feminist empiricist' approach, believing that 'social biases [are] correctable by stricter adherence to the existing methodological norms of scientific inquiry' (Harding 1986, p. 24), others began to question the particular forms of scientific inquiry sanctioned by the economics profession. Issues of dependence, interdependence, tradition, social beliefs, power, and oppression did not seem to fit well into mainstream forms of analysis, and feminist scholars began to search for more adequate means with which to understand the fullness of economic life.

During this same period, feminist physical scientists, philosophers, and historians (e.g., Harding) were revealing the gender biases deeply embedded in modern conceptions of science. The identification of objectivity with detachment, and the consequent model of an active, knowing subject exerting control over a passive object of knowledge, are metaphorically connected with images of male/mind activity and female/matter passivity (see *Rationality and Feminist Thought*). Within the economics profession, some attention to the rhetoric of economics (McCloskey 1985) simultaneously helped bring to light the social nature of economic discourse, allowing for the potential discussion of gender influences within the discipline.

Table 1

Contrasts defining neoclassical economics.

Core	Margin
Public	Private
Market	Home
Individual	Social
Efficiency	Equity
Hard	Soft
Fact	Value
Rigor	Intuition
Precise	Vague
Objective	Subjective
Detached	Committed
Self-interest	Altruism
Autonomy	Dependence
Reason	Emotion
Masculine	Feminine

Contrasts such as those shown in Table 1 highlight the gendered construction of neoclassical economics, as those domains, methods, and assumptions chosen to be central to the discipline all have stereotypical masculine associations, while the neglected and denigrated areas are socially considered 'feminine' (see, e.g., McCloskey 1985, Nelson 1992, Jennings 1993). As such a definition of economics can be seen as one manifestation of an identification of feminine associations with inferiority, feminists have sought ways to deconstruct these gender dualisms.

The formation of the International Association for Feminist Economics in 1991 and the publication of *Beyond Economic Man: Feminist Theory and Economics* in 1993 (Ferber and Nelson 1993) signaled the beginning of feminist studies in economics as a body of thought. Continued conferences and the publication of a journal, *Feminist Economics* (started 1995) illustrate the dynamism of the field at the turn of the century.

3. *Major Developments in the Field*

Three major areas have absorbed the attention of many feminist scholars: first, exploration of the definition and methodology of economics; second, analysis of gendered labor market outcomes; and third, inquiries into work traditionally done by women.

3.1 *Definition and Methodology of Economics*

Most feminist work in economics recognizes that the discipline is socially constructed, that is, that it reflects the interests of the people who developed it and the worldviews and prejudices of its formative historical epochs, rather than simply a set of trans-

historical ideals. The dualisms in Table 1, then, reflect social, intellectual, and even psychic biases of the creators of the discipline.

One response might be to think that the point of a 'feminist' economics would be to add more women economists, and, therefore, 'women's ways of knowing.' This emphasis, however, has relatively little support among feminist economists, since it puts the causal story at the level of biology (or sex) rather than at the level of social understandings (gender). Most feminists are less interested in finding 'essential' differences between males and females, and more interested in examining the heavy cultural influences that go into creating observed differences and schemes of categorization. Since, when given the chance, female economists have proven quite adept at the types of thinking required by current disciplinary strictures, a feminist approach looks elsewhere.

The more important point is that the creators of mainstream economics lived in societies characterized by a culture-wide belief in masculine superiority. One graphic tool for working beyond the usual associations between gender and value was proposed by Nelson (1992). A vertical axis marks off the dimension of value, positive (+) or negative (−), while a horizontal axis marks off gender associations, masculine (M) or feminine (F). While sexist thinking recognizes the M+ and F− quadrants, examination of the unusual M− and F+ cells may reveal new possibilities. For example, the idea that economics should be a 'hard' science rests on an identification of hardness with (masculine-associated) strength, and softness with weakness (and femininity). Yet hardness can also mean rigidity (M−), while softness can mean flexibility (F+). Too much effort in economics has been spent on defending the masculine/feminine boundary, and too little on improving the value of economic studies. As a result, the models, methods, and pedagogy of mainstream economics tend to be characterized by rigidity, obsession with mathematical reasoning, detachment from human suffering, and narrow forms of explanation (M−). While recognizing the validity of ambitions to be strong, logical, scientific, and precise (M+), one can simultaneously recognize the value of flexibility, analogical and dialectical reasoning, concern with human life, and richness of analysis (F+).

Avoiding the hyper-rationalist mind/body split implicit in a definition of economics in terms of rational choice, and the marginalization of women's traditional activities implicit in a definition in terms of markets, feminists have suggested a definition of economics as being related to the subject matter of the provisioning of human life. Such a definition in terms of subject matter allows for the use of a larger toolbox of models and methods. Instead of models of individual autonomous in choice, agency is recognized as being limited by physical embodiment and social embeddedness. Rather than identifying knowledge

solely in the forms of theorems and proofs, or statistical tests on second-hand data, the value of more qualitative and even experiential data, and of rigorous attention to the gathering and interpretation of all data, receives increased attention. Objectivity is not guaranteed by strict adherence of an individual researcher to some prescribed model or method, but rather is a characteristic obtainable only through the shared understandings of an expanding knowledge community.

Some feminist economists take this analysis a step further, and, seeing the importance of language and the partiality of positions thought previously to reflect universal truths, conclude that reliable knowledge is not possible. In the extreme deconstructionist postmodern view, the social construction of economics is taken to mean that there is no reality outside of the linguistic structure. While the feminist-postmodernist conversation (see *Feminist Theory: Postmodern*) has influenced feminist studies in the 1990s, such *relativism* is not a necessary consequence of a feminist approach, and most scholars in economics adopt at least a limited form of philosophical realism. That is, while economics is seen as socially constructed, it is also recognized that the external world puts limits on the interpretations that can be supported.

3.2 Gendered Labor Market Outcomes

The movement of women into paid labor in most industrialized countries supplied the raw material for a large literature on gendered occupational distributions and the gap between average male and female wages. Much of the work uses mainstream methods, most notably regression analysis, to determine how much of the wage gap might be attributable to sex *per se*, after taking into account the effects of observable, presumably productivity-related, factors (e.g., education, seniority, etc.). Such analyses typically find these factors to explain about half of the wage gap. Yet mainstream economists who prefer to believe that 'market forces' are stronger than 'social biases' still find ways to avoid interpreting these results as evidence of discriminatory employer behavior. The remaining gap is usually attributed, instead, to variations in some unobserved characteristics of the worker (such as 'effort,' thought to favor men) or at most to 'societal' discrimination that effects the aspirations or dedication that individual workers bring to their (presumably unbiased) employers.

Feminist economists have pointed out the limitations of using a methodology based on an *a priori* faith in markets and a narrow view of economic data (Bergmann 1989). Alternative sources of data such as that available through court cases, interviews, case studies, ethnography, and detailed-level occupational analysis, while less acceptable given mainstream methodological biases, portray a situation of continuing

sex discrimination, and the need for more integrated social and structural analysis of gendered labor market outcomes. Many feminist economic studies along such lines have been done by feminist sociologists (see *Sex Segregation at Work*) and historians.

In studying the newly industrialized countries, feminist scholars have taken a more structural approach, asking how policies of development and trade have created gendered patterns of market work. A rising question in the 1980s was whether the availability of cheap, flexible female labor had led to a pattern of 'feminization' in export-oriented sectors of many of these economies (see *Economic Globalization and Gender*).

Recent work in this area has included more attention to issues of race and sexual preference (see *Lesbians: Social and Economic Situation*) in studies of occupational segregation and wage gaps, and attention to the impact of structural change in industrialized countries.

3.3 Women's Traditional Work

While labor markets are within economists' customary domain of analysis, the historical importance of the nonmarket areas of home and family to women's economic well-being forces an expansion of the domain (Folbre 1994). Taking families as places of interesting economic production and allocation, how can these phenomena be studied? How do home and market life interact? When women move into the labor market, what happens to the tasks that they used to do? Can serious study of traditionally 'feminine' realms also change how we look at markets?

The notion that there could be something called 'home production' filtered into the discipline by the 1970s, usually in reference to constraints affecting women's labor force participation. The early models tended to perpetuate, rather than challenge, gender biases, however, by assuming that women had a comparative advantage in household work, and that the traditional allocations of home and market work reflected rational decisions given unitary 'household' preferences. A few forays into examining the inner workings of households in the early 1980s became a booming literature of bargaining models and intra-household allocation by the late 1990s (Seiz 1991, Agarwal 1997). Recognizing that households are made up of individuals who might have different and potentially conflicting interests, and different well-being outcomes, such models allow at least a limited exploration of the issues of power and perception within households (see *Family Bargaining*). The increased attention within the mainstream to intra-household phenomena may represent the most significant effect on it, so far, of feminist criticism.

In developing countries where formal markets are a smaller part of the economy, women's traditional

work is likely to play an even bigger role in determining well-being (see *Economic Development and Women*). A substantial literature has developed focusing on women's 'reproductive' as well as 'productive' labor (Elson 1991, Benería 1995). In particular, macro-economic and structural adjustment policies have come under feminist scrutiny for their differential impact on women, since the 'reproductive' sector is relied on to cushion the shocks caused by, for example, reduced public spending on health care.

Yet the division between paid and unpaid work is somewhat artificial, and the experiences of most women cannot be adequately analyzed by studying each in isolation. Sex segregation in the labor market, and the relatively low wages paid for nurturing, people-oriented activities (see *Comparable Worth in Gender Studies*), illustrate the way in which traditional expectations of, and domination of, women extends from the household into the market sphere (Hartmann 1976). Women's poverty in industrialized countries is often a combined outcome of low-paying, female-stereotyped jobs and the invisibility of needs for childcare (see *Poverty and Gender in Affluent Nations*). The need for paid childcare (and care of the elderly or ill), while most often studied as a constraint on mothers' labor force participation, is also increasingly looked at in terms of its own characteristics in creating employment—often of women less educated and more likely to be immigrants or of color compared to their employers. The division between paid and unpaid work is often also associated with a stereotypical distinction between presumably self-interested, arms-length relations in the market, and altruistic, intimate relations among family and friends. While the notion of household bargaining stretches this old dualism in the direction of recognizing individual interests within families, recent literature on 'caring labor' (see *Care and Gender*) stretches it in the direction of studying affective and socially significant relationships in the monetized sector. As family and emotional labor becomes less 'invisible' and 'natural,' and more a subject of debate and study, the myth of (male) autonomy becomes more difficult to sustain.

4. The Future of the Field

Spanning a variety of epistemological loyalties, political viewpoints, and geographical continents, feminist work promises to be a dynamic, though not unified, influence on the economics discipline for years to come. With the critique of standard economic schools relatively far along, more scholars are working towards greater development of positive feminist alternatives to the traditional paradigms. Promising routes to the development of a strong alternative to the neoclassical school include cross-fertilization from feminist work in the other social sciences, perhaps in combination with insights from institutional econo-

mics, pragmatist or postanalytical philosophy, and new views of scientific methods arising from developments in the theories of chaos and complexity. Progress is hampered, however, by resistance rooted in the traditional ideologies. While mainstream economists now tend to tolerate more study of women, or the occasional methodological article, or feminist work when it takes place in departments of cultural studies or sociology, the challenges to the core of the discipline make most economics departments and institutes reluctant to provide intellectual or institutional support for feminist studies.

See also: Gender and Feminist Studies; Gender and Feminist Studies in Geography; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Gender, Economics of

Bibliography

- Agarwal B 1997 'Bargaining' and gender relations: Within and beyond the household. *Feminist Economics* 3: 1–51
- Benería L 1995 Toward a greater integration of gender in economics. *World Development* 23: 1839–50
- Bergmann B R 1986 *The Economic Emergence of Women*. Basic Books, New York
- Bergmann B R 1989 Does the market for women's labor need fixing? *Journal of Economic Perspectives* 3: 43–60
- Elson D 1991 *Male Bias in the Development Process*. Manchester University Press, Manchester, UK
- England P 1982 The failure of human capital theory to explain occupational sex segregation. *Journal of Human Resources* 17: 358–70
- Ferber M A, Birnbaum B G 1977. The 'new home economics': Retrospects and prospects. *Journal of Consumer Research* 4: 19–28
- Ferber M A, Nelson J A (eds.) 1993 *Beyond Economic Man: Feminist Theory and Economics*. University of Chicago Press, Chicago
- Folbre N 1994 *Who Pays for the Kids: Gender and the Structures of Constraint*. Routledge, London
- Harding S 1986. *The Science Question in Feminism*. Cornell University Press, Ithaca, NY
- Hartmann H. 1976 Capitalism, patriarchy, and job segregation by sex. *Signs* 1: 137–69
- Jennings A L 1993 Public or private? Institutional economics and feminism. In: Ferber M A, Nelson J A (eds.) *Beyond Economic Man*. University of Chicago Press, Chicago, pp. 111–29
- McCloskey D N 1985 *The Rhetoric of Economics*. University of Wisconsin Press, Madison, WI
- Nelson J A 1992 Gender, metaphor and the definition of economics. *Economics and Philosophy* 8: 103–25
- Seiz J A 1991 The bargaining approach and feminist methodology. *Review of Radical Political Economics* 23: 22–9
- Waring M 1988 *If Women Counted: A New Feminist Economics*. Harper and Row, San Francisco

J. A. Nelson

Gender and Feminist Studies in Geography

Gender studies in geography address ways in which the material and symbolic lives of women and men are shaped by aspects of place and spatial relations and, in turn, how considerations of gender are implicated in the creation of place and space. Central within gender studies are feminist perspectives stemming from concerns about inequalities of power between women and men. In seeking points of contact between feminism and geography, Susan Hanson (1992) has identified three core analytic traditions: finding significance in everyday life; appreciating the importance of context; and thinking about difference. Bringing these traditions together, feminist geography has produced a dynamic scholarship which has introduced new themes about women's lives into the discipline, recognized the importance of space and place in understanding gender, and increasingly elaborated its focus to include the intersections between gender and other social inequalities such as those originating in 'race' and ethnicity, class, age, physical ability, religion, and nationality. The sensitivity to context that informs feminist scholarship implies not only that expressions of gender are shaped by context, but also an understanding that the creation and uses of knowledge are 'situated,' that is, they are closely tied to the cultural and political contexts in which the research is located. Given this position, interpretation of gender and feminist studies in geography has to recognize that the field is not unitary but diverse, incorporating multiple perspectives and ways of knowing.

1. The Initiation of Feminist and Gender Studies in Geography

Geographic research on gender began in the early 1970s as an academic expression of the contemporary women's rights movement and its concerns with equality, a phenomenon recognized internationally with the establishment in 1975 of the United Nations Decade for Women. Early publications came from North America, the UK, and The Netherlands, the first widely known pieces appearing in the radical journal *Antipode* in 1973. This writing reflected its political origins, addressing the unequal status of women in society and also drew attention to women's marginal positions within the geographic profession (Zelinsky et al. 1982). Studies reflected the different contexts in which they were created. Whereas Dutch initiatives emerged from student critiques of Marxist writing and relatively quickly turned to the applied and policy-related work that characterizes much Dutch geography, early research in the USA followed empirical, positivist traditions, documenting differences between women's and men's spatial behavior, for example, in daily travel patterns or migration.

From these beginnings, gender studies in geography have become global in scope (Monk 1994). Formal or informal study groups hold meetings and/or produce newsletters in Argentina, Australia, the UK, Canada, Catalonia, German-speaking countries (Austria, Germany, Switzerland), New Zealand, and perhaps other locations; within the European Union, an Erasmus program has linked feminist geographers in Denmark, Greece, The Netherlands, Spain, and England. A strong international network comes together through the Commission on Gender in the International Geographical Union. Within the latter, symposia were organized as early as 1981, a formal Study Group was approved in 1988, and the Commission was recognized in 1992. It links over 400 scholars in more than 50 countries. Feminist work is well represented in a number of mainstream geographic journals, especially in the English-language literature and in Spain (García-Ramon and Caballe 1998). A specialized journal of feminist geography, *Gender, Place and Culture* was founded in 1994. Reviews of current feminist literature also appear regularly in *Progress in Human Geography*. Nevertheless, the field remains barely visible in some parts of the world, notably eastern Europe and east Asia. The range and diversity of participation, however, means that gender studies address numerous topical themes and that research is conducted in multiple paradigms. Still, there are common strands, primarily the understanding of the concept of gender as a complex and varying social, cultural, historical, and geographic construction, not simply a binary biological distinction between women and men, that power relations are implicated in gender relations, and that space, place, and gender are interdependent.

1.1 Feminist Critiques

Important from the beginning was critique of the extent and nature of the masculine assumptions underlying the content, theories, methods, and purposes of most geographic research (Monk and Hanson 1982). Scholars pointed out that research on migration, for example, generally presented men as key actors and decision-makers and women as passive followers, despite an absence of studies confirming how decisions were made. Data also showed women as numerically dominant in migration streams at particular times and places, for example, in rural to urban migration in many Latin American countries. Urban travel models dealt with the journey to work assuming full-time employment (implicitly male), rather than attending to the more complex patterns of part-time (usually female) workers, or people combining responsibilities for household and paid work. Economic geographers usually studied formal economies in which men were the major participants (such as industrial work or export-oriented agriculture), rather

than the service sector, or the informal or subsistence sectors in which women often worked. Feminist geographers advocated attention to arenas more central to women's lives, and introduced issues such as child care, housework, clerical employment, or women's fears in public spaces.

Methodologically, problems were identified with the limitations of secondary data sources such as censuses for their failure to capture women's lives, with primary data collection practices such as interviewing the male 'head' of the household, and with the use of survey methods that did not offer the subjects of research the opportunities to frame their own experiences. Feminist scholars advocated qualitative and in-depth investigations with the hope of learning how women saw their own worlds and in an attempt to modify hierarchical power relations between the researcher and those being studied.

1.2 Theorizing and Documenting Women's Material Lives

Going beyond critique, early theorizing, especially in British geography, elaborated the concept of patriarchy—social control by men originating in the authority of the father—as the source of women's oppression. Included in the arguments about patriarchy were claims that its intersections with capitalism created dual and interlocking sources of oppression. This line of theorizing was compatible both with the socialist tradition in British radical geography, with which early feminists identified, and with the substantive emphasis in research on women's material experiences, especially in urban settings. Distinctions were drawn between the city, as the site of male paid employment and public privilege, and the suburb as the site of domesticity, a private space that served as a basis for women's secondary status. This work begins to show how feminist geographers link the material world with the symbolic and ideological as they strive to understand gender-based inequalities. Though the theorizing initially emphasized the urban scale, other scales were recognized as important. The research also highlighted the importance of place as well as space in the dynamic conjunction of capitalist and patriarchal processes. An enduring piece representing this perspective is McDowell's and Massey's (1984) analysis of ways in which different forms of economic development in four regions of Britain perpetuated male dominance, even as the specifics of economic patterns changed between the nineteenth and late twentieth centuries.

While the theoretical writing emphasized patriarchal and capitalist ideologies, a great deal of the empirical work was undergirded by related concepts useful for understanding the material aspects of labor. Numerous studies have been devoted to documenting and understanding gender divisions of labor and the

ways in which productive and reproductive work were gendered and interdependent. Attention to women's reproductive work brought new subject matter into the geographic literature, for example, on domestic employment, child care, and community-organizing. Research has been conducted in many geographic contexts, for example, on the work of Spanish rural women which demonstrates marked regional differences in its nature and extent (García-Ramón et al. 1991), and especially in the Third World where scholars have identified relationships between the gendering of work and development. This research has brought into gender studies not only scholars from North America, Europe, and Australasia, but has significantly engaged African and Asian geographers (Momsen and Kinnaird 1993). Documenting the details of daily and seasonal space-time divisions of labor in rural settings, they reveal the substantial economic contributions of women, identify relationships between household structures and labor supply and exploitation, demonstrate how the introduction of contract labor schemes, of the rearrangement of work hours, and of development schemes promoting agroforestry or irrigation have frequently increased women's burdens and fostered changes in household and community relations without improving economic outcomes for local people. Research on service and industrial employment at local scales similarly shows that from Malaysia to Massachusetts and Mexico, contemporary labor recruitment practices are implicitly and explicitly gendered with consequences for regional and international flows of migrants, daily patterns of movement, the organization of household work, and the persistence of gender-based occupational segregation (Cravey 1998, Hanson and Pratt 1995, Momsen and Kinnaird 1993).

2. Gender and Women's Agency

Accompanying the focus on women's oppression, disadvantages, and gender inequalities has been research on women's agency in shaping the spaces and places of their lives. Research in political geography and on environment and development provide good examples. Both exemplify ways in which feminist scholarship has contributed to revealing the linkages between not only production and reproduction but also between 'public' and 'private' spaces, questioning the validity of envisioning these as binary divisions. Studies of political struggles in Latin America (Radcliffe and Westwood 1993) show women invoking their 'private' identities as mothers to enter the public spaces to protest against state violence. These women have transgressed in public spaces, for example, seeking out clandestine cemeteries, taking and publishing photographs of the tortured and 'disappeared,' and visiting prisons. Yet they also incorporate tra-

ditional 'private' caring responsibilities of women into their public projects—organizing health and child care services and providing food and clothing for refugees. Similarly, research on Palestinian women under Israeli occupation argues that their resistance in the public political arena has been sharpened by military incursions into the secluded private spaces of the home, and by their needs to defend their children in the streets (Mayer 1994).

The motivations and tactics of environmental protests have also been interpreted in gendered terms. Women's activism may be stimulated by concerns for the health of their children and families; their public protests may draw on domestic symbols in public protests (for example, the Spanish women who nightly banged their *cacerolas* (saucepans) in protest against toxic waste dumps (Bru, in Rocheleau et al. 1996)). However, the feminist geographic literature on the environment reveals complex relationships. In this vein, Rocheleau et al. (1996) elaborated the concept of feminist political ecology. They demonstrate the importance of gender differences in access to and control over resources; how gendered knowledge is reflected in a science of survival; and the gendered nature of environmental politics and grassroots activism that is linked with women's redefinition of their identities and recognition of new environmental problems.

3. Difference, Identity, and Embodiment

In the 1990s, feminist geography has been strongly influenced by interdisciplinary currents, especially by the linguistic turn in the social sciences which has foregrounded discourse analysis, and by theoretical perspectives emanating from postmodernist, post-colonialist, and 'minority' feminist scholarship. Scholars have questioned grand narratives, the presumed centrality of white, Western feminisms, and the essentialism of the category 'woman.' They have focused on destabilizing dichotomous categories, such as man/woman, public/private, self/other, mind/body, and global/local (Duncan 1996). This work is particularly developed in the English-language literature, though is also attracting attention in, for example, Spanish and German-language feminist geography.

The implications of the differences among women on such bases as 'race,' ethnicity, class, sexuality, life stages, nationality, and so on are explored, as are ways in which identities based in these sources of difference are mutually constituted, situationally fluid, formed through discourse and everyday actions, and vary across specific historical and geographic contexts and spatial scales. Research explores hegemonic processes, resistances to them, and the negotiation of identities. Linda McDowell (1999) synthesizes this work, emphasizing life in post-industrial societies. Her discussion

attends to home and workplaces, the public spaces of the street and pleasure, and the nation-state. She challenges naturalized assumptions about gender and about these spaces, for example, depicting the home as simultaneously a place of safety and a trap for women with young children, a haven after work or a place where relations between gender and age have to be continually negotiated; she identifies public spaces of parks and streets as sites that can be liberatory or generative of fear and danger.

This literature recognizes spaces and places, as well as identities as simultaneously open and bounded, within and across geographic scales. How are 'hybrid' identities created, for example, by immigrants who move across international boundaries to settle in specific local places? The fluidity and contingent nature of identity in relation to place is highlighted. While some work looks to the present, past imperialism has received significant attention as the historical discourses of women's travel narratives are re-read. Geraldine Pratt provides multiple examples of the fluid and contextual expressions of identity (Fincher and Jacobs 1998). Among the cases she presents is that of a Victorian British woman traveller in west Africa who is bound by gender at home but enabled by class and race privileges in Africa; in another example, she explores the shifting value positions on ethnicity and expressions of gender roles elaborated by Canadian couples in a multicultural neighborhood in Vancouver as they struggle with making decisions about caretakers (and care-taking locations) for their children; yet again, she speculates on the impacts on the political values of women in Worcester, Massachusetts, of moving daily between homes that are middle class (reflecting their husbands' status) and workplaces where they are side-by-side with women of working class backgrounds.

Perhaps the most notable departure in the 1990s from previous geographic work is the focus on the scale of the body and on links between sexuality and space, especially homosexuality. Research examines ways in which particular bodies—the pregnant, the homosexual, the disabled, and the non-white—are marked in space often in exclusionary ways, and also how 'performances' of gender and sexuality are place-related (Duncan 1996, McDowell 1999).

While this new work introduces exciting possibilities, cautionary notes are sounded. On the one hand, it links feminist geography with other critical scholarship concerned with challenging oppression (Jones et al. 1997). The recent collection, *Cities of Difference* (Fincher and Jacobs 1998), for example, brings together feminist work on gender (inflected by other sources of difference such as life stage) with comparable studies relating to disability, gay male identity, race/ethnicity, religious convictions, and class. On the other hand, concerns exist that decentering gender could lead to a depoliticized relativism in which a hard-won feminist political stance is

silenced. So long as the research retains its focus on inequalities and power relations, however, and seeks possibilities for collaborations and coalition formation, the development may be more effective in achieving feminist goals than concentration on a unitary, and some would argue exclusionary, identity politics.

4. Reflections on Methods and Methodologies

Critical perspectives on the creation of knowledge have been enduring issues in gender and feminist studies. While some discussions address specific issues of methods, particularly the relative merits of qualitative and quantitative research, the fundamental concerns are with the politics involved in creating and representing understandings of the world and the partiality of all knowledge, in both senses of this word—that is, no story is considered complete or objective. Overviews (McDowell 1999), collected essays (Jones et al. 1997) and special sections of mainstream journals have addressed the theme, such as the series of articles in *The Canadian Geographer* in 1993 on aspects of ethnographic methods, in *The Professional Geographer* in 1994 on dilemmas of field research, and in the same journal in 1995 on appropriate approaches to, and the values of, quantitative work. They highlight issues of power and ethics. How does one set a research agenda/define a research question if the goals include being responsive to the people being researched and contributing to the feminist agenda for social change, while remaining constrained by the expectations and conventions of the academy? Rocheleau (Rocheleau et al. 1996) describes her involvement of rural people in the Dominican Republic in a multi-faceted research project on agro-forestry that incorporated both 'top down' and 'bottom up' generation of questions, categories of analysis, and interpretations. Gibson-Graham reports her use of an interactive, dialogic method in which women in Australian mining towns were involved in implementation of the research, engaging their domestic and community politics (Duncan 1996).

Numerous scholars reflect on power issues of their own relationships with 'subjects' especially as their interviews involve people who are of different class, 'race,' ethnic, gender, or sexual orientation from themselves and who may express views that are problematic for the researcher (for example, considering them to be sexist or racist) or who reveal knowledge that could compromise the subject or researcher (Jones et al. 1997, McDowell 1999). While such questions have been widely raised in interdisciplinary feminist literature, as Hanson (Jones et al. 1997) points out, perhaps the distinctive and most interesting contributions from geography concern the roles of location and place within the research process—in

what settings does the recruitment of 'subjects' take place, where are interviews held (homes, workplaces, community contexts)? How does the choice of context shape not only the way subjects present themselves but how they interpret the researcher's intentions and meanings? In a revealing account, Nagar identifies multiple ways that she was viewed and presented herself in research carried out in religiously, regionally, and ethnically diverse social spaces in Asian communities in Dar-es-Salaam (Jones et al. 1997). Overall, these discussions point to the importance of reflexivity that requires researchers to continually evaluate their assumptions and political positions. They also lead to a conclusion that favors combining multiple methods and exploring ways to engage the 'researched' in the validation of one's work.

Discussions about representation also figure prominently in feminist geography—not only asking whose voices are heard in the research, but also how categories are imposed on what we hear, and on where and how research is reported and read. Pulido, for example, writing on the activist group, Mothers of East L. A. (Los Angeles) indicates that while feminist work has identified women's environmental activism with their gender identities, the women involved may not see themselves and their motivations in gendered terms, but rather in relation to ethnic or community identities (Jones et al. 1997). Monk notes the irony of the hegemony of British and US scholarship in the reviews of literature by feminist geographers who simultaneously promote resistance to hegemonic discourses and listening to a plurality of voices (Jones et al. 1997) while overlooking publications written outside their own milieu. One of the most interesting examples of attempts to present research in ways that will speak to different audiences is Pulsipher's analysis of the diverse interpretive texts she has created to communicate her research in historical geography/archeology on a slave plantation in Montserrat to multiple audiences including local project workers and outside volunteers assisting in the research (and especially to engage locals in defining concepts), as well as museum curators and visitors at the Smithsonian Institution. She has also written several fictional accounts using the research material to convey the bodily feelings of life on the plantation to local schoolchildren, community members, outside visitors, and academic audiences (Jones et al. 1997).

Surprisingly rare in these methodological discussions, especially given the proliferation of cartographic metaphors in contemporary post-modern and feminist writing, are reflections on the place of mapping in feminist geography and gender studies, though several feminist atlases (for example, of women in the USA, the *Atlas of Women and Men in India* (Raju et al. 1999), and especially Seager's (1997) world atlas) present examples of the possibilities and difficulties of representing gender themes cartographically. Huffman's critique of the masculine values of traditional car-

tography, and his reading of feminist endeavors, opens up methodological questions, such as the ways maps and other visual representations can present fragmented data, reveal hidden perspectives, and contribute to interpretations of gendered bodies (Jones et al. 1997).

5. Status and Emerging Directions

In little over two decades, gender and feminist studies have become well established in geography despite their marginal origins. In particular, their critical perspectives have contributed to reshaping epistemological understandings in human geography. In addition to the substantive themes mentioned above, work is expanding on gender and the body (including sexuality and health), on gender and human rights, trans-national cultural, political, and economic processes and their intersections, and constructions of masculinity. There is every indication that gender and feminist studies will remain a vital field within geography.

See also: Feminist Theory; Nation-states, Nationalism, and Gender

Bibliography

- Cravey A J 1998 *Women and Work in Mexico's Maquiladoras*. Rowman and Littlefield, Lanham, MD
- Duncan N (ed.) 1996 *Bodyspace: Destabilizing Geographies of Gender and Sexuality*. Routledge, London
- Fincher R, Jacobs J M 1998 *Cities of Difference*. Guilford Press, New York
- García-Ramón M D, Caballe A 1998 Situating gender geographies: A bibliometric analysis. *Tijdschrift voor Economische en Sociale Geografie* **89**: 210–16
- García-Ramón M D, Cruz Villalón J, Salamaña L, Valdovinos N, Vilariño M 1991 Women and farm households: Regional variations in gender roles and relations in Spain. *Iberian Studies* **20**: 81–112
- Hanson S 1992 Geography and feminism: Worlds in collision? *Annals of the Association of American Geographers* **82**: 569–86
- Hanson S, Pratt G 1995 *Gender, Work, and Space*. Routledge, London
- Jones J P III, Nast H J, Roberts S M 1997 *Thresholds in Feminist Geography: Difference, Methodology, Representation*. Rowman and Littlefield, Lanham, MD
- Mayer T 1994 *Women and the Israeli Occupation: The Politics of Change*. Routledge, London
- McDowell L 1999 *Gender, Identity and Place: Understanding Feminist Geographies*. University of Minnesota Press, Minneapolis, MN
- McDowell L, Massey D 1984 A woman's place? In: Massey D, Allen J (eds.) *Geography Matters*. Cambridge University Press, Cambridge, UK
- Momsen J H, Kinnaird V (eds.) 1993 *Different Places, Different Voices: Gender and Development in Africa, Asia and Latin America*. Routledge, London

- Monk J 1994 Place matters: Comparative international perspectives on feminist geography. *The Professional Geographer* 46: 277–88
- Monk J, Hanson S 1982 On not excluding half of the human in human geography. *The Professional Geographer* 34: 11–23
- Oberhauser A M, Rubinoff D, De Bres K, Mains S, Pope C (in press) Geographic Perspectives on Women Specialty Group. In: Gaile G, Wilmot C (eds.) *Geography in America at the Dawn of the 21st Century*. Oxford University Press, Oxford
- Radcliffe S A, Westwood S (eds.) 1993 *Viva: Women and Popular Protest in Latin America*. Routledge, London
- Raju S, Atkins P J, Kumar N, Townsend J 1999 *Atlas of Women and Men in India*. Kali for Women, New Delhi, India
- Rocheleau D, Thomas-Slayter-B, Wangari E 1996 *Feminist Political Ecology: Global Issues and Local Experiences*. Routledge, London
- Seager J 1997 *The State of Women in the World Atlas*. Penguin Reference, London
- Zelinsky W, Monk J, Hanson S 1982 Women and geography: A review and prospectus. *Progress in Human Geography* 6: 317–66

J. Monk

Gender and Feminist Studies in History

Gender history, feminist history, and women's history identify overlapping areas of inquiry whose relationships to each other have become more complex and more contested in the past decade. Emerging as an autonomous subfield by the early 1970s, women's history was a direct outgrowth of the feminist movement in the USA, Canada, Australia, and Western Europe. Certainly, earlier generations of women and of feminists had written on the historical experiences of their sex, but most did so outside the institutions and practices of professional history (Smith 1998). Only in the 1970s did women historians gain access to professional training and teaching positions in colleges and universities in significant numbers and in a sufficient array of countries to nurture this new field of study. A small number of men participated in launching women's history and a growing number employ gender analysis, but women remain the mainstay of the field.

Feminist history was initially conceived as nearly synonymous with women's history in the sense that most practitioners of the latter embraced the politics of the former, including critiques of patriarchal institutions and practices and of maldistributions of power and resources more generally. But feminist history could also define an area of investigation—the origins and development of movements for women's rights and sex equality. Gender analysis appeared as a distinct but related approach to history in the mid-1980s, though a concern with the 'social relations of the sexes' and the relative status of women and men

had certainly been introduced earlier. This approach both complemented and challenged existing work in women's and feminist history by asserting the critical importance of exploring historical changes in individual and collective subjectivities and the centrality of gender in signifying forms and relationships of power (Kelly 1986, Scott 1988, *Gender & History* 1989).

The emergence of gender history as a distinct area of inquiry paralleled and converged with a number of other theoretical and professional developments that assured the transformation of feminist history. These included the rapid increase in women's history scholars, journals, books, and conferences; the resurgence of conservative and right-wing politics around the world; the rise of postmodern, postcolonial, and subaltern epistemologies; and the critique of early feminist frameworks by scholars of color and by historians of the working class, lesbians, and non-western regions and countries (Brown 1992, Meyerowitz 1992, Alexander and Mohanty 1997, Pierson and Chaudhuri 1998).

1. The Rise of Gender History

From the 1960s on, women's and feminist historians had incorporated some gender perspectives in their work. Not until the mid-1980s, however, did the history of gender become a substantial field in its own right. Proponents of gender analysis argued that 'the creation and reproduction of gender is a process, changing over time' and place, and operating 'at every level, and in every sphere' of historical study (*Gender & History* 1989, pp. 2–3). The ubiquity of gender—as a system of symbols and signifiers as well as a cultural construction with implications for society, the economy and politics—demanded 'a theory of meaning that assumes a multiplicity of references, a resonance beyond literal utterances, a play across topics and spheres' (Scott 1988, p. 60). In addition, gender analysis was seen as a critical weapon in transforming traditional historical narratives. Women's history could be accepted as a parallel but separate area of inquiry, but gender history—which recognized gender 'as a key axis of power in society' and provided 'a crucial understanding of how any society is organized and structured'—demanded the recasting of the past *in toto* (Hall 1992, p. 12).

Responding to earlier challenges that feminist history had long been too white, too middle-class and too heterosexual, many gender historians engaged issues of class, race, nationality, and sexuality from the beginning. They asked how class was gendered, how gender was raced, and how nationality was raced, gendered, and sexualized. Davidoff and Hall, for instance, explored the critical role of women and the family in the formation of the British middle class; Brown suggested that jazz and gumbo ya-ya were appropriate models for capturing the multivocal,

relational character of an American women's history riven with race and class differences; and Guy analyzed the ways that fears of 'white slavery,' that is, the international trade in women for sex, shaped notions of citizenship, and nationality in Argentina (Davidoff and Hall 1987, Brown 1992, Guy 1992)

There was considerable resistance, however, to the emergence of gender analysis among some feminist historians. Scholars' emerging fascination with men and masculinities led some to worry that women would quickly be subsumed within this new gender history. Other questions arose when pioneering advocates of gender analysis embraced postmodernism and deconstruction. Postmodernist critiques of grand narratives and the deconstruction of unified subjects and subjectivities posed challenges to fields like women's and feminist history that were seeking to create new narratives by building on the identity politics of the 1960s and 1970s (*International Labor and Working Class History* 1987). Nor was this concern limited to battles among white women historians. Scholars of color, committed to integrating race and nationality into studies of women and gender, were divided over the utility of postmodernism for creating new political insights and practices (Christian 1987, Alexander and Mohanty 1997). Students of sexuality, too, were alternately attracted to and repelled by theoretical perspectives that denied unitary subjects, homosexual and heterosexual alike (Meyerowitz 1992). Critics focused on the political dangers of destabilizing categories of individual and group identity at a moment when identity politics seemed to promise a potent weapon in the fight against racism, sexism, classism, homophobia, colonialism, and imperialism. In a more purely historical vein, the attack launched by postmodernists threatened not only older grand narratives, but also newer ones that celebrated the lives of women, workers, African Americans, indigenous peoples, and anticolonial activists. Yet despite such concerns, gender analysis has increasingly shaped work in women's and feminist history. And taken together, these approaches have begun to transform the study of the past in several significant ways.

2. *The Transformative Power of Feminist and Gender History*

Certain themes and interpretations provide common threads in the history that is being rewoven through analyses of women and gender. Perhaps most importantly, private life, domesticity, reproduction, and sexuality have been reclaimed as subjects. Not only do we now recognize that these topics have a history that parallels that of public life, politics, production, and war, but we also are beginning to see the ways that these two histories are deeply intertwined—private with public lives, domesticity with politics, reproduction with production, sexuality with war.

In recognizing this broader tapestry, studies of women and gender have transformed the meanings of terms central to historical analysis. Politics no longer refers only to the formal institutions of government and law, but also to mass movements, voluntary associations, maternalist campaigns, spontaneous uprisings, and alternative understandings and conceptualizations of power in workplaces, schools, cultural sites, and homes as well as courts, legislatures, parliaments, and partisan arenas. Work is no longer limited to labor performed by men in exchange for land, goods, or wages, but embraces the production of goods and services by women and men, regardless of payment or the system of exchange, and includes the labor of physical and social reproduction and the contributions of informal as well as formal economies. In a similar fashion, studies of war are no longer confined to the battlefield or the arenas of military strategy and diplomatic negotiation, but are instead open to analyses of the homefront, of low-intensity but chronic conflict, and of rape, disappearances, and torture as weapons of domination.

The history of labor—productive and reproductive—offers numerous examples of how attention to women and gender recasts our understanding of the transition to capitalism. For example, women were critical economic agents in many arenas during the medieval and early modern periods, particularly in domestic production, social reproduction, informal sectors, and petty trade. Bennett's (1996) *Ale, Beer and Brewsters in England*, illustrates the centrality of women to the production and sale of alcoholic beverages in the early medieval period, when brewing was a largely local affair. She then analyzes the process by which female brewsters gradually were pushed to the margins as larger markets, new technologies, government regulation, and guild formation turned brewing into a masculine venture. In this transition to commercial capitalism, women remained tied to local, low-paying, low-status work, while the skills they had honed in home-based shops were taken over by men once they proved sufficiently profitable. In similar fashion, Ulrich's exploration of a midwife's contributions to household sustenance and social interaction in a small town in late eighteenth- and early nineteenth-century Maine illuminate the significance of women's labor in maintaining families and communities and its contributions to larger transformations in the gendered economies in which they were embedded (Ulrich 1990). Studies of slave women across time and place and of female service, factory, and home-based workers in the industrial age add to these early portraits of women's place in both formal and informal economies (Berger 1992, Morgan 1997)

In a similar fashion, the study of politics has been transformed by the inclusion of women and of gendered perspectives. Initially feminist scholars shifted the focus away from electoral, parliamentary, and judicial activities, nearly all of which excluded

women—at least formally—until the last half century. Instead women's historians studied the means by which women and other marginalized groups shaped the political terrain through voluntary associations, petitions, food riots, mass demonstrations, and reform and revolutionary movements. Interpretations of the most significant periods of reform in the USA—the 1830s–1850s, the 1890s–1920s, and the mid-1950s–mid-1970s—have now been rewritten to highlight the public activities of women and their larger social, political, and cultural consequences (Hewitt and Lebsack 1993). Scholars of the French Revolution and the revolutions of 1848 across Europe have highlighted both women's participation and the use of women as symbols, alternately, of liberty and degradation. And interpretations of nationalist movements and national liberation struggles in South Africa, India, Spain, Cuba, and other areas of the world are being rewritten to include women activists and gendered images (Landes 1988, Riot-Sarcey 1994, Pierson and Chaudhuri 1998).

In developing new areas of inquiry and new perspectives on traditional subjects, feminist and gender historians have been joined by historians of other groups long marginalized in historical narratives—racial and national minorities, workers, peasants, slaves, and those ostracized for sexual practices or by sexual traumas. Race, class, and sexuality were never absent from women's history, but only from the mid-1980s on did explorations of differences among women gain equal attention with investigations of differences between women and men (Brown 1992, Alexander and Mohanty 1997, Gordon 1999). In recent years, scholars have developed more and more sophisticated analyses of the interplay among race, class, nationality, and sexuality, with a growing number of feminist historians working to reveal their mutually constitutive character. Even studies of feminism and women's emancipation have been recast in terms of race, class, and nationality. In particular, the ways that many white feminists excluded women of color from their ranks and/or wielded racist or ethnocentric logic to advance their own cause has now been sharply critiqued even as historians of women's movements continue to search for moments of alliance and collaboration across these divides (Burton 1994, Pierson and Chaudhuri 1998).

Many fields have been influenced by this attention to race, class, and nationality within women's history. Research on the history of Africa and the African diaspora has been especially important in excavating the intricate links among women's multiple forms of identity. Gendered images of Africa, particularly images of African women as beastly and oblivious of pain in childbirth, justified European conquest, colonization, and enslavement. The skewed sex ratios that characterized the slave trade—with far more men than women being sent to plantations in the Caribbean and the Americas—necessitated transformations in family

life and in productive and reproductive labor and over time shaped new patterns of ethnic, regional, and even religious identity in both Africa and the Americas. Moreover, while acknowledging the overwhelmingly racial character of slavery, historians have now demonstrated the powerful ways that gender shaped the experiences of bondage and of freedom. And in recent years, studies of twentieth-century Africans and African Americans have explored the ways that class, religion, and region as well as gender shaped divisions of labor, movements for social reform and social justice, and struggles for national liberation (Berger 1992, Morgan 1997, Diouf 1998).

As the mosaic of women's and gender history has expanded, a growing number of scholars have turned their attention to conservative women and women on the right. They challenge earlier feminist histories that highlighted women's engagement with progressive, emancipatory, and pacifist struggles or that treated women's engagement with conservative, right-wing, and fascist politics as solely a function of their domination by men. Rather, building on scholars' efforts to reveal women's agency in all aspects of their lives, historians of conservative and right-wing women have insisted that we recognize many women's willing embrace of antifeminist, racist, nativist, and elitist policies and projects. Beginning in the 1980s with works such as Koontz's (1987) *Mothers in the Fatherland: Women, the Family and Nazi Politics*, a number of historians have argued that women who accept traditional family and community roles often feel threatened by efforts to transform their place in society and in response support movements that celebrate motherhood, domesticity, and female subordination (Gravenhorst and Tatschmurat 1990, Sarkar and Butalia 1995). Most of these studies focus on women's engagement with the far right—facism, the Ku Klux Klan, and religious fundamentalism. Only recently, have scholars explored women's participation in more mainstream movements, demonstrating not only their support of conservative political and social agendas but also the ways that that support has undermined feminist claims for women's universal support of peace, social welfare, and other progressive causes (Delegard 1999).

These studies, though their subjects are mainly members of dominant racial and ethnic groups, reinforce claims made by historians of people of color, workers, and lesbians, for the significance of differences among women. They also hark back to earlier work on social reform that demonstrated that women and men who shared economic and racial status often held common views on social order and social change. And they fit neatly alongside more recent studies in which the social relations of the sexes take center stage in analyzing the dynamics of social, political, and cultural change (Kelly 1986, Davidoff and Hall 1987, Guy 1992, Berger 1992, Roberts 1994, Sinha 1995, Bennett 1996, Gordon 1999).

3. Globalization and Revitalizing Gender and Feminist History

The issues raised by gender history were complicated further in the 1990s by the call for more global histories. Certainly earlier work focused on diverse areas of the world. Nonetheless, the increasing globalization of the economy and of communications and the emergence of women's movements and women's histories in a growing number of nations has demanded renewed attention to international and transnational perspectives. Many nations gained their independence in the final decades of the twentieth century, even as they remained financially and culturally colonized by Western industrial powers. The emergence of women's and gender histories in newly independent nations has provided not only additional stories of women's past struggles but has also demanded a rethinking of the roles of women, and feminists, in imperial and colonial politics and culture from the seventeenth century to the present (Davis 1995). Studies of women, gender and sexuality as they shaped colonial relations in South Asia have been especially important in exploring this rich new terrain (Burton 1994, Jayawardena 1995, Sinha 1995).

Interest in comparative and transnational work and in queer theory, which extends postmodernist critiques of unitary subjects into the realm of sexuality and beyond, promise even more complex renderings of women and men, femininities and masculinities in motion around the world (Alexander and Mohanty 1997). Finally, with a profusion of work on women, feminism, and gender now appearing in a wide array of countries and regions and on a broad range of racial, ethnic, national, and sexual communities, differences within as well as between groups of women are increasingly highlighted, making generalizations more difficult even as it often makes comparisons more fruitful. Thus, for instance, differences among women of color rooted in class, color, and sexual orientation, are being explored in greater depth. Similarly, ideological differences among feminists are being analyzed at the same time that their international connections—past and present—are being highlighted and the differences in conceptualizations of women's rights and priorities are being explored in a more global context. The challenges that new and forthcoming work pose to existing interpretations combined with the rich array of research by scholars claiming identities as feminist, women's and gender historians suggests that these fields will continue to have profound effects on the development of history.

See also: Beauvoir, Simone de (1908–86); Bernard, Jessie (1903–96); Boserup, Ester (1910–99); Feminist Economics; Feminist Epistemology; Feminist Legal Theory; Feminist Political Ecology; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in Political

Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Gender and Place; Gender, Feminism, and Sexuality in Archaeological Studies; Klein, Melanie (1882–1960); Mead, Margaret (1901–78); Multicultural Feminism: Cultural Concerns; Population Dynamics: Two-sex Demographic Models

Bibliography

- Alexander M J, Mohanty C T 1997 *Feminist Genealogies, Colonial Legacies, and Democratic Futures*. Routledge, New York
- Bennett J 1996 *Ale, Beer, and Brewsters in England*. Oxford University Press, New York
- Berger I 1992 *Threads of Solidarity*. Indiana University Press, Bloomington, IN
- Brown E B 1992 What has happened here? The politics of difference in women's history and feminist politics. *Feminist Studies* 18: 295–312
- Burton A 1994 *Burdens of History*. University of North Carolina Press, Chapel Hill, NC
- Christian B 1987 The race for theory. *Cultural Critique* 6: 51–63
- Davidoff L, Hall C 1987 *Family Fortunes*. Hutchinson, London and University of Chicago, Chicago
- Davis N Z 1995 *Women on the Margins*. Harvard University Press, Cambridge, MA
- Delegard K M 1999 Women patriots. Ph.D. Thesis, Duke University
- Diouf S 1998 *Servants of Allah*. New York University Press, New York
- Editors 1989 Why gender and history. *Gender & History* 1: 1–4
- Gordon L 1999 *The Great Arizona Orphan Abduction*. Harvard University Press, Cambridge, MA
- Gravenhorst L, Tatschmurat C 1990 *Töchter Fragen*. Kore, Verlag Traute Hensch, Frieberg, Germany
- Guy D 1992 White slavery. In: Parker A, Russo M, Sommer D, Yaeger P (eds.) *Citizenship and Nationality in Argentina, Nationalisms and Sexualities*. Routledge, New York
- Hall C 1992 *White, Male and Middle Class*. Routledge, New York
- Hewitt N, Lebsack S 1993 *Visible Women*. University of Illinois Press, Urbana, IL
- Jayawardena K 1995 *The White Woman's Other Burden*. Routledge, New York
- Kelly J 1986 The social relations of the sexes. In: *Women, History, and Theory*. University of Chicago Press, Chicago
- Koonz C 1987 *Mothers in the Fatherland*. St. Martin's Press, New York
- Landes J 1988 *Women and the Public Sphere in the Age of the French Revolution*. Cornell University Press, Ithaca, NY
- Meyerowitz J 1992 American women's history. *Canadian Review of American Studies* 23: 27–52
- Morgan J 1997 Some could suckle over their shoulder. *William and Mary Quarterly* 54: 167–92
- Pierson R, Chaudhuri N 1998 *Nation, Empire, Colony*. Indiana University Press, Bloomington, IN
- Riot-Sarcey M 1994 *La démocratie à l'épreuve des femmes*. Michel, Paris
- Roberts M L 1994 *Civilization without Sexes*. University of Chicago Press, Chicago
- Sarkar T, Butalia U 1995 *Women and the Hindu Right*. Zed Press, London

- Scott J 1988 *Gender and the Politics of History*. Columbia University Press, New York
- Sinha M 1995 *Colonial Masculinity*. Manchester University Press, Manchester, UK
- The Language of Class* 1987 International labor and working class history. 31 (Special issue)
- Smith B 1998 *The Gender of History*. Harvard University Press, Cambridge, MA
- Ulrich L T 1990 *A Midwife's Tale*. Random House, New York

N. Hewitt

Gender and Feminist Studies in Political Science

There are two main issues which feminism continues to pose for political science. The first is the questioning of any conception of social science that sees disciplines as neutral and universal bodies of knowledge. The second, and closely related issue, is the limit of the political. In short, not only have feminists argued that traditional political science was very far from gender neutral; they have also argued that, in order to analyze politics adequately, and incorporate gender issues, it is necessary substantially to extend more traditional political science understandings of what constitutes 'politics.' In the process, feminist work has connected with wider analyses of gender, emphasizing the need to analyze gender politics in terms of the construction of both masculinities and femininities.

1. Political Science as Neutral

From the 1970s on, feminist academics increasingly challenged both the gender neutrality of the focus of political science and the gender neutrality of the discipline itself. As Squires has pointed out: 'until the emergence of feminist theory as a recognized academic perspective, contemporary political theory was largely assumed to be gender-neutral in focus' (Squires 1999). Not only did that neutrality come to be challenged but so did women's largely subordinate position in the political science profession. Giving her Presidential Address to the Australasian Political Studies Association in 1981, Pateman noted that: 'Women's position in the profession of political studies is very similar to our position in other professions and in public life generally; we are marginal participants in organizations that are numerically and structurally dominated by men' (Pateman 1982). While matters have improved in many respects since the time when Pateman was writing, a recent survey of the position of women in the American discipline of political science reported that women held only 21.1 percent of full-time faculty positions in 1997–8 and constituted only 11.2 percent

of full professors. Improving the position of women in the discipline has been made harder by a general decline in tenure-track job opportunities (Sarkees and McGlen 1999). A survey of the impact of feminism on Canadian political science found that fewer than 640 of the nearly 17,770 pages published in the *Canadian Journal of Political Science/Revue canadienne de science politique* in the 22 years since its inception discussed women. Fortunately, the situation had improved markedly in the previous seven years (Arscott and Tremblay 1999).

2. The Limits of 'Politics'

Most challenging of all to conceptions of political science as 'neutral' were the challenges to the subject matter of the discipline. In particular, with the increasing impact of academic feminism, the very subject of the 'political' in political science was thrown into question. Politics, as Pateman had pointed out in her 1981 Address, tended to stop at the garden gate. Traditionally, politics was about government, the state, political parties, and other manifestations in the public sphere. Politics was not about gendered power relations in personal life, domestic labor, or women working in the volunteer welfare sector, or organizations representing sex workers in the third world, or representations of women in the media, or domestic violence or rape. Indeed, Squires has argued that it is the limited subject matter of mainstream political science that has most influenced its androcentric nature: 'The adoption of the narrow institutional or instrumental conception of the political within most dominant renderings of the discipline' is what 'accounts for the fact that the study of politics has been one of the last to take up the challenge of feminist scholarship, and most recently men's studies, and modify the cannon' (Squires 1999). Given the need to move beyond limitations of what has been defined as the political, it is not surprising that the engagement between feminists and political scientists has been continually informed by sociologists, philosophers, historians, linguists, and literary theorists rather than confined to insights generated from within the discipline (see e.g., Butler and Scott 1992).

The archetypal second-wave feminist slogan of 'the personal is political,' therefore, had a particular meaning for political science and challenged the discipline's boundaries. Initially, the tasks of feminist political science seemed relatively straightforward. Neglected areas of women's participation could be documented. The welfare state was reconceived in terms of its role in reinforcing established gender relations rather than merely analyzed in terms of the influence of Keynesian economics or postwar European social democracy. (For example, governments frequently assumed that women who were married or in *de facto* relationships should not be eligible for state

benefits. Rather, it was the job of their male breadwinner to support them.) There could be speculation regarding the possibility of undocumented feminist debates in classical Greece that may have influenced Plato and outraged Aristotle. Leading female political thinkers could be rediscovered and reassessed from fifteenth-century writers such as Christine de Pizan to nineteenth-century writers such as Flora Tristan or Harriet Taylor. Voting studies could concentrate on female voters and analyzing gender difference. The under-representation of women in parliament and the bureaucracy could be documented and critiqued. The lack of provision of services targeted for women, from childcare to women's health, could be highlighted. Feminists could analyze attempts to influence state policy and services, for example, the Australian experiment with 'femocrats,' feminist bureaucratic advisers to government (Eisenstein 1996).

Such tasks still remain a central focus of feminist political science, for example, they are amongst the aims of the *International Feminist Journal of Politics*, established in 1999. They have long been the aim of older journals such as *Women and Politics*. However, it is not a matter of just adding women on to existing analyses; feminists have argued that political science also needs to readdress some of its most fundamental theoretical and analytical underpinnings.

3. The Reassessment of Traditional Political Theory

Many of the issues raised by late twentieth-century feminists were not of course new. As already indicated, predecessors from fifteenth-century Christine de Pizan to better known feminists such as Mary Wollstonecraft had raised issues regarding women's relative exclusion from public life. Nonetheless, from the 1970s on, feminist political theorists along with academics from related disciplines began to reassess traditional political theory. Okin (1979) was one of the first 'second wave' feminists to argue that the fathers of modern liberal democratic thought tended to construct the citizen as a male head of household. The role of women was seen to revolve around the private sphere of the family. Women were to have only limited, if any, involvement in public, political, and economic life. Even liberal supporters of female suffrage such as John Stuart Mill assumed that, while women should be able to choose either a career or marriage, the majority of women will choose marriage. Having made the latter choice, the woman's primary commitment would be to managing the family and rearing children. Such views have led Coole (1988) to argue that, despite tendencies towards egalitarianism, liberal democrats expected women to defer to their husbands on political matters well into the nineteenth century. The issue of marriage also has a deeper significance. Pateman (1988) has argued that the marriage contract

was indissolubly intertwined with the social contract. In other words, when contract theorists postulate an original social contract in which free men, living in the state of nature, joined together and agreed to give up certain individual and collective powers to government, they were telling only part of the story. Underlying the social contract was a marriage contract in which men continued to have power over their wives and their wives' property. The wage contract was similarly gendered as the employee tended to be assumed by both employers and governments to be a male wage earner supporting his family.

While there is debate about how relevant Pateman's analysis of contract is to analyzing contemporary political discourse, there is a widespread feminist view that the citizen subject is still being constituted in masculine terms. Brown (1995) paints a complex picture of how this occurs. She draws on both a conception of ideology as an obfuscating narrative and a conception of discourse as a narrative that is constitutive of a social order. Brown notes that liberalism simultaneously produces gendered subjects whose differences are seen as natural (an ideological moment in her terms), and abstract individuals who appear to be genderless (a discursive moment in her terms). The liberal individual may appear superficially to be gender neutral but is still gendered masculine in crucial ways. As numerous feminist theorists, including Brown, have pointed out, gendered dualisms operate throughout political discourse. For example, in much neoliberal discourse, categories such as self-interest are associated with the masculine, and categories such as selflessness are associated with the feminine. The citizen is meant to be an abstract, self-reliant, and entrepreneurial individual. However, far from the self-interested liberal individual being autonomous and self-reliant he is in fact dependent upon the labor and emotional support of the subordinated female 'other.' It is, therefore, hardly surprising that cutting state welfare services often results in labor being transferred to unpaid female labor in the domestic sphere. For example, as jobs and services are cut in the public sector, it is frequently women who then have to increase their level of care for sick, young, old, or disabled relatives at home. Of course, such problems are not confined to some forms of market liberalism but also appear in some of the classic works of social democracy. For example, Pateman (1996) has argued that T. H. Marshall's classic statement of the 'social' rights to which citizens were entitled was quite literally based on a conception of the citizen as male breadwinner and worker. Here, as elsewhere, women's exclusion from various 'social rights' was not addressed adequately. Not only were married women frequently denied eligibility for welfare payments but 'social' rights were conceived of according to 'male' norms. Rights to services most affecting women such as childcare were neglected. Female under-representation in parliaments and legislatures was not seen as

a major barrier to women's citizenship that needed to be urgently addressed. In short, as Phillips (1991) has pointed out, far from transcending gender, the apparently gender neutral figure is frequently based on a male norm. Feminists have therefore argued that the projects of traditional political theory were often inherently androcentric.

While the situation has improved, feminists argue that many of these problems are still apparent in contemporary western political and social theory. So, for example, Arscott and Tremblay (1999) argue that key Canadian theorists of citizenship such as Jim Tully and Will Kymlicka 'have paid next to no attention at all to gender' despite its relevance and the fact that they do deal with relevant ethnic or indigenous issues.

4. Women and Political Representation

The critique of traditional political science and political theory has had a strong influence on more empirical feminist analyses of women and political representation. It is not just that feminists have studied forms of politics neglected by traditional political science, but feminist analyses have also contributed new insights into mainstream political science concerns such as how politicians are selected. For example, in her analysis of Women National Leaders, Francine D'Amico reminds us that many prominent female political leaders have taken 'the widow's walk,' literally walking into the political leadership job previously occupied by their deceased husbands. Corazon Aquino, former President of the Philippines, would be a typical example. In another form of 'surrogate' politics, daughters have occupied positions occupied by their deceased fathers, for example, Benazir Bhutto, former President of Pakistan. D'Amico points out that traditional studies of political leadership tended to neglect such paths to office because of their emphasis on the 'public' rather than the 'private,' for example, on recruitment of political leaders from a background in business, the military or the professions (D'Amico 1995).

However, concentrating on a few prominent female political leaders should not blind us to the fact that, as of January 2000, only eight countries in the world, Sweden, Denmark, Germany, Finland, Norway, Iceland, the Netherlands, and South Africa had reached the UN sanctioned target of women's share of seats in national parliaments being 30 percent or above. New Zealand had reached 29.2 percent, Australia 22.4 percent, the UK 18.4 percent, and the US 12.5 percent (UNIFEM 2000). Christy has drawn attention to some of the most obvious national differences that can influence women's level and forms of political participation, even where there are not apparently huge cultural or religious differences. For example, as Christy points out in a comparison of Europe and the US, women's attitudes to politics may

be different in countries with more extensive welfare states, where women can easily see the relevance of government provision in areas such as caring for children, the sick, and elderly. Electoral systems that utilize proportional representation, with party listing of candidates, tend to elect more women (Christy 1994). Multimember electorates are, therefore, a particularly important factor (Norris 1994).

The impact of women's increased political representation is still subject to debate, particularly since some prominent political leaders, such as Margaret Thatcher, have been identified with antifeminist politics. Nor can one assume that women will behave 'better' than male politicians. Nonetheless, Conway and co-workers conclude in their study of American politics that the increasing number of women in the legislatures and the political elite has already influenced policy. Female legislators are more likely to emphasize social welfare issues and support equal rights policies. They are also likely to place more emphasis on helping constituents. Their presence has 'probably led' to male members of the political elite being aware that women's issues are important (Conway et al. 1997). The claimed difference in the attitudes of female and male legislators might well reflect the persistent so-called 'gender gap' on issues such as social welfare which a number of commentators have found in longitudinal voting and opinion poll studies (see e.g., the famous study by Shapiro and Mahajan 1986).

5. Analyzing the Politics of Masculinity

While feminism has played a crucial role in exposing traditional political science's neglect of women, it has also drawn attention to the need for political science to analyze the role of the state and political institutions in constructing masculinity. Indeed, one could argue that political institutions are just as implicated in the construction of particular forms of masculinity as they have been in marginalizing women. If, for example, the citizen is being conceived in terms of the male citizen breadwinner, what are the implications of this for conceptions of masculinity? Do such conceptions of citizenship encourage not just a predominantly heterosexual masculinity but also forms of masculinity that emphasize financial self-reliance and paternalism? Do they encourage forms of masculine self-esteem that are dependent upon female subordination? In short, are various forms of masculinity being privileged over others here and what are the implications for men as well as women? Hearn's (1992) study analyzes constructions of masculinity within the hierarchies of patriarchy pointing out that some men oppress other men on the basis of class, race, sexuality, disability, and other categories. Work such as Yuval-Davis's (1997) has drawn attention to the ways in which nationalism is gendered in both masculine and femi-

nine terms. For example, masculinity can be mobilized for national defence while femininity can be mobilized for training children (and even resistant adults) in appropriate cultural values and national identities. However, increasingly, work on the politics of masculinity has also had to deal with the phenomenon of male organizations engaging in 'backlash' politics, as more traditional forms of masculinity are seen to be under threat from feminism (Messner 1997). There is certainly a need for more political science work on the wider normalizing political discourses that help to construct particular forms of both masculinity and femininity. However, it should be noted that mainstream political scientists sometimes have been even less keen to analyze the politics of masculinity than they have been to analyze so-called women's issues. After all, problematizing the self can be even more threatening than acknowledging the other.

6. Analyzing the Politics of Feminism

The analytical tools which feminist theory developed to critique gendered discourses drew on a variety of sources from sociology of gender to French theory's critique of those forms of dualistic thinking that subordinated the feminine 'other.' Increasingly, these analyses began to be turned upon the subject of feminism itself. While early feminist engagements with political science concentrated on documenting women's subordination, marginalization, and exclusion, the issues have now become far more contentious. It is not just the constitution of the 'political' which has been found to be problematic but also the very constitution of categories inherent to feminism such as 'woman.' Difference between women has been emphasized. Feminism itself has become the subject of deconstruction and critique as racial, ethnic, cultural, and class biases in particular versions of feminism have been highlighted. Sometimes, these critiques have a direct impact upon analyses of feminism and political science. For example, the authors of the study on women in Canadian political science (Arscott and Tremblay 1999) note the continued impact of Anglophone and Francophone divisions within the discipline and amongst feminists themselves—many Anglophone scholars rarely citing or teaching the work of their Francophone colleagues. Meanwhile, queer theory raises issues that earlier feminist analyses of traditional political theory tended to neglect such as the extent to which constructions of the citizen as male head of household were not only patriarchal but also constructed the citizen in heterosexual terms. The public/private conception, once so crucial to feminist critiques of traditional political theory, has been criticized for being a Western conception. Women's subordination is now as likely to be seen as being due to the continued influence of gendered binary linguistic categories as being due to more materialistic factors such as the gender division of labor.

Not surprisingly, these issues are also reflected in debates over feminist political practice. For example, there are major debates over whether feminist strategies should revolve primarily around cultural issues of 'recognition' or whether the battles should be over issues of economic 'redistribution.' There are also debates over whether feminist politics has revolved around 'states of injury' which government has then been called upon to redress, rather than upon more empowering political strategies (Brown 1995). Many 1970s' versions of feminism have been seen to result from forms of 'identity politics' which, theorists such as Butler argue, both have strategic importance and can reproduce forms of exclusion. Butler (1991) states that the constitution of the identity of the subject has wider implications since 'identity categories tend to be instruments of regulatory regimes, whether in the normalising categories of oppressive structures or as the rallying points for a liberatory contestation of that very oppression.' Consequently, political scientists such as Dean (1997) have argued for developing reflective political strategies that take issues of difference seriously while negotiating forms of feminist solidarity and political action. Dean's collection on *Feminism and the New Democracy* argues for the importance of increasing democracy not only at the level of the state but also in personal relationships.

7. Feminism, Political Science, and Policy

In short, contemporary feminists are still debating, and stretching, the limits of the political. In Vickers's phrase (1997), feminists need to 'reinvent' political science. However, political science also poses a number of challenges to feminist theory and practice. Key amongst these is the challenge posed by the policy-making process. For example, critiques of the universalizing category 'woman' can be useful in the policy-making process when the resulting sensitization to difference draws attention to the need to develop programs that address a range of women's problems, experiences, and lives. They can be less helpful if the differentiation goes so far that the desirability of feminists establishing women-focused programs is brought into question, or if feminists deny the insights that can be drawn across cultures from old feminist concepts such as gender stereotyping or the public/private divide (Lee 1995, Herrick and Sapieva 1997). Similarly, queer theory can be useful for drawing attention to the ways in which both minority and mainstream identities exclude those who are not seen to belong. The insights of queer theory can be less useful if the need for specifically lesbian advocacy or services is ignored. Arguing for a plurality of interpretations may well be a useful feminist political strategy in the fields of literature or film that prevents gendered meanings being fixed. Eventually, however, policy-makers will need to make definite decisions and implement specific programs. Debates over recog-

nition and redistribution usefully draw attention to the need for programs that address both aspects but are less useful if they lead to the privileging of one aspect over the other. Critiques of dual categories can be useful for drawing attention to the discursive subordination of the feminine but are less useful if dichotomous thinking is seen as the fundamental source of women's inequality that must be addressed. Here, as elsewhere, a combination of judicious eclecticism and sensitivity to difference and multiplicity can be helpful in negotiating the complex engagements between feminism and political science.

The relationship between feminism and political science has also been complicated by the times. The backlash against so-called 'political correctness' in many English-speaking countries has made it more acceptable to not merely dismiss feminist critiques of gender bias but to accuse feminists of silencing and marginalizing men. As indicated previously, the institutions and organizations of gender politics now include various aggrieved men's groups—groups mobilizing over issues of child custody, divorce settlements, and parental support—as well as women's organizations and gay and lesbian organizations. The impact of economic globalization and rapid social and technological change has contributed to an environment in which appeals to nostalgia, including gendered forms such as bringing back dependent stay-at-home Mums and reliable male breadwinners, can be electorally attractive to women as well as men. Meanwhile, the influence of neoliberalism, particularly in English-speaking countries, has undermined those feminist critiques of male dominance and female disadvantage that relied on state action to redress the balance. While there are forms of neoliberalism that attempt to incorporate some women into conceptions of the self-reliant individual, others flirt with forms of social conservatism which privilege more traditional gender relations.

All of this has made the engagement between feminism, gender studies, and political science more complex. It has also meant that analyses must be updated constantly. While appeals to nostalgia ensure that older feminist analyses are unlikely to be totally superseded, there is a need to develop analyses that can deal with the evolving political discourses and practices of the twenty-first century. Earlier feminist analyses concentrated on analyzing the politics of exclusion and marginalization. However, increasingly, feminists may need to develop sophisticated analyses of new forms of political inclusion and incorporation, (and new forms of masculinities and femininities), as changes in gender politics are played out in numerous sites and discourses. In the process, mainstream political science could usefully learn to be as self-critical and reflective as feminism has become.

See also: Beauvoir, Simone de (1908–86); Bernard, Jessie (1903–96); Boserup, Ester (1910–99); Feminist

Economics; Feminist Epistemology; Feminist Legal Theory; Feminist Political Ecology; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Gender and Place; Gender, Feminism, and Sexuality in Archaeological Studies; Klein, Melanie (1882–1960); Mead, Margaret (1901–78); Multicultural Feminism: Cultural Concerns

Bibliography

- Arneil B 1999 *Politics and Feminism*. Blackwell, Oxford, UK
- Arscott J, Tremblay M 1999 Il reste encore des travaux à faire: Feminism and political science in Canada and Québec. *Canadian Journal of Political Science/Revue canadienne de science politique* XXXII: 125–51
- Brown W 1995 *States of Injury: Power and Freedom in Late Modernity*. Princeton University Press, Princeton, NJ
- Bryson V 1999 *Feminist Debates: Issues of Theory and Political Practice*. Macmillan, London
- Butler J 1991 Imitation and gender insubordination. In: Fuss D (ed.) *Inside/Out: Lesbian Theories, Gay Theories*. Routledge, New York
- Butler J, Scott J (eds.) 1992 *Feminists Theorize the Political*. Routledge, New York
- Christy C 1994 Trends in sex differences in political participation: A comparative perspective. In: Githens G, Norris P, Lovenduski J (eds.) *Different Roles, Different Voices: Women and Politics in the United States and Europe*. HarperCollins, New York
- Conway M, Steuernagel G, Ahern D 1997 *Women and Political Participation*. CQ Press, Washington, DC
- Coole D 1988 *Women In Political Theory: From Ancient Misogyny to Contemporary Feminism*. Wheatsheaf, Brighton, UK
- D'Amico F 1995 Women national leaders. In: D'Amico F, Beckman P (eds.) *Women in World Politics: An Introduction*. Bergin and Garvey, CT
- Dean J (ed.) 1997 *Feminism and the New Democracy*. Sage, London
- Eisenstein H 1996 *Inside Agitators: Australian Femocrats and the State*. Allen and Unwin, Sydney, Australia
- Hearn J 1992 *Men in the Public Eye: The Construction and Deconstruction of Public Men and Public Patriarchies*. Routledge, London
- Herrick R, Sapieva A 1997 Perceptions of women politicians in Kazakhstan. *Women and Politics* 18(4): 27–40
- Lee A-R 1995 Korean women's politicization: The effects of public/private constructions. *Women and Politics* 15(2): 53–83
- Lovenduski J, Hills J 1981 *The Politics of the Second Electorate: Women and Public Participation*. Routledge and Kegan Paul, London
- Messner M 1997 *Politics of Masculinities: Men in Movement*. Sage, Thousand Oaks, CA
- Nelson B, Chowdhury N (eds.) 1994 *Women and Politics Worldwide*. Yale University Press, New Haven, CT
- Norris P 1994 The impact of the electoral system on election of women to national legislatures. In: Githens G, Norris P, Lovenduski J (eds.) *Different Roles, Different Voices: Women and Politics in the United States and Europe*. HarperCollins, New York

- Okin S M 1979 *Women in Western Political Thought*. Princeton University Press, Princeton, NJ
- Olson K 1999 *Adding Insult to Injury: Social Justice and the Politics of Recognition*. Verso, London
- Pateman C 1982 Women and political studies: Presidential address to the 23rd annual meeting of the Australian Political Studies Association, Canberra, August 1981. *Politics May*: 1–6
- Pateman C 1988 *The Sexual Contract*. Blackwell, Oxford, UK
- Pateman C 1996 *Democratization and Citizenship in the 1990s: The Legacy of T H Marshall*. Vilhelm Aubert Memorial Lecture 1996. Institute for Social Research and Department of Sociology, University of Oslo, Oslo, Norway
- Phelan S 1997 The shape of queer: Assimilation and articulation. *Women and Politics* 18(2): 55–73
- Phillips A 1991 *Engendering Democracy*. Polity Press, Cambridge, UK
- Rae S (ed.) 2000 *International Perspectives on Gender and Democratisation*. Macmillan, Houndmills, UK
- Sarkees M R, McGlen N E 1999 Misdirected backlash: The evolving nature of academia and the status of women in political science. *Political Science and Politics* 32: 100–9
- Shapiro R, Mahajan H 1986 Gender differences in policy preference: A summary of trends from the 1960s to the 1980s. *Public Opinion Quarterly* 50: 42–61
- Squires J 1999 *Gender in Political Theory*. Polity, Cambridge, UK
- Turshen M, Holcomb B 1993 *Women's Lives and Public Policy: The International Experience*. Praeger, Westport, CT
- UNIFEM 2000 *Targets and Indicators: Selections from Progress of the World's Women*. UNIFEM, New York
- Vickers J 1997 *Reinventing Political Science*. Fernwood, Halifax
- Yuval-Davis N 1997 *Gender and Nation*. Sage, London

C. Johnson

Gender and Feminist Studies in Psychology

1. The Emergence of a Field

Feminist studies in psychology emerged in the late 1960s, in the context of the second wave of the women's movement (see *Feminist Movements*). A group of activists, mostly graduate students and newcomers to psychology, founded the Association for Women in Psychology in 1969. At about the same time, a group of somewhat more established psychologists worked within the American Psychological Association to create a division for the psychology of women. Divisions on ethnic minority psychology and lesbian/gay psychology were established later.

Division 35, inaugurated in 1973 and now known as the Society for the Psychology of Women, is one of the largest divisions of APA. AWP continues as an activist organization with a collective governance structure and no formal ties to APA, although Division 35 and AWP cooperate on various projects and there is

considerable overlap in membership. Similar organizational development took place in Canada and the UK, where the British Psychological Society now houses a Psychology of Women Section and a Lesbian Psychology Section. For the first time in the century-long history of psychology, the structure of its major professional organizations acknowledges the presence and professional identity of diverse women. The study of women and gender, and women themselves, now have a political and organizational base within psychology.

With these developments came the founding of feminist journals devoted to research from psychology and related disciplines: *Sex Roles* (1975), *Psychology of Women Quarterly* (1977), and *Feminism and Psychology* (1991) are perhaps the most widely circulated. Their founding, along with the growth of interdisciplinary women's studies journals, provided new publication outlets for research on women and gender. In 1974, there was not a single journal article on women and achievement published, although achievement motivation was a lively research field; in 1993, there were 161. Similar growth has occurred for other topics, notably sexual violence and feminist therapy.

The growth of feminist psychology can also be seen in the number of psychology departments offering courses in the psychology of women or gender. Before 1968, there were virtually none. By 1990, 51 percent of departments that responded to an APA survey included such courses, and there were 172 offered at the graduate level. The first two textbooks appeared in 1971; today, there are many in print.

Along with facilitating the recognition of women's presence in psychology, the women's movement gave rise to a feminist critique of psychological theory, research, and practice and to the development of new approaches in all these areas. Organized activism by AWP and Division 35 led to a surge in feminist symposia and papers at APA national conferences starting around 1970. Critical books and journal articles objected to psychology's views of women. Naomi Weisstein, among the first second-wave critics, famously claimed that psychology had nothing to say about women because it knew nothing about women. Carolyn Sherif analyzed how the needs of the cold-war military had shaped psychology's research agenda. Phyllis Chesler argued that clinical psychology and psychiatry were instruments of control directed at women.

These, and the many other voices of second-wave feminist psychology, were not cautious or timid ones. The 1970s feminist critique quite openly challenged psychologists' choice of research topics and methods, their theoretical constructs, and their modes of diagnosis and therapeutic intervention, and it attacked the values underlying them. It had an assertive tone and an activist agenda that probably contributed to its exclusion from 'mainstream' journals. A similar confrontational approach had last been seen some 50 years

earlier, during the first wave of the women's movement, when some of the first American women to receive higher degrees in psychology questioned the biases of the psychology of their day, such as the ongoing search for sex differences in personality and ability. Hollingworth (1916) wrote a powerful critique of the social devices for impelling women to bear children, while Wooley (1910) characterized research on sex differences in mental abilities as 'riddled with flagrant personal bias, logic martyred in the cause of supporting a prejudice ... and sentimental rot and drivel.'

A fundamental aspect of second-wave feminist critique was the claim that psychological knowledge was androcentric, or male-centered. Men were studied much more often than women, and research questions were formulated from gender-biased perspectives. For example, research on violence and aggression focused on male-male aggression ranging from animal territoriality to human warfare. Violence by men against women, particularly in the context of intimate relationships, was overlooked and under-studied.

Theories were constructed from androcentric perspectives, and women's behavior was judged against a (white, middle-class) male norm. Theories of lifespan psychological development, for example, were often derived entirely from research on middle-class men's life patterns. Women's behavior was more likely than men's to be explained in biologically determinist terms, with gender-differentiated social contexts ignored. For example, explanations of the greater incidence of depression in women focused on hormonal causes, not environmental ones like prior physical or sexual abuse. Psychological adjustment for women was articulated in terms of conformity to gender norms. However, because femininity (but not masculinity) was defined in opposition to agency, autonomy, and competence, these norms created many double binds for women. Feminist critics documented gender biases in theories of intelligence, attribution, social influence, learning, memory, thinking, identity, clinical diagnosis, and therapy, among others.

As a result of this critique, psychology began to develop new ways of thinking about women and gender, expand its research methods, and offer more gender-sensitive approaches to therapy and counseling. Contemporary feminist psychology, like most areas of contemporary psychology, is not a unified subject. Meanings are created in a community of practice: they change over time and as members of the community intersect with their other communities. Not all scholarship on the psychology of women is grounded in feminist perspectives, and no single feminist perspective unites all the work that is self-identified as feminist. Rather, an array of political stances is implicit in the scholarship in this area. Feminist psychology means different things to developmental, clinical, and social psychologists; to older second-wave activists and younger people enter-

ing psychology today; to researchers, educators, and practitioners. Most generally, feminist psychology is a form of critical psychology: it interrogates the moral, political, and scientific claims of psychology and attempts to influence the directions of the discipline. Its goal is the systematic analysis of women's psychology and of the impact of gender on human subjectivity and experience.

Subjective definitions are also important. Academic psychologists who self-identify as feminists define their identity in terms of a set of values—respecting women and their experiences, a concern with equality of power, the need for social activism, and the idea of gender as a social construct—and report that they experience feminism as a transformative personal experience.

2. Development and Contributions

One of the important early contributions of feminist psychology was the conceptual distinction between sex (conceived as biological differences between males and females) and gender (conceived as cultural elaborations based on biological differences). The distinction was important because it called into question the naive biological determinism and essentialism of much psychological theorizing. However, this approach later came to be seen as limiting because of its individualism. Although it distinguished social influences from biological ones, it still conceived of gender as a static set of attributes or personality traits, formed by socialization and relatively permanent within individuals. Moreover, the sex/gender distinction rested on the assumption that causation can be determined and is restricted to either nature or nurture, with little allowance for complex interactions of biological and social influences. And it foreclosed the possibility of examining the social construction of biological sex, only to be confronted with the transgender and intersex activist movements. In hindsight, the usefulness of distinguishing gender from sex was that it opened the way to conceiving of gender as a principle of social organization which structures the relations between women and men and governs access to resources. More recent formulations of gender stress its performative aspects; that is, gender is viewed as a set of agreed-upon transactions that are enacted in mundane social relationships.

Another early focus of research was replacing women in the history of psychology. The devaluation, stereotyping and invisibility of women in history had been played out in psychology; indeed, it was the recognition of the 'womanless' state of the discipline that energized the feminist critique of psychological research. Not only were women excluded from the profession, women and their characteristic experiences and activities were thought too unimportant to be a focus of inquiry. Although it originated as attempts to

restore the 'neglected foremothers' of psychology, this line of work catalyzed exploration of the structural and institutional mechanisms of exclusion. An underrepresentation of women in any scholarly discipline points to the importance of gender as a social and political category, prompting questions about how knowledge is sought, validated, and disseminated.

As more research centered on women and gender, new areas of inquiry were developed. Perhaps the most important contribution of feminist and gender studies in psychology is the wealth of new research generated. Uniquely female experiences such as menarche and childbirth have become legitimate topics of study and new feminist perspectives have been applied to them. A feminist critique of autonomy as the endpoint of psychological development grew into theories of the self-in-relation (see *Feminist Theory: Ecofeminist and Cultural Feminist*), and a feminist critique of masculinity/femininity grew into the construct of psychological androgyny (see *Androgyny*). Work on gender socialization, stereotyping, stigma, and objectification proliferated. Violence against girls and women was named and documented. Its causes, incidence, and effects were analyzed (see *Domestic Violence: Sociological Perspectives; Rape and Sexual Coercion; Sexual Harassment: Social and Psychological Issues*). The pervasiveness of gender as a cognitive category, and gender polarization as a cognitive mechanism, were explored. Developmental issues of childhood, adolescence, sexuality, motherhood, midlife, and aging were examined from women-centered perspectives (see *Motherhood: Social and Cultural Aspects; Sexuality and Gender*). New research examined the interactions of multiple roles (spouse, parent, and worker) for both sexes. Sexual identity and orientation were no longer taken for granted, but became active areas for developmental theory-building and empirical inquiry. Women's health issues, long neglected in behavioral and medical research, also became an active site of study. In clinical psychology, a new importance was attached to issues of power in social and therapeutic contexts: the power of diagnosis and labeling, power imbalances in close relationships (including the therapeutic relationship), and the effects of women's social subordination on psychological adjustment. Psychological problems that differed in incidence or characteristic manifestation according to gender (eating disorders, depression, substance abuse, and agoraphobia, among others) received new attention from feminist perspectives.

In interdisciplinary feminist theory there are many variants, including radical, socialist, psychoanalytic, postmodern, and womanist (woman of color). However, in psychology, liberal and cultural feminist approaches have predominated (see *Feminist Theory*). Both these theoretical frameworks are individualistic, leading to conceiving gender largely as a constellation of individual traits. This view of gender as rather static

and internalized has resulted in prolonged debates about the extent, importance, and mutability of sex differences. Liberal feminists have tended to minimize sex differences, arguing that they are small, relatively inconsequential, and easily eliminated with corrective socialization and education. This viewpoint has predominated in debates about cognitive abilities such as mathematics performance. Cultural feminists, on the other hand, have maximized differences, arguing for essential and fundamental divergences between women and men in such areas as identity, relatedness, and moral reasoning.

As the field has matured, there has been a turn away from these debates, which seem intractable, and an increasing interest in gender processes rather than gendered traits. In other words, gender currently is conceived not only as a phenomenon inside individuals but a system of agreed-upon meanings used in social interaction. An increased interest in how gender inequity is recreated and maintained in mundane social interaction is one result, with emphasis on such topics as cognitive bases of stereotype maintenance, gender-biased attributions of others' behavior, tokenism, self-fulfilling prophecies, sex discrimination, sexual harassment, hostile and benevolent sexism, and how the illusion of equity is sustained despite inequalities of status and power (see *Small-group Interaction and Gender*). This reconceptualization has brought social constructionist theory to the forefront and underlies a turn to language and discursive processes, particularly in the UK, where discursive psychology is more developed than in the US.

Feminist psychology has brought about a notable decrease in androcentrism in psychological research, education, and practice. Tangible changes include guidelines for nonsexist research and language in APA publications, the development of courses and curriculum integration resources, and the availability of nonsexist and feminist clinical training. In the US, women now account for 60 percent of Ph.D.s awarded annually in psychology.

There is also an increased presence of women, including lesbians and women of color, in APA's structure and on its policy-making boards and committees. In over 100 years as a professional organization, APA has elected only eight women (all white) to its presidency. The impact of second-wave feminism is evident: six of the eight were elected after 1970. In the UK, the first (woman-organized) psychology of women conference took place in 1987. A decade later, feminist psychology was a theme for the annual conference of the British Psychological Society.

3. Methodological and Theoretical Developments

The characteristic research methods of a discipline reflect its values and the assumptions it makes about its objects of study. Both first- and second-wave

feminists criticized psychology for sexist biases in research, pointing out problems with studying women less often than men, designing research materials from masculinist perspectives, using restricted samples, failing to compare women and men under similar conditions, and exaggerating or overgeneralizing about sex differences. With the second wave, however, came a more fundamental criticism, not of flaws and biases in the use of (otherwise satisfactory) methods, but of the value of the discipline's most characteristic and central methods themselves. Psychology, it was charged, over-relied on experimental methods, which strip behavior from its social context and position the experimenter as an expert manipulator of the participant and situation. Second-wave feminists have also criticized what they see as the naive belief that psychology can discover universal laws of behavior.

There is today considerable variability in the methods used to study women and gender within psychology. The majority of articles published in the two leading US feminist journals, *Psychology of Women Quarterly* and *Sex Roles*, as well as those published in general psychological journals, use quantitative methods such as experimental and correlational designs and meta-analysis. Only the UK journal *Feminism & Psychology* routinely publishes more qualitative and discursive work. However, even outside of feminist studies, psychology has a long tradition of inquiry that transcends laboratory manipulation of isolated variables: field research, observational techniques, content analysis, participant observation, and case studies are a few examples. A significant minority of feminist researchers draws on these traditions as well as newer, postmodern influenced forms of analysis. The feminist call for methodological plurality has had some effects. Special issues of *Psychology of Women Quarterly* have focused on method and theory (1989), transforming psychology (1994), and innovative methods for feminist research (1999).

In contrast to the positivist epistemology that permeated prefeminist research, feminist methodology is characterized by at least three innovations. First, feminist methods are reflexive: they involve recognition of the social identity and involvement of the researcher. Personal reflexivity involves disciplined self-reflection on how one's identity influences one's work and, in turn, how one's work influences the self. Moreover, a broader kind of reflexivity is also apparent—the maintenance of a critical perspective on the discipline and how its dominant paradigms are sustained by powerful institutions.

Second, feminist methods are meant to serve—not drive—the inquiry. There is little methodological elitism, and a great deal of creativity, within feminist psychology. The use of multiple methods within a single research project, and adaptations of methods more characteristic of other disciplines, are emerging practices. These innovations go beyond psychology's traditional concept of convergent validity, in which

different methods of measurement are applied to the same problem in order to strengthen the evidence for a hypothesis by accumulating different kinds of data. Instead, they are ways of creating dialogue between different approaches; they recognize that a method is a theory, and that different methods may lead to different—and incompatible—truths.

Finally, feminist research has a social change orientation. Within feminist psychology, advocacy and scholarship are not viewed as incompatible. Innovative work erodes the polarizations such as basic/applied and theoretical/practical that are maintained elsewhere in psychology. While the original goal of second-wave feminist critics was to eliminate gender bias and make psychology more objective, that feminist empiricist project has evolved into a more sophisticated one. Feminist psychology is no longer merely corrective or oppositional; it is transformative. Science is conceived as a social practice, and the tenets of empiricism, such as objectivity and validity, are subjects of feminist analysis. Alternative modes of research are being developed and recognized, and there is a great deal of practical research oriented toward understanding and resolving social problems.

4. Current Constraints and Future Directions

Despite the growth and gains of feminist psychology, the field has not developed adequately in some key areas. Theoretical initiatives have lagged behind empirical studies. The extended debate between cultural and liberal feminists over the extent and importance of fundamental sex/gender differences delayed recognition of the importance of studying gender as a set of interactional processes and a system of social meanings.

Moreover, despite the wealth of empirical research generated in the last three decades of the twentieth century, the knowledge base of feminist psychology is insufficiently integrated into psychology as a whole. Citation and content analyses indicate that the textbooks, standard courses, and major theories of psychology have been little changed by the availability of its new perspectives. Feminist work is still too easily marginalized and treated as optional.

Feminist psychology has not yet encompassed adequately the diversity of women, despite ongoing attention to this issue. Its research base and its practitioners are still largely white, middle-class, heterosexually identified, and British or North American. And its success in persuading the discipline as a whole to recognize that psychology's claims about human nature and human behavior are culturally and historically situated has been limited.

The relatively successful institutionalization of the psychology of women and gender (and of women) within psychology is not an unmitigated gain. At one extreme, some feminist critics have expressed the fear

that there can be no entity such as 'feminist psychology' without either distorting and weakening feminist commitments beyond recognition, or transforming psychology entirely. They claim that the former, and not the latter, is taking place: that the research that comes to be assimilated is a weak, individualistic 'psychology of women' and decontextualized 'sex differences,' and that there has been a loss of political activism within feminist psychology over time. However, others point to the large variety of topics, methods, and epistemological starting points represented in contemporary feminist psychology. The multiplicity of approaches allows for healthy dialogue and cross-fertilization, and the reflexivity that increasingly characterizes the field assures that the politics underlying the methods, questions, and practice of feminist scholarship will remain open to analysis.

See also: Beauvoir, Simone de (1908–86); Bernard, Jessie (1903–96); Boserup, Ester (1910–99); Feminist Economics; Feminist Epistemology; Feminist Legal Theory; Feminist Political Ecology; Feminist Psychotherapy; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Sociology; Gender and Place; Gender, Feminism, and Sexuality in Archaeological Studies; Klein, Melanie (1882–1960); Mead, Margaret (1901–78); Multicultural Feminism: Cultural Concerns

Bibliography

- Bem S 1993 *The Lenses of Gender*. Yale University Press, New Haven, CT
- Burman E (ed.) 1998 *Deconstructing Feminist Psychology*. Sage, London
- Caplan P J, Crawford M, Hyde J S, Richardson J T E 1997 *Gender Differences in Human Cognition*. Oxford University Press, Oxford, UK
- Chesler P 1972 *Women and Madness*. Doubleday, New York
- Crawford M 1995 *Talking Difference: On Gender and Language*. Sage, London
- Crawford M, Marecek J 1989 Psychology reconstructs the female: 1968–88. *Psychology of Women Quarterly* **13**: 147–65
- Crawford M, Unger R 2000 *Women and Gender: A Feminist Psychology*, 3rd edn. McGraw-Hill, New York
- Fine M, Gordon L 1989 Feminist transformations of/despite psychology. In: Crawford M, Gentry M (eds.) *Gender and Thought*. Springer-Verlag, New York, pp. 146–74
- Gilligan C 1982 *In a Different Voice*. Harvard University Press, Cambridge, MA
- Hare-Mustin R T, Marecek J (eds.) 1990 *Making a Difference: Psychology and the Construction of Gender*. Yale University Press, New Haven, CT
- Hollingworth L S 1916 Social devices for impelling women to bear and rear children. *American Journal of Sociology* **22**: 19–29

- Kimmel E B Crawford M in press *Innovations in Feminist Psychological Research*. Cambridge University Press, Cambridge, UK
- Landrine H (ed.) 1995 *Bringing Cultural Diversity to Feminist Psychology*. American Psychological Association, Washington, DC
- Morawski J 1994 *Practicing Feminisms, Reconstructing Psychology: Notes on a Liminal Science*. University of Michigan Press, Ann Arbor, MI
- Parlee M 1997 Feminism and psychology. In: Gergen M M, Davis S N (eds.) *Toward a New Psychology of Gender*. Routledge, New York, pp. 65–95
- Russo N F, Dumont B A 1997 A history of Division 35 (psychology of women): Origins, issues, activities, future. In: Dewsbury D A (ed.) *Unification Through Division: Histories of the Divisions of the American Psychological Association*, 2nd edn. American Psychological Association, Washington, DC
- Sherif C 1979 Bias in psychology. In: Sherman J A, Beck E T (eds.) *The Prism of Sex: Essays in the Sociology of Knowledge*. University of Wisconsin Press, Madison, WI, pp. 93–133
- Tiefer L 1991 A brief history of the Association for Women in Psychology: 1969–91. *Psychology of Women Quarterly* **15**: 635–49
- Weisstein N 1968 *Kinder, Kirche, Kuche as Scientific Law: Psychology Constructs the Female*. New England Free Press, Boston
- Wilkinson S 1988 The role of reflexivity in feminist psychology. *Women's Studies International Forum* **11**: 493–502
- Wilkinson S 1997 Feminist psychology. In: Fox D, Prilleltensky I (eds.) *Critical Psychology: An Introduction*. Sage, London, pp. 247–64
- Wilkinson S, Kitzinger C (eds.) 1995 *Feminism and Discourse: Psychological Perspectives*. Sage, London
- Wooley H T 1910 Psychological literature: A review of the recent literature on the psychology of sex. *Psychological Bulletin* **7**: 335–42
- Worell J 1996 Feminist identity in a gendered world. In: Chrisler J C, Golden C, Rozee P D (eds.) *Lectures on the Psychology of Women*. McGraw-Hill, New York
- Worell J, Johnson N G 1997 *Shaping the Future of Feminist Psychology: Education, Research, and Practice*. American Psychological Association, Washington, DC

M. Crawford

Gender and Feminist Studies in Sociology

When feminists first started using the term gender in the early 1970s, the point was to make a clear distinction between how we are born (sex), from how our society shapes who we become (gender). In the process of studying how we become gendered people, feminist sociologists have changed our sense of both what gender is and of how we can best study social phenomena.

Like all sciences dominated by positivist epistemology, sociology traditionally has assumed that, appropriately conducted, it was objective, that is, had no point of view (see *Inferences, Psychology of*). As feminists began to struggle to make social science

useful for understanding the lives of women, we also began to uncover systematic biases in the tradition we inherited, for example, in the kinds of questions asked, the metaphors underlying governing theories, dominant conceptual frameworks, and the strategies used in gathering and analyzing data. These biases tended to follow a pattern. Sociology (like other sciences) was organizing knowledge from the perspective of the experiences and worldview of racially and economically privileged men (Sprague 1997).

As feminist sociologists have struggled to understand gender from the standpoint of women, we have found that it operates on at least three planes of human experience: as a personality structure, as a set of interactional practices, and as an aspect of social organization. Further, we have moved from seeing gender as something about all women and all men to discovering what gender is, how it works, and, therefore, how we might work to subvert it, varies in different intersections with race and class (see *Race and Gender Intersections; Social Class and Gender*). In sum, we have learned that what might appear to many people as a naturally occurring dichotomy of two opposite sexes is the complex outcome of a great deal of social effort and organization.

1. Gender as Something We Are

Most people who use the term gender probably think of it as an individual attribute, a personality structure that we are socialized into fairly early in life. There is a lot of evidence that we respond to children in gender-specific ways. Parents and other significant adults recognize different qualities in girls than in boys, for example, we tend to see girls as cute and loveable and boys as energetic and mischievous. Since we can only support the development of skills and tendencies that we can perceive, seeing girls as different from boys becomes a self-fulfilling prophecy. In attempting to please children, we give them the toys and experiences that our culture has led us to believe they will like and, eventually, that gender-specific advertising directed at children teaches them they should like. In the process, boys have been more likely to learn to like roughhousing and manipulating objects and girls have been more likely to learn to like taking care of others and dressing up and wearing makeup.

Sociologists operating in the tradition of feminist psychoanalysis (Chodorow 1989) say that the development of gendered personalities in contemporary Western society runs much deeper than developing skills and tastes that gender organizes our very sense of self and how we relate to others. Gendered selves are a consequence of gender beliefs in two ways. First, we assign primary responsibility for care-taking of the young to women and not to men. Second, we have learned to use gender as a key indicator of how we should respond and relate to one another. As a result,

children form their first attachments and develop their basic sense of self and their place in the larger community in very different kinds of relationships depending on their gender.

Boys develop their first attachment in a relationship with someone who has learned to understand herself and her charge as opposite sexes. The sexual opposition propels boys into seeing themselves as separate beings earlier in life and masculinity becomes an important part of how they understand their separateness and identity. As development leads to needs for increasing autonomy, the relative absence of males from primary care means that the people boys most need to separate from are women. Thus, an important means by which boys assert their independence involves rejecting the feminine within themselves.

On the other hand, mothers are not prohibited culturally from seeing girls as extensions of themselves, making the mother-daughter attachment much stronger and keeping it in force much longer. When girls do begin to differentiate themselves from their mothers, it is not as complete as it is for boys. Girls' struggle to develop a sense of an autonomous self is not based on gender identity and does not require a complete break from psychological connection with their mothers.

These gendered psychodynamics can be exacerbated in the learning of social roles. Children spend most of their time in domestic and care-taking sites. If girls learn that their most important roles involve taking care of others, they are in places and sets of relationships in which those roles can be practiced. If boys learn that their important roles occur outside their immediate domestic environment, they must acquire a sense of their adult roles by denying their connection to their immediate context and identifying abstractly with the public roles of their fathers and other men.

Their early experiences can lead to gender differences in how people think about themselves and their orientation to the world. Men tend to develop a highly individuated sense of self, an abstract orientation to the world, and contempt for women in general. Women, in contrast, are more likely to develop a connected sense of self embedded in concrete relationships, a more concrete orientation to the world, and hostility toward their mothers in particular.

The psychoanalytic account of gendered personality structure has given feminist sociologists a good deal of insight on how gender works. It explains why gender has such emotional power; why it generates so much anger, hostility, and fear, both interpersonally and in public debate. Men's psychic need to maintain an individuated sense of self defined by a fairly abstract notion of masculinity suggests why emotional intimacy can be threatening, and why men might be more worried about appropriately displaying (proving) their masculinity. If women have a relational sense of self, founded on a relationship with a female caretaker, we can expect they will seek to form strong emotional

bonds, particularly with other women. And, if our sense of who we are depends on distancing ourselves from a powerful mother, that might explain why women in powerful positions make many of us uncomfortable.

The kind of masculine self-psychoanalysis describes is also consistent with the pattern of conceptual organization that feminists have found runs through the organization of knowledge in Western culture, from religion to science (Harding 1998). Common patterns include organizing phenomena into isolated units, arranging them in hierarchies, particularly dichotomies, privileging abstract universals over contextualized understandings, and adopting explanations that best fit the experience of the privileged.

A major shortcoming of the approach that sees gender as something we are, is that it tends to reproduce these same patterns. It tends to represent masculine and feminine as abstract universals, and people as individual bearers of gendered selves, and it generalizes from the experiences of the privileged. As a result, the picture it gives us of gender is distorted.

While all children in the West and many children elsewhere are inundated with mass mediated messages about gender that reflect the beliefs and values of affluent whites, there is reason to suspect that families and communities in other classes and ethnic communities may communicate competing notions about gender to their children. For example, poor and working-class African-American communities in the US are more likely to teach girls to be strong and self-reliant rather than the dominant view of femininity as passive and dependent. Thus, children growing up in culturally marginalized communities are very likely getting mixed messages about their gender's roles and tastes.

Similarly, the psychoanalytic approach assumes a nuclear male breadwinner/female care-taker family structure, one in which few children in the world and increasingly few in West live. This bias is partially the result of the kind of data psychoanalysis tends to rely on: clinical interviews that select for a relatively affluent group, people who have the time and financial resources to undertake psychoanalytic treatment. While it is still the case that men are rare among those who do the work of caring for the young and it seems likely that female mother-surrogates will also relate to girls as like them and boys as the opposite sex, we need to ask how a variety of gendered care-taking relationships shape the development of personality structure.

Finally, seeing gender as an individual attribute makes it seem natural, almost voluntary, downplaying the degree to which gender differences in beliefs and behavior can be outcomes of differences in power and access to resources. Standpoint epistemology prompted some feminist sociologists to ask how people understand themselves and their lives and that drew attention to the impact of situations on behavior. It is well known that people act differently than they

feel in some situations, and act differently from one situation to the next.

Attempting to understand what gender means from the standpoint of actual people also made it apparent that the kinds of situations people are placed in—the constraints they face and the resources they have—differs markedly depending on what class and race/ethnicity they are. Conceiving of people as conscious actors creatively and strategically responding to variable situational demands led these scholars to uncover the situational and enacted character of gender.

2. Gender as Something We Do

From a social constructionist perspective, gender is not an internalized personality structure as much as it is a social fiction we work to maintain through our everyday social practices. We use gender as a lens through which to perceive and evaluate one another. We do not see one another as generic individuals but as men and women, and we have gender-specific expectations for what people are and should be like. We hold one another accountable for acting in ways our culture says are appropriate for our gender (West and Zimmerman 1987).

Knowing that others have gendered expectations of us, we take that into account in deciding how we will act. We work to construct an appearance that will make us classifiable into one of two opposed sex categories: we make choices about what we can and cannot wear, how we cut our hair, and if and where we shave. We take actions to alter the shape of our bodies: we diet, girdle, pad, exercise, and even undergo surgery. We monitor how we conduct ourselves. We understand the appropriate range of gestures and postures for our gender: more expansive and resolute for men, more constrained and passive for women. Traditionally the burden on women has been much higher than for men: the range of attractive faces and bodies, the expectations for clothes and comportment, have been much narrower, requiring more careful self-monitoring and adjustment (Kessler and McKenna 1978).

Being held accountable for performing in gender-appropriate ways in a value system that privileges a certain kind of masculinity puts women in a double-bind. At the same time women are expected to do more work on their bodies, display, and comportment, they are devalued for spending excessive time in trivialities. At the same time women are expected to do the work of facilitating conversation by asking questions, seeking to find topics that men will find interesting, they are derided for talking too much.

The nature of gendered expectations and the resources for meeting them are shaped by class and race/ethnic relations. For example, in the face of centuries of racist barriers to Black men's employment in the US, Black women often have been breadwinners

and have been criticized as matriarchs; economically discriminated against and forced into domestic work, they have been trivialized as 'mammies' (Collins 2000). Asian men immigrating to the US in the nineteenth and early twentieth centuries were 'demasculinized,' denied the opportunity to have wives by restrictions on the immigration of Asian women and laws barring interracial marriage, and forced into traditionally feminine work in domestic service (Espiritu 1997). Feeding the family, a central expectation for women in many cultures, is a constant struggle financially and time-wise for poor women, but it is constructed as trivial work assigned to lower status others by affluent women (DeVault 1991). Motherhood occupies a position of virtue and protection among the economically and racially privileged and is defined as a social problem among poor women, especially members of racial minority groups, who lack access to good paying jobs or husbands with steady employment, and often must rely on state support to raise children. Gendered interactional practices are embedded in every social institution. In schools, for example, responding to children in gender-specific ways influences how, and how much we teach them (Sadker and Sadker 1994). We are more likely to help girls by giving them the answer and to help boys by giving them clues that support their discovering the answer independently. Because girls have been trained to be more cooperative and complacent while boys are expected to be more rowdy and outspoken, engaging boys' interest becomes a higher priority for teachers. The need to keep the boys interested facilitates the development of male-centered curricula and class experiences. Boys learn they are doers, should have high expectations for performance, and develop independent problem-solving skills; girls learn they are observers rather than actors, and have lower and narrower expectations of their own ability.

Classrooms are not the only place where gender is enacted in school. Thorne (1993) has analyzed the ways that elementary school children impose gender-specific expectations on one another on school playgrounds. By fourth or fifth grade, children are imposing gender segregation. For example, boys taunt boys who like to play with girls or who do not act 'manly' enough. Boys play in large groups, evading close adult supervision while developing physical and team-building skills. Girls form shifting pairs of 'best friends' and organize themselves into emotionally charged status hierarchies based on who is friends with whom. They are verbally intimate and much of the content of conversation is about romance and boys, about 'liking' and 'going with,' that is, who can win what boy's approval.

By middle school and high school, the expectation for feminine behavior makes high academic performance a threat to displaying appropriate femininity. Many girls who previously earned self worth through school performance learn that they must switch

attention to winning the recognition of an acceptable male through being sexually appealing. They must walk a narrow line, however, if they are too sexual, they are derogated as 'sluts.' Boys must demonstrate their masculinity through their ability to use force skillfully, through sports and/or technology, and through their conquest of girls. Boys who do not display some sexual power over girls have their own sexuality called into question. Again, performance resources and standards vary by class and race. Economically privileged boys may be academic achievers without undercutting their masculinity; the poorest boys may have to resort to fights to enact their masculinity. African-American girls' self-esteem does not drop as far as that of white and Latina girls (see *Education (Primary and Secondary Schools) and Gender*).

Gendered interactions permeate adult lives. We make many of our decisions about allocating resources and responsibilities in social groups and these groups use gender-specific expectations for determining who will exert influence on decision-making (Ridgeway 1993). We are much more willing to be influenced by men than women: in the absence of information, we will assume that men have competence but that women do not. Men's ability to exert influence seems independent of their interactional style or their apparent motivation. Whether and on what terms we are willing to be influenced by women depends on what gender we are. Men are more willing to be influenced by women who speak tentatively and express priorities on group interests. Thus, women interacting with men are forced to choose between having influence and being seen as competent. All-woman groups, on the other hand, are more willing to be influenced by women who speak assertively and appear to be self-interested, but women like group-oriented women more so women in female groups must choose between having influence and being liked (see *Small-group Interaction and Gender*).

Evaluation is particularly central to workplace experiences—evaluations determine hires, pay, promotions, and the distribution of opportunity—and evaluation is a thoroughly gendered process. Martin (1996) finds that managers tend to see men as having more ability and potential, thus important to groom and mentor. This assessment is shared by male workers who are more likely than women to see themselves as entitled to and thus seek out opportunities and special treatment. In evaluating performance, managers emphasize men's successes and downplay their failures while they tend to do the opposite for women. We see men as entitled to authority but find authority in women problematic, especially when women have authority over men. We are much more comfortable with openly criticizing female colleagues than male ones.

One concrete and pervasive consequence of our holding people accountable for gender-appropriate behavior is the gender gap in pay (see *Sex Differences*

in Pay). Standard job evaluation ratings systematically discount the skills and responsibilities involved in jobs primarily held by women. If those jobs involve taking care of others, they are paid even less, presumably because it is assumed that meeting human needs is somehow a natural extension of an inherent ability to nurture among women.

Knowing others hold us accountable for gender-appropriate behavior and holding ourselves accountable leads us to act in ways that may not seem to be in our individual interests. For example, moving up a career ladder often requires relocating for a promotion or a better job, but women are less willing than men to disrupt their families by changing locations for personal advancement. Also, even women who work outside the home and bring in a substantial portion of the household income will do well more than their share of household labor and feel that the arrangement is fair. Men who might want to be more involved in domestic life are reluctant to be evaluated as less serious about their work. Seeing gender as something we do exposes its malleability. If our acting in gendered ways is at least partly the outcome of gendered expectations and evaluations, we can expose those practices and make choices about their appropriateness. For example, feminist scholarship uncovering biases in job and performance ratings has facilitated the feminist political campaign for comparable worth to develop a more equitable system (England 1992). There are constraints on what we can do to change gender. A standpoint approach to methodology and situational understanding of gender as something we do have helped feminist sociologists to see that situations are not randomly and spontaneously constructed. There are some larger regularities that shape systematically the kinds of situations we find ourselves in. Over the 1980s and 1990s, feminist sociologists increasingly have been talking about gender as a social structure or institution.

3. Gender as a Form of Social Organization

The idea that gender is a social structure or form of social organization has not yet been developed fully. The basic picture seems to be this: gender is the outcome when a society divides labor by sex, assigns the work of meeting human needs to women, and devalues that work. While dividing labor does not lead necessarily to differences in power, this has been the pattern in what are usually thought of as the 'great civilizations,' those social formations that have dominated the world. As we divide labor, we have created a hierarchy of power and authority, largely by separating authority and power from the work of taking care of people, directly meeting their physical and emotional needs.

Historically in the West and in many other cultures, men have been assigned to different social positions

than have women. During the period of industrialization, this sexual division of labor was institutionalized in a division between public and private spheres. Work directed at generating profit and rewarded with wages and other important benefits became defined as a male activity. The work of feeding, clothing, nursing, and otherwise directly meeting people's daily needs for subsistence and emotional support was defined as private, unpaid domestic work and assigned to women.

The public/private distinction has always been more fantasy than fact. The work of taking care of people is not purely a private matter: it occurs in every domain of life (see *Care and Gender*). There are whole professions built around the work of care-taking: education, medicine, and social work. Many debates about social service spending, for example, spending for childcare, healthcare, elder care—are really debates about the degree to which we should take public responsibility for paying for care-taking or allocate it to unpaid domestic labor. In most jobs, some workers make coffee, remember birthdays, listen to troubles, bring treats, and organize social events. This, too, is care-taking work and it, too, is assigned usually to women and not recognized as paid work.

The notion that women had been confined to the private sphere is also inaccurate. Even in the nineteenth century, many women, particularly poor and immigrant women, worked for wages (see *Household Production*). Also, women have done the labor of creating and sustaining communities through volunteer work for churches, charities, and social clubs. The sexual division of paid labor has been resilient over time (Reskin 1988). Even though more and more women have been entering paid labor over the last half of the twentieth century so that now a great majority of women work for wages, we still have primarily men's jobs and primarily women's jobs—there are few jobs where the proportion of women workers is even close to the proportion of men. When women have entered professions previously dominated by men, those professions frequently re-stratify themselves into gender hierarchies (see *Sex Segregation at Work*).

What seems to be a constant in the sexual division of labor is that directly meeting people's needs, especially basic bodily needs, is devalued and assigned to women, and, in some cases, to low-status men (Glenn 1992). The work of care-taking is unpaid or underpaid, unrecognized, or assumed to be instinctive. One indicator of privilege in a gendered form of social organization is the degree to which one can distance oneself from the 'dirty work' of care-taking. All women are held responsible for domestic work, but not all women end up doing it. At the turn of the twenty-first century, most women work for wages and still do the 'second shift' of domestic work (see *Family and Gender*). Those who are in more privileged economic situations hire poorer women to do some or all of the domestic work, either in their homes or

through the modern, commodified domestic labor of the burgeoning service sector with its restaurants, cleaning services, and services that care for and entertain children. Thus, those who are most removed from care-taking work are economically well-off men, typically white men. Those who are the most likely to be involved in it are women, particularly poor women, disproportionately women of color. Increasingly, men and women in higher status jobs are opting out of parenting altogether.

The sexual division of labor is increasingly organized internationally (Mies 1986; see *Economic Globalization and Gender*). Women in countries of the South produce much of the food, clothing, and other commodities that women in the North put to work in their domestic labor. Male-dominated multinational corporations collaborate with male-dominated governments to maintain local policies that restrict the rights of laborers. The extremely low wages and bad working conditions that result from this collaboration enable the generation of high profits. The super-exploitation of third world workers, especially young women, undercuts the situation of working-class men and women in the West who see good jobs and secure futures disappear as more and more factories move offshore.

The sexual division of labor helps support men's political and economic power: their freedom from responsibility for domestic work has allowed them the time and energy for participating in the organizations and activities that underwrite their continued domination (Connell 1987). Men control industry, central state planning, and the means of institutionalized violence, the military and paramilitary institutions. Men dominate the institutions that produce official culture: science, the academy, the mass media. Even though women do the great majority of the work in the world, wealth is concentrated in male hands.

Their command over institutional resources gives some men a base from which to form alliances in solidarity with one another and against the interests of women (see *Male Dominance*). They have used this power to limit the public resources that go to the work of raising children and caring for the frail (see *Social Welfare Policies and Gender*). They allow wages in the workforce to be determined by 'the market,' thus reproducing the existing pattern of valuing men's work over women's. They have restricted women's ability to control their own fertility via birth control and abortion (see *Reproductive Rights in Developing Nations*; *Reproductive Rights in Affluent Nations*). They have raised obstacles against women's solidarity through sanctions against lesbianism (see *Lesbians: Social and Economic Situation*). This combination of practices pushes women into dependence, either on individual men or on the collective male-dominated state (see *Poverty and Gender in Affluent Nations*; *Poverty and Gender in Developing Nations*).

Gender as a form of social organization has until

recently involved two overlapping dichotomies. The social division between those who control and those who do the work of meeting human needs has until recently overlapped with a sexual dichotomizing of men and women. Perhaps what we are seeing in the West is the development of a reorganized gender order, one in which men and women are being further subdivided based on whether or not they have care-taking responsibilities.

4. Conclusion

Feminist sociologists have found that gender organizes our sense of self, is a set of expectations we encounter and must deal with in our daily interactions, and is a form of social organization shaping opportunities and constraints. These are all aspects of the same process. People's selves are the products of relationships and experiences shaped by where they are in a social order organized by gender, as well as race and class. That social order is the crystallization of myriads of daily interactions in which people evaluate and hold one another accountable for behaving in gendered ways. The nature of the expectations individuals confront and their ability to negotiate the outcomes of these interactions depend on the kinds of personal strengths, cultural advantages or disadvantages, and material resources they bring to each. Understanding how this multi-faceted process works, and how we might work to subvert it, requires looking at it from diverse perspectives, including variations in sex, race, class, and nation. The more we do, the more it seems that gender organizes us into many more than two categories.

See also: Bernard, Jessie (1903–96); Boserup, Ester (1910–99); Feminist Economics; Feminist Epistemology; Feminist Legal Theory; Feminist Political Ecology; Gender and Feminist Studies; Gender and Feminist Studies in Economics; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Place; Gender, Feminism, and Sexuality in Archaeological Studies; Klein, Melanie (1882–1960); Mead, Margaret (1901–78); Multicultural Feminism: Cultural Concerns; Population Dynamics: Two-sex Demographic Models

Bibliography

- Chodorow N J 1989 *Feminism and Psychoanalytic Theory*. Yale University Press, New Haven, CT
- Collins P H 2000 *Black Feminist Thought* 2nd edn. Unwin Hyman, Boston
- Connell R W 1987 *Gender and Power: Society, the Person and Sexual Politics*. Polity (in association with B. Blackwell), Cambridge, UK

- DeVault M L 1991 *Feeding the Family: The Social Organization of Caring as Gendered Work*. University of Chicago Press, Chicago
- England P 1992 *Comparable Worth: Theories and Evidence*. Aldine de Gruyter, New York
- Espiritu Y L 1997 *Asian American Women and Men: Labor, Laws, and Love*. Sage, Thousand Oaks, CA
- Glenn E N 1992 From servitude to service work: Historical continuities in the racial division of paid reproductive labor. *Signs* 18(1): 1–43
- Harding S 1998 *Is Science Multicultural? Postcolonialisms, Feminisms, and Epistemologies*. Indiana University Press, Bloomington, IN
- Kessler S J, McKenna W 1978 *Gender: An Ethnomethodological Approach*. Wiley, New York
- Martin P Y 1996 Gendering and evaluating dynamics: Men, masculinities, and managements. In: Collinson D L, Hearn J (eds.) *Men as Managers, Managers as Men*. Sage, Thousand Oaks, CA
- Mies M 1986 *Patriarchy and Accumulation on a World Scale: Women in the International Division of Labour*. Zed Books, London
- Reskin B F 1988 Bringing the men back in: Sex differentiation and the devaluation of women's work. *Gender & Society* 2: 58–81
- Ridgeway C L 1993 Gender, status, and the social psychology of expectations. In: England P (ed.) *Theory on Gender/Feminism on Theory*. A. de Gruyter, New York
- Sadker M, Sadker D 1994 *Failing at Fairness: How America's Schools Cheat Girls*. Scribner, New York
- Sprague J 1997 Holy men and big guns: The can(n)on in social theory. *Gender & Society* 11(1): 88–107
- Thorne B 1993 *Gender Play: Girls and Boys in School*. Rutgers University Press, New Brunswick, NJ
- West C, Zimmerman D H 1987 Doing gender. *Gender & Society* 1(2): 125–51

J. Sprague

Gender and Health Care

The healthcare system is a deeply gendered social institution, often affecting men and women in dramatically different ways. In the United States and many European countries, healthcare occupations are sex segregated: men constitute a majority of physicians, while women are concentrated in occupations that are less prestigious and poorly paid. Because women are more likely than men to seek medical treatment, they are the principal consumers of healthcare. However, research has demonstrated differences in the kinds of healthcare men and women receive, differences that often place women at a disadvantage. Together with inequalities of race, socioeconomic status, and age, gender shapes both the healthcare delivery system and the quality of healthcare men and women receive (see *Socioeconomic Status and Health; Minority Access to Health Services: United States*). Focusing on developments in the United States, this

article discusses men and women's encounters with the healthcare system, as both practitioners and recipients of healthcare.

1. Origins of Interest in Women and Healthcare

Interest in gender and healthcare developed from the engagement of the social sciences with the second wave of feminism that flourished in the 1970s. Feminists made medicine the centerpiece of their analysis, arguing that no institution so clearly epitomized women's subjugation. They developed a far-reaching critique of the healthcare system, which attacked: medicine, for its exclusion of women; the overuse and misuse of drugs, surgery, and technology; and physicians, for withholding information from their patients and trivializing women's concerns. Feminists asserted the right to be full and equal participants in medical decisions.

The women's health movement was particularly influential in the social sciences with, historians, anthropologists, and sociologists testing, applying, and extending feminist ideas. The result was a body of theory and research that reshaped how disciplines conceived of healthcare. In medical sociology, for example, theory and research on gender and healthcare has challenged fundamental assumptions of the medical model (Mishler 1981). Researchers have challenged the assumption that definitions of health and illness are unaffected by the social context by showing them to be culturally variable and historically contingent. Anthropologists, for example, have shown considerable cross-cultural variation in how menstruation and menopause are defined and treated (Lock 1993). Studies revealing gender showing that male and female patients presenting the same symptoms receive different diagnoses and treatments have challenged the assumption that medicine is scientifically neutral. Research on gender and health has been part of the move to create a social perspective on health, illness, and medicine.

2. Gender and the Medical Division of Labor

2.1 Gender and Healing: A Brief Historical Overview

Both men and women have been healers, although their roles have varied historically. By the end of the Middle Ages, the medical division of labor had assumed a form that was to continue in many European countries until the nineteenth century: a largely male stratum, consisting of guilds and dominated by university-trained physicians; and midwives and folk healers, usually women, who served the rest of the population (Ehrenreich and English 1978).

In nineteenth century America, multiple paradigms of healing existed in competition. The predominately male allopaths or 'regulars' served a wealthy clientele in Eastern cities. Regulars were known for treating disease with harsh remedies, such as purging and bleeding. Women were often rejected from regular medical schools and forced to attend women's colleges or proprietary colleges. Women (as well as some African-Americans) were also more likely to enter the schools of the sects that proliferated during that period. In contrast to regulars, sectarians avoided harsh remedies, and most served a rural or working-class clientele (Ehrenreich and English 1978).

At the beginning of the twentieth century, the leadership of the regulars, influenced by progressivism, began a campaign to make medical education more 'scientific,' culminating in the Flexner Report of 1910, a study of medical education commissioned by the Carnegie foundation. As a result of both licensing and the Flexner Report, medical education became university-based and grounded in laboratory science instruction. There were, however, additional consequences. Lacking the resources to build laboratories, proprietary colleges, including sectarian colleges, saw their enrollments dwindle. With the closure of the proprietary colleges, most women, African-Americans, and members of the working class, could no longer obtain a medical education. Medicine became a profession of a white, male elite. The process continued as many states outlawed midwives (Ehrenreich and English 1978) (see *Medical Profession, The*).

Excluded from medicine, many women chose the fledgling profession of nursing. In both the United States and England, the first professional nursing programs were established in the nineteenth-century after Florence Nightingale successfully commanded a contingent of nurses during the Crimean War. Some of Nightingale's views were to become a mixed blessing in the twentieth century. During the Crimean War, she ruled that no nurse could act without explicit orders from a doctor. Throughout the twentieth century, nursing has struggled to regain the autonomy from medicine that Nightingale had forfeited (Freidson 1970).

2.2 The Contemporary Medical Division of Labor

For most of the twentieth century, the gender composition of the American healthcare occupations has followed a single principle: the higher the prestige, power, and pay of the occupation, the smaller the proportion of women. Medicine, the most lucrative and prestigious occupation in the United States, remains a predominately male profession. Although the number of women physicians increased fivefold since 1970, in 1998 less than 23 percent of all physicians were women. Only three percent of all women physi-

cians—or less than one percent of all physicians—were African-American (Pasco et al. 2000–1).

Medicine also remains internally stratified by gender. Women generally are concentrated in the medical specialties having the lowest incomes, such as pediatrics, psychiatry, and family practice—specialties involving considerable direct patient contact. Conversely, the lowest proportions of women are found in the most lucrative specialties, radiology and surgery—specialties that involve medical procedures and require less patient contact. Moreover, the medical subspecialties are themselves internally stratified by gender. Very few women enter the most lucrative surgical subspecialties, such as orthopedic surgery. Within internal medicine, the proportion of women is lowest in cardiology and gastroenterology, subspecialties similar to surgery in their use of procedures and their salary structures (Pasco et al. 2000–1, Wassenaar and Thran 2001–2). The practice patterns of men and women physicians place women at a disadvantage. Women physicians are more likely than men to work in the least lucrative practice arrangements, that is, to be salaried employees (Wassenaar and Thran 2001–2). Less than five percent of women physicians enter administration, teaching, and research—sectors in which the leadership of the profession is found (Pasco et al. 2000–1).

Discrimination against women physicians extends to medicine's informal structure. Disturbing proportions of women physicians report experiencing sexual harassment (Olesen 1997). Women physicians are often excluded from sponsorship networks that channel medical students into prestigious specialties and enable younger physicians to establish practices (Lorber 1984).

Contrasting with medicine is a middle level consisting almost entirely of women, most of whom are white women (Olesen 1997). This level includes the allied health professions—registered nurses, pharmacists, dietitians, and therapists—as well as licensed medical technicians. These professions are also stratified according to race, with the largest percent of women of color found in the occupations with the lowest salaries.

Women allied health professionals and technicians usually advance more slowly and than men in those occupations. For example, the small number of men who have entered nursing have been promoted very quickly to administrative positions (Williams 1992). Women earn less than men who do the same work, and they also earn less than men in predominately male occupations who do different but comparable work. Pharmacy, the only predominately male allied health profession, is also the most lucrative. Pharmacists earn more than nurse supervisors even though researchers found the latter job to be more demanding (Achtenberg 1991) (see *Health Professionals, Allied*).

At the bottom of the occupational hierarchy are the health service occupations, consisting primarily of aides, orderlies, and attendants. Women represent

nearly 90 percent of this group, which also has the greatest proportion of black workers. As care of the chronically ill moved out of the hospital and into the home, the number of home health aides has increased. These jobs, typically performed by women of color and immigrants, are compensated at levels at or below the poverty line (Olesen 1997).

Women also perform most of the unpaid, informal caregiving for sick relatives in the home. Caregiving entails a number of strains, including: loss of social relationships outside the family; interference with work roles; and the strain of performing multiple caregiving roles simultaneously (Olesen 1997). Perhaps the greatest strain arises from the assumption that responsibility for providing care rests exclusively with women (see *Caregiver Burden*)

3. Gender and Medical Encounters

Although the number of female physicians increased dramatically, the most typical medical interview is likely to involve a male physician and a female patient. This pattern raises the question of whether the gender of either physician or patient affects the physician-patient relationship. *Are Male and Female Patients Treated Differently?* Differential treatment can assume two forms. First, physicians can interact differently with female patients by adopting a more controlling communication style or by giving them less information. Second, physicians may give a different diagnosis or recommend a different treatment to male and female patients with similar medical problems.

Some of the first studies to identify the structure of the medical interview were conducted in obstetrical and gynecological clinics. These studies found that physicians: persuaded patients to use or forego specific treatments by presenting information selectively; made recommendations that reflected cultural assumptions about gender (e.g., that only married women should have children); and gave poor women fewer choices than educated, middle-class patients (Fisher and Todd 1986).

These studies show that assumptions about gender affect the content of the medical encounter but do not demonstrate that male and female patients are treated differently. Studies comparing physicians' communication with male and female patients present contradictory findings. On the one hand, physicians spend more time with female patients and give them more explanations, possibly because women present more symptoms and ask more questions (Elderkin-Thompson and Waitzkin 1999). On the other hand, communication between male doctors and female patients is likely to be viewed as problematic: physicians report liking male patients better, and women are more likely to report that physicians talked down to them and trivialized their concerns (Elderkin-Thompson and Waitzkin 1999).

The second question, whether gender influences medical decision making, has been studied extensively. Research undertaken in the 1970s and 1980s tested the hypothesis that primary-care physicians share the commonsense stereotype of women as more emotional than men and are more likely to interpret the women's physical complaints as psychogenic. Consequently, physicians may overlook a physical disease or incorrectly diagnose a psychological disorder. Studies of routine decisions of internists and family practitioners provide strong support for the hypothesis that physicians stereotype female patients as more emotional, and weaker, more equivocal support for the hypothesis that stereotypes result in unequal treatment (Weiss and Lonnquist 1997).

In the late 1980s and 1990s, research on gender bias focused on specialists' decisions, particularly those involving 'high technology medicine.' Even after adjusting for confounding variables, many studies have found male and female patients with heart disease to be treated differently at one or more of the following decision points: women are less likely than men to receive prompt attention in the emergency room; less likely to be hospitalized; less likely to receive noninvasive diagnostic studies, such as treadmill tests, less likely to be treated with thrombolytics ('clot busters'), and less likely to receive invasive diagnostic procedures, such as catheterization. Women, particularly black women, are less likely to undergo angioplasty or undergo coronary bypass surgery. When they do receive angioplasty or surgery, they are referred later in the course of their disease—a pattern that may contribute to women's higher surgical mortality (Beery 1995). Recently, women have also been found to be less likely than men to receive kidney transplants (Elderkin-Thompson and Waitzkin 1999).

These findings should be interpreted cautiously for two reasons. First, a substantial number of studies has failed to find evidence of gender differences in one or more of the treatments of heart disease that have been mentioned. Secondly, because these studies use information extracted from written documents, they do not reveal the actual decision-making process culminating in differential treatment. Several explanations of these findings regarding heart disease have been proposed (Beery 1995). (a) Perceived risk of procedures. Physicians may be reluctant to perform angioplasty and cardiac bypass surgery on women because they view them as more risky for women. In fact, women who undergo bypass surgery are older, have higher comorbidity, and are at a later stage in their disease. However, this pattern may also reflect physicians' reluctance to diagnose and treat women with heart disease at earlier stages. (b) Beliefs about heart disease. The misconception that heart disease is a 'man's disease' may prevent both physicians and women from recognizing the symptoms of heart disease. (c) Different symptoms. Whereas men experience chest pain, women are more likely to have

shortness of breath or epigastric pain, more subtle symptoms that might be not be correctly identified. (d) Cultural assumptions about gender. This explanation receives empirical support from a study of chest pain evaluation in an emergency room, which found men more likely to receive emergency cardiology consults, nitroglycerin, aspirin, and thrombolytic agents but women more likely to receive anti-anxiety drugs (Lehmann et al. 1996). This finding suggests that physicians may fail to diagnose heart disease in female patients because of a belief that their symptoms are psychosomatic. (e) Presentation style. Some writers suggest that women's greater tendency to introduce contextual information into the medical interview may cause their complaints to be discounted or interpreted as psychosomatic (Elderkin-Thompson and Waitzkin 1999).

3.1 Do Male and Female Physicians Differ?

Many studies have examined whether male and female physicians differ in attitudes and values, choice of medical treatment, or interactive styles. This research presents a mixed picture. On the one hand, male and female family practitioners do not differ in their evaluations of common medical problems (West 1993). On the other hand, women physicians are more likely than their male colleagues to order preventive screening; to use a more collaborative mode of interaction; to engage in discussion of psychosocial issues; to deal with emotions; and to create opportunities for their patients to participate in the medical exchange (Roter and Hall 1998). When giving instructions to patients, male physicians are more likely to use the imperative, whereas women are more likely to use requests rather than commands, thereby creating more symmetry in the doctor-patient relationship (West 1993). In most cases, then, the physician's gender is more likely to affect the *style* of communication than to affect the *content* of medical decisions.

Although women first-year medical students show greater commitment to the interpersonal dimension of healthcare, by the time they have completed medical training, women are no more patient-oriented than their male colleagues. Sex role socialization, then, is only one of many forces that shape physicians' attitudes and behavior. Its effects diminish as men and women undergo a similar training and encounter pressures arising from similar work environments (Weiss and Lonnquist 1997, Olesen 1997).

These considerations suggest that increasing the number of women in medicine is a necessary, but ultimately insufficient condition for creating a humane healthcare system responsive to the diverse needs of both men and women. It is necessary because of what women contribute to the medical encounter. It is insufficient, however, because of the absence of women

of color and because very few women physicians enter teaching, research, and administration. What is needed are women who occupy positions of leadership within medicine and are able to reshape the medical curriculum, revise the textbooks, and influence funding priorities.

4. Gender and the Social Construction of Health and Illness

Research on gender and health has demonstrated that conceptions of health and illness are historically contingent, culture-bound, and ineluctably tied to the social context out of which they arise. These changing conceptions have affected the ways women's medical problems have been treated.

In the nineteenth century, for example, Victorians viewed women's psychological complaints as diseases of the uterus or ovaries caused by deviation from their social roles. Their treatments ranged from 'rest cures' to sexual surgery to remove the uterus or ovaries (Ehrenreich and English 1978). Ostensibly scientific ideas were used to make claims about women's proper place.

Although these ideas no longer hold sway, twentieth-century conceptions of illness, like those in the nineteenth, continue to be socially created. In the twentieth century, social and cultural forces have led to the increasing medicalization of women's lives—a process that proved consequential for women's healthcare.

4.1 Medicalization

A persistent trend in Europe and America is the extension of the medical model to an increasing number of deviant behaviors and natural processes. Women, particularly middle-class women, have been deeply affected by the tendency to treat the reproductive phases in the female life course—childbirth, menstruation, and menopause—as illnesses (Riessman 1983).

Since the 1800s, childbirth in the United States has been transformed from a natural event, taking place at home and assisted by midwives into a medical event managed by obstetricians. By the middle of the twentieth century, childbirth in America had been medicalized to an extent unparalleled elsewhere in the world. Women were subjected to myriad medical procedures, such as epidural anesthesia, forceps, and fetal monitors. However, US infant mortality rates continue to surpass those of other European countries in which medicalization has been less extensive (Riessman 1983).

Medicalization is also behind the controversial definition of premenstrual syndrome (PMS) as a

psychiatric disorder. Despite the lack of consensus as to how the condition should be defined and despite the lack of evidence that women's moods vary according to their menstrual cycles, PMS was included in the revised third edition of the American Psychiatric Association's *Diagnostic and Statistical Manual* and has even been proposed as a criminal defense (see Tavris 1992). Some doctors have treated women diagnosed with PMS with hormones. The medicalization of PMS can be a double-edged sword. On the one hand, women may be relieved to learn that they suffer from a real illness. On the other hand, the medicalization of PMS in the criminal law may revive the nineteenth-century view that women's mental lives are controlled by their reproductive systems.

There is reason to believe that the origins of PMS are cultural rather than hormonal. Since 1931, descriptions of this condition have varied considerably. In periods in which women were expected to work, the effects of PMS were depicted as mild. During periods in which women were expected to leave the labor market, the effects of PMS were described as debilitating. Further, although middle-class women interviewed in one study viewed themselves as suffering from PMS, working-class women did not believe in its existence.

In Europe and the United States, menopause, once viewed as a natural process has been defined as a disease requiring medical treatment. Inequalities of age and gender combine to create a highly negative view of menopause in medical texts, which depict menopause 'hormonal failure' having far-reaching consequences for the heart, the skin, and the skeleton and requiring treatment with hormone replacement therapy (Lock 1993).

These views of menopause are deeply culture-bound. In Japan, views of the aging woman are shaped by concerns with population aging, cultural views of the women's roles rooted in ancient Japanese society, and the preoccupation with work. In Japan, menopause has not been medicalized although aging women are nevertheless stigmatized as unemployed caretakers. Since women are viewed as 'natural' caretakers, government officials have suggested they should become full-time caregivers of both aging parents and aging in-laws. It is these issues, rather than their hormones, that concern Japanese women (Lock 1993).

4.2 Medicalization and Medical Technology

Medicalization is often accompanied by the growth of medical technologies. In fact, medicalization is encouraged by the drug companies, who have a stake in promulgating the ideas that pregnancy, PMS, and menopause are diseases requiring treatment.

Early studies of estrogen replacement therapy, DES, and the Dalkon shield focused on their rapid diffusion into medical practice before their consequences were

known and the drug companies' delay in recalling these technologies even after their risks had become apparent. (see Riessman 1983). This pattern, however, has occurred with other medical technologies unrelated to women's health.

More recent research has focused on the social consequences of new reproductive technologies, such as genetic screening, *in vitro* fertilization, and fetal surgery healthcare. Some new reproductive technologies have diminished women's control in decision making, while augmenting the power of professionals, social movements or the state. For example, the use of implantable contraceptives by the poor or by women in developing nations raises the spectre of a powerful group or society controlling the fertility of the less powerful (see Ruzek et al. 1997). New diagnostic technologies diminish the importance of women's experience of their bodies and force them to depend on experts to interpret the new technologies. With the development of ultrasound, for example, women became less reliant on their own experience of fetal movements and more reliant upon experts to help them 'see' the fetus in an otherwise obscure sonogram (Duden 1993).

Some commentators fear the potential of some new medical technologies to devalue the mother while increasing the social value of the fetus. For example, the new specialty of fetal surgery results in the creation of an unborn patient—the fetus—while the mother becomes peripheral to the treatment process (Casper 1998). This trend has culminated in the fetal rights movement, in which African-American mothers accused of using drugs have been prosecuted for allegedly endangering fetal health or even forced to undergo Caesarian sections. In these cases, women lose control of reproduction, which has been taken over by the state.

Most critics of new medical technologies do not simply assume that all medical technologies are intrinsically harmful to women. Rather, most take the more nuanced view that the consequences of any technology depend largely on the social context in which it is deployed.

4.3 Demedicalization of Reproduction

In the United States, broadly-based social movements have attempted more recently to demedicalize childbirth. In the 1960s, some middle-class parents sought alternatives to the medical model, even to the extent of bypassing the hospital altogether and turning to lay midwives. At the same time, the women's health movement challenged professional dominance even to the extent of usurping what had been previously considered professional prerogatives (Ruzek et al. 1997).

As a result of these social movements, alternatives to labor and delivery have developed, including home

births using lay midwives or nurse midwives and delivery by nurse-midwives in free-standing birth centers (Rothman 1983, Riessman 1997). These alternatives encourage mothers and families to become actively involved in the birth process, avoid technological interventions, and rely on hospitals only when complications arise.

While organized medicine has opposed home births and attempted to regulate freestanding birth centers, hospitals have accommodated the demands of many middle-class parents for a less medicalized childbirth by creating alternative birth centers in hospitals and by offering Lamaze classes. These changes, however, do not fundamentally challenge the technological and interventionist model of conventional medicine.

Despite opposition, home births and freestanding birth centers have continued through the end of the twentieth century. However, these social movements have appealed exclusively to a small, educated, middle-class clientele—a limitation that will be discussed in the next section.

5. *Intersecting Inequalities*

In the 1980s and 1990s, research on gender and health moved away from an exclusive focus on gender and toward understanding how gender intersects with race, class, and other forms of inequality. Academic feminist influenced by postmodernism criticized treating gender as a unitary category and eliding differences among women. As scholars and activists struggled to understand why the women's health movement had appealed only to white, middle-class women, it became apparent that many women had perspectives and interests that diverged from those of white, middle-class women. For example, disability activists oppose prenatal testing and the selective abortion of 'defective' fetuses—policies that, in their view, symbolically devalue the disabled. Nonwhite women and women of low socioeconomic status, lacking access to basic healthcare, seek to the very high-technology medicine that white, middle-class women reject. Women of color, for whom forced sterilization campaigns in the 1970s are a recent memory, seek protection from the enforced use of the very technologies white women view as a social good. Because medical technologies may benefit some women to the detriment of others, it is misleading to extrapolate from the experiences of one group of women and generalize about women as a whole (Ruzek et al. 1997) (see *Minority Access to Health Services: United States; Socioeconomic Status and Health Care*).

The focus on differences among women has opened avenues of inquiry. One line of research examines how the divergent worldviews of middle and working-class women result in different perspectives on abortion or reproduction (Luker 1984). Another line of research explores the role of race, class, and gender in shaping

healthcare policies. For example, the early birth control movement emphasized the potential of contraception to enhance reproductive freedom of all women. As control of the movement shifted to physicians and eugenicists, its focus shifted to population control, directed toward controlling the fertility of those deemed 'unfit' (Gordon 1983). Future research should also examine the differential impact of healthcare policies on different groups of women.

The mandate to study differences among women consumers of healthcare has yielded a richer and more complex picture of diverse women, whose perspectives and interests sometimes collide and sometimes coincide. Once these differences have been identified, it may be possible to draw more meaningful generalizations about the kind of healthcare that serves diverse women.

See also: Care and Gender; Caregiving in Old Age; Gender and Cardiovascular Health; Gender and Physical Health; Gender and Reproductive Health; Gender Role Stress and Health; Health Care Markets: Theory and Practice; Health Care Delivery Services; Medical Sociology; Medicine, History of; Men's Health; Minority Access to Health Services; United States; Reproductive Medicine: Ethical Aspects; Reproductive Rights in Developing Nations; Socioeconomic Status and Health Care; Women's Health

Bibliography

- Achtenberg J 1991 *Woman as healer: A panoramic survey of the healing activities of women from prehistoric times to the present*. Shambhala, Boston, MA
- Beery T 1995 Gender bias in the diagnosis and treatment of coronary artery disease. *Heart and Lung Journal of Critical Care* 24(6): 427–35
- Casper 1998 *The Making of the Unborn Patient: A Social Anatomy of Fetal Surgery*. Rutgers, New Brunswick, NJ
- Duden B 1993 *Disembodying Women: Perspectives on Pregnancy and the Unborn*. Harvard University Press, Cambridge, MA
- Ehrenreich B, English D 1978 *For Her Own Good: 150 Years of the Experts' Advice to Women*. Anchor, New York
- Elderkin-Thompson V, Waitzkin H 1999 Differences in clinical communication by gender. *Journal of General Internal Medicine* 14(2): 112–41
- Fisher S, Todd A 1986 Friendly persuasion: The negotiation of decisions to use oral contraceptives. In: Fisher S, Todd A (eds.) *Discourse and Institutional Authority: Medicine, Education, and Law*. Ablex, Norwood, NJ
- Freidson E 1970 *Profession of Medicine*. Aldine, Chicago
- Gordon L 1983 The politics of birth control, 1920–1940: The impact of professionals. In: Fee E (ed.) *Women and Health: the Politics of Sex in Medicine*. Baywood, Farmingdale, NY
- Lehmann J, Paulette S, Lehmann C, Savory L 1996 Gender bias in the evaluation of chest pain in the emergency department. *American Journal of Cardiology* 77(8): 641–4
- Lock M 1993 The politics of mid-life and menopause: Ideologies for the second sex in North America and Japan. In: Linden-

- baum S, Lock M (eds.) *Knowledge, Power and Practice: The Anthropology of Medicine and Everyday Life*. California, Berkeley, CA
- Lorber J 1984 *Women Physicians: Careers, Status, and Power*. Tavistock, New York
- Luker K 1984 *Abortion and the Politics of Motherhood*. California, Berkeley, CA
- Mishler E 1981 Critical perspectives on the biomedical model. In: Borwn P (ed.) *Perspectives in Medical Sociology*. Wadsworth, Belmont, CA
- Olesen V 1997 Who cares? Women as informal and formal caregivers. In: Ruzek S, Olesen V, Clarke A (eds.) *Women's Health: Complexities and Differences*. Ohio State University Press, Columbus, OH
- Pasco T, Silverman B, Birkhead S (eds.) 2000–1 *Physician Characteristics and Distribution in the United States, 2000–2001*. American Medical Association Center for Health Policy Research, Chicago
- Riessman C 1983 Women and medicalization: A new perspective. *Social Policy* 14: 3–18
- Riessman C 1997 (1981) Improving the health experiences of low income patients. In: Conrad P (ed.) *The Sociology of Health and Illness: Critical Perspectives*, 5th edn. St Martin's, New York
- Roter D, Hall 1998 Why physician gender matters in shaping the physician-patient relationship. *Journal of Women's Health* 7(9): 1093–7
- Rothman B 1983 Midwives in transition: The structure of a clinical revolution. *Social Problems* 30(3): 262–71
- Ruzek S, Clarke A, Olesen V 1997 What are the dynamics of differences? In: Ruzek S, Olesen V, Clarke A (eds.) *Women's Health: Complexities and Differences*. Ohio State University Press, Columbus, OH
- Tavris C 1992 *The Mismeasure of Woman*. Simon and Schuster, New York
- Wassenaar J, Thrane S 2001–2 *Physicians Socioeconomic Statistics, 2002 Edition*. American Medical Association Center for Health Policy Research, Chicago
- Weiss G, Lonnquist L 1997 *The Sociology of Health, Healing, and Illness*, 2nd edn. Prentice Hall, Upper Saddle River, NJ
- West C 1993 Reconceptualizing gender in physician-patient relationships. *Social Science Medicine* 36(1): 57–66
- Williams C 1992 The glass escalator: Hidden advantages for men in the 'female' professions. *Social Problems* 39: 53–67

R. R. Anspach

Gender and Language: Cultural Concerns

Research on gender and sex in sociolinguistics and discourse analysis started systematically in the early 1970s, probably as a result of women's liberation movements. Investigators examined two domains of language behavior in particular: speech behavior of men and women on the phonological level, and interactions (conversational styles) between women and men in discourse (for overviews and the history of the field see Wodak and Benke 1996, Holmes 1996,

Wodak 1997, Kotthoff and Wodak 1997, Wilkinson and Kitzinger 1995, Crawford 1995, Hall and Buchholtz 1995).

1. Setting Out the Problems

Studies of gender-specific language behavior are often contradictory and depend on the authors' implicit assumptions about sex and gender, methodology, and samples used, etc. For example, some studies suggest that women use more standard language than men because they try to adapt to social norms (Trudgill 1974). On the other hand, other large-scale surveys have shown that young women tend to speak dialect and nonstandard language as well (Wodak and Dressler 1978), due to specific situations like family conversations and peer group talk.

The socialization of infants is often regarded as primarily the duty and work of mothers; mothers use 'motherese,' a very specific discourse adapted to the needs of infants (see Wodak and Rindler-Schjerve 1985); as a result of caretaking, children acquire the regional dialects and vernacular which their mothers use at home. Other investigations illustrated that professional women are under more pressure than their male colleagues; therefore they adopt high standard sociolects or even hypercorrect features (see Labov 1966, Moosmueller 1997). However, hypercorrectness is also clearly related to social class: Working-class and lower-middle-class men also tend to speak hypercorrectly in certain organizational contexts (Leodolter 1975). Very detailed case studies of code switching, moreover, have shown that the use of either standard or nonstandard styles as well as minority and majority languages depends highly on the context of the utterances as well as on social class (Gal 1989). Nonstandard lects can also be the manifestation of certain political opinions and attitudes as in the case of Ebonics (see Baugh 1999).

The same variation of correlations is true for studies on conversational styles: In some cultures, men clearly interrupt women more frequently and dominate topic choice, but in other cultures this need not be the case (see Shieffelin 1990). As a result, women's language has been said to reflect their conservatism, prestige consciousness, upward mobility, insecurity, deference, nurturance, emotional expressivity, connectedness, sensitivity to others, solidarity, and men's language is heard as evincing their toughness, lack of affect, competitiveness, independence, competence, hierarchy, control (Eckert and McConnell-Ginet 1992, p. 90).

Many empirical studies have thus neglected the context of language behavior and have often analyzed gender by merely looking at the speakers' biological sex (see the arguments in Nichols 1983, Cameron and Coates 1990, Henley and Kramarae 1991). By focusing

only on the biological sex, the whole group of women is contrasted to the whole group of men without taking differences in these holistic categories into account (see below).

Such a dichotomization clearly does not make sense when looking at many studies which have shown significant differences that occur with women or with men due to culture, social class, ethnicity, etc. (Wodak and Schulz 1986, Kotthoff and Wodak 1997). Some of the research has isolated the variable of sex/gender from other sociological or situational factors and has made hasty generalizations about genderlects (Chambers 1992). A context-sensitive approach, however, which regards gender as a social construct leads to more fruitful results (see Harres 1996, Wetschanow 1995). Moreover, gender is related to and overlaps with the sociocultural and ethnic background of the interlocutors, with their age, their level of education, their socioeconomic status, their emotions, and the specific power-dynamics of the discourse investigated.

The focus nowadays lies on the discursive construction of identities in interaction where gender is but one part of this attempt (Wodak et al. 1999). Thus, research on gender should take the whole interaction and context into account; of course, the biological sex will always be part of a categorization but other categories might prove more valid in a specific research which makes it possible to operationalize gender roles (the inclusion of professional roles and organizational roles with gender is of utmost importance as well as culture and ethnicity; the socialization of women who come from Islamic cultures is so different than the background of Western European women that intercultural problems clearly arise in interactions, see Martin-Rojo 2000, Wodak 2000a, Essed 1992, 1997). All these results suggest that gender has an enormous impact in many very relevant situations, like schools, hospitals, courtrooms, etc., but only in combination with other variables. However, it is not possible and is fruitless to try to decide which variable has more impact generally. Only very detailed qualitative studies can provide insight into such complex theoretical questions.

2. Feminist Linguistics

Analogous to the term 'racism,' the word 'sexism' was invented in the 1960s. It refers to discrimination within a social system on the basis of sexual membership. With the concept of 'sexism,' women defined themselves for the first time as a social group and as a suppressed minority expressing basic political interests of being judged and valued in equal terms as men. Of course, the history of women's fight for equal rights is much older (see Saurer 2000) but the political movements in the 1960s, together with innovations like the invention of the contraceptive pill and birth control on

a larger scale started a qualitative new time for women's emancipation. As such, they sought to reveal the mechanisms of suppression, making others aware of and fighting these devices. Social groups often define themselves by means of their common language, which plays an important role in identity creation and, for subcultures, serves as a means of differentiating themselves from the outside world. This specific identity manifests itself in certain conversational styles, manifestations of emotions, etc. (see Coates 1997, Buchholtz et al. 1999).

According to Hellinger (1990), feminist linguistics (FL) is distinguished from all other disciplines by the following three aspects, which similarly analyze the relationship between language and gender:

(a) FL places female and male linguistic behavior and the linguistic phenomena connected with the designations of women and men at the center of its considerations.

(b) FL interprets person-related asymmetries in the field of language systems and language use as expressions of the linguistic discrimination of women (sexism) and links these directly to the plane of social discrimination. Traditional studies usually make do with descriptive results. Clearly, languages differ in their linguistic realizations of such inequality. In German or Polish, for example, suffixes and certain prefixes manifest gender. Kargl et al. (1997) have submitted a whole register of possibilities of equal treatment of women and men in many everyday genres in communication, starting with names and labels up to whole reports in newspapers, as well as conversations. These attempts are valid in written as well as oral speech and have led to huge debates in the media (see Pusch 1984); in German, specifically, the use of capital 'I' in written texts was at the center of such political discussions (i.e., '*LehrerInnen*' vs. '*Lehrer*' and '*Lehrerinnen*' or '*Lehrer/jinnen*').

(c) FL does not accept phenomena as given, but seeks alternatives in keeping with the principle of the linguistic equal treatment of women and men. It pursues explicitly political goals by criticizing ruling linguistic norms and understanding the linguistic change it advocates as part of an overall change in society (Hellinger 1990, p. 12).

3. Sociolinguistic Studies of Sex/Gender

In the 1970s, 'sex' was established within sociolinguistic research as a social variable next to the already existing variables of social stratum, age, nationality, ethnic affiliation, religion, class, and region. In correlative-quantitative sociolinguistic investigations on the prestige and stigma variants of languages (see Kotthoff 1992) the sex variable became a factor which significantly affected the use of language. The best-known representatives of this quantitative-correlative approach in sociolinguistics,

with its focus on urban groups of persons, are William Labov, Peter Trudgill, and Lesley Milroy. The working methods within this sociolinguistic approach diverge considerably from one another in certain aspects (unlike Trudgill and Labov, Milroy works with network analyses) but in one respect they are identical: their methodology. All of them operate quantitatively, that is, linguistic variables are defined that are realized in different forms in different varieties. Then, the different variants are correlated with sociological parameters.

Several other studies have taken different approaches and have used new and other methodologies. In particular, a lot of criticism related to the implicit chauvinistic ideologies in these first studies and also to the undifferentiated analysis of the two sexes which totally neglected the range of variation in each of the genders under observation. In a study of the language of mothers and daughters in Vienna, for example (Wodak 1984, Wodak and Schultz 1986), it could be shown that, apart from these sociological parameters, other variables, such as psychological ones, are also responsible for the variation concerning each gender. This study stressed in particular the variation that was found among the women and girls investigated, one that relativized the factor of 'sex.' It could be shown that mothers with a positive relationship to their daughters spoke differently than mothers with a conflicting relationship. The same is true for the daughters (30 mother-daughter pairs were interviewed and tested). These results illustrate that emotions and relationships as well as educational styles and the content/discontent of mothers are more important than social class in an isolated way. Moreover, this study proves that the differences in one category of sex can be statistically significant due to all these parameters.

Such results and tendencies are of utmost importance for the whole debate on gender and sex as well as social class. First, it could be shown that significant differences exist in one category (either women or men). Second, many assumptions about the differences between working-class and middle-class socialization could be contradicted—educational styles do not vary so much by social class but more by the self-assessment of mothers and their attitudes to their own life and achievements. Third, it could be shown that no significant differences appear between mothers who work at home and mothers who have professional lives outside of their homes. Therefore, many claims that children need mothers who only stay at home are wrong; the satisfaction of women with their own lives is the most important factor that affects their relationship with others, in their families or outside of their homes. Fourth, generalizations about the use of vernacular and dialect could also be contradicted: Even daughters from upper-class families used Viennese dialect to distinguish themselves from the phonological styles of their mothers and to be able to

manifest their autonomy. This study illustrates the necessity of a very complex and interdisciplinary approach to gender research.

Another approach of considerable importance in recent sociolinguistics is the concept of 'communities of practice' (Eckert and McConnell-Ginet 1992, p. 92ff). Communities of practice are defined as 'An aggregate of people who come together around mutual engagement in some common endeavor. Ways of doing things, ways of talking, beliefs, values, and power relations—in short, practice—emerge in the course of their joint activity around that endeavor' (Eckert and McConnell-Ginet 1992, p. 95).

This approach relates to the proposals of context sensitivity made above: While researching gender, the emotions, attitudes, values, intergroup and intragroup relationships have to be accounted for as well. Specifically, Eckert and McConnell-Ginet studied peer groups of girls and boys and could illustrate why big differences on a phonological and conversational level occur between two groups due to all the above-mentioned factors. Eckert and McConnell-Ginet argue very convincingly and provide many examples to support their view that survey studies are too general and their level of abstractness too great because all the important sociological and psychological parameters cannot be integrated. Moreover, and this criticism applies to survey studies in general about other topics as well (e.g., research on anti-Semitism and racism, Wodak et al. 1990), surveys usually manifest self-assessments of the interviewees. A bias due to education must be considered as well: Better educated interviewees usually know what kind of answers are expected from them. They answer accordingly even though this might not be their opinion at all. This bias is very important in all studies related to stereotypes and prejudice research. Therefore, such quantitative studies must be complemented with qualitative case studies using authentic data tape-recorded in everyday interactions as well as ethnographic observations (see Titscher et al. 2000).

4. Sexism and the Language System

Feminists conceive language as 'a symbolical reflection of androcentric structures' (Günthner and Kotthoff 1991, p. 7), as one of the means of patriarchal society to discriminate, disregard, and incapacitate women. In their view, the language system already reflects the patriarchal structure of Western societies. Such an assumption manifests a basic consideration of sociolinguistics, namely that language and social norms are connected and that language manifests and also constructs social practices (Fairclough and Wodak 1997). If this claim is accepted (and such theories stem from Wittgenstein's 1967 concept of 'language games'), then the manifestation of women's domination must also be manifest in language.

The assumption that an individual language system has lexical elements and morphological and grammatical rules that are already sexist is based on the premise that 'due to their long history as public decision-makers, men not only determine the economic, political and social orientation of social life, but also influence the functioning and the semantic contents of each individual language' (Postl 1991, p. 89).

Once the language system has become the object of investigation for feminist linguists, it is interesting to note how the linguistic structure of an individual language is connected to the structure of society, how the structure of the language is conditioned by the structure of society, and vice versa. Unlike studies investigating divergent gender linguistic behavior, studies on the sexist use of language focus on the possibilities of reference to both genders or their practice that exist in an individual language. According to Pusch (1984, p. 7), the objects of 'feminist system linguistics' are 'patriarchalisms in diverse language systems.'

5. *The Analysis of Conversation and Discourse*

Unlike the above-mentioned research field, FL—which examines the use of language—concerns itself with gender differences in conversation and discourse (oral as well as written). Differences are investigated in the following fields of language use: voice, pronunciation, intonation, choice of words, argumentation, lexicon, syntax, interactional and conversational behavior, as well as visual features and modes and nonverbal communication. The gender-induced differences in the use of language are not treated simply as divergent variants standing side by side. Due to a patriarchally organized society, the interpretation of the differences ascertained is of major social and political significance.

The interpretation of the diverse linguistic indicators (like turn taking, indirectness, interruptions, and overlaps) varies according to the specific gender theory and ideology, which underlie the studies. (Due to the fact that there are no comparable standardized studies across cultures and social classes as well as contexts and ethnicities, no generalizations about any universal gender or sex-based differences in conversational styles are possible.) Probably the two best-known theories about the gender-induced use of language are those of 'difference' and 'dominance' (see Cameron 1997, Kendall and Tannen 1997).

Whereas investigations following the 'deficit model' have interpreted the use of such indicators as manifestations of 'powerless language of women,' research in the paradigm of the 'difference model' views the linguistic behavior of men and women as originating in opposed modes of socialization and equal in their meaning and impact. Hedging, tag questions, or indirectness can be viewed 'negatively' or 'positively,'

either as signs of female insecurity or as supporting conversational work, depending on the context of the discourse and the theoretical approach adopted. This controversy led to many polemical debates which in the end shows the invalidity of such dichotomous distinctions. Thus, nowadays, other research questions have become more important, viewing gender as constructed due to context.

6. *Recent Theoretical Approaches in the Feminist Study of Discourse*

6.1 *The Construction of the 'Sexes'*

In Judith Butler's arguments against the tenability of the categories of 'woman' and 'sex,' she refers to postmodernist theories. In the manner of Foucault she ascribes normative power to the idea of gender identity and the attempt to describe it. By the mere act of defining a gender identity many bodies, practices, and discourses are excluded or devalued, whereas the constructed and hence upright character of this gender identity is simultaneously concealed.

Thus, in such feminist theories (Butler 1990), the idea of two sexes is criticized as being a construction, a prediscursive factor taking the place of reality. Following this interpretation, 'sex' cannot be separated into a biologically inherent aspect and a socially acquired one. The category of 'sex' itself, in Butler's view, is a purely cultural product of discourse. Such an opinion, of course, relates well to Foucauldian theories of discourse; 'sex,' therefore, cannot be viewed as a 'universal, suprahistorical and extrasocial entity'; it is understood to be an 'integral component of every form of life' (Butler 1990, p. 23); such a view relates well to the assumptions of context sensitivity discussed above. On the other hand, 'sex,' of course, possesses a material basis: In many relevant contexts in our societies, women and men are still evaluated, judged, promoted, or paid according to their biological sex alone. These facts are not accounted for in Butler's approach. More recently, some authors have differentiated Butler's dogmatic position: It is a question of revealing 'the reproduction mechanisms, networks and institutional compulsions that ensure that constructions become persistent and resistant and appear timeless, immovable and identical with themselves' (Hirschauer 1992, p. 333). It is a matter of exposing the arbitrary construction of this binary opposition and hence also its mutability, and not of criticising the binarily organized perception as unreal.

6.2 *Doing Gender*

Unlike a research approach that accepts sexual differences as an aggregation of qualities and deals with the qualitative behavioral tendencies of women and men,

ethnomethodologically oriented studies produced a new focus of research: 'doing gender.' Such an approach complements traditional theories which viewed gender as an aggregation of static attributes and are concerned with investigating and displaying the peculiarities of women and interpreting them as 'gender-specific' or 'gender-typical' attributes so as to reveal the asymmetry of the difference between the sexes, to criticize it and to make it politically visible. The use of a bundle of characteristics and attributes to define gender complicates or renders impossible an interactional approach since attributes are 'entities' and not processes. This concept entails that definitions of gender cannot change, transform, and are not influenced by, spontaneous interactions. A further problem raised by the concept of gender as a concept of attributes is the possibility of individualization; that is, the individual who has been seen to possess or not to possess certain attributes becomes the centre of focus, and the level of the social system is neglected.

Unlike this noninteractive approach, 'doing gender' regards membership of a sex not as a pool of attributes 'possessed' by a person, but as something a person 'does.' In this sense, membership of a gender constitutes a performative act and not a fact. Gender is continually realized in interactional form. Gender is created not only in the everyday activities that characterize 'doing gender,' but also in the asymmetry of the relationship between the sexes, the dominance of the 'male' and its normativeness. Patriarchal inequality is produced and reproduced in every interaction. This concept of 'doing gender' stresses the creative potential and the embedding of gender-typical behavior in a social context.

Female Members of the European Parliament, for example, have created very different ways of positioning themselves in interactions with male colleagues in the European Commission or in the European Parliament itself. Interviews with female MEPs (Wodak 2000a) illustrate that they are very conscious of their gender roles and that they tend to construct very different ways of self-presentation in interactions: in particular, one MEP stated the following, in talking about her first experience as a rapporteur.

when I—entered the parliament— *Orientation* (lines 1–3)
on my first report it was about Leonardo
I don't know if you know:
((laughing)) well—I said 'I'm going to speak to the commissioner'
and—I—I knew—he only speaks very bad French
and my er my French was very bad as well.
so I said 'I want to have interpretation'
So—I went to the commissioner *Complicating Actions* (lines 4–14)
with a very good int/int/interpreter
and I/I/I/I talked more than an hour with him.
because we talked the same about it
and at the end he said—
'well: I have here the advice of my: civil servants but I—agree with you:

and this and this and this all goes through—'
so you have to be: —er—
I don't know/h/ow do we call it in English in /
in the Netherlands we say [brutal]
so you have to: ((laughing)) be polite *Evaluation* (lines 16–20)
but you have to—you:/you mustn't be/
you mustn't *sit* behind your -/your desk—
because that doesn't help. ((laughing))
but then then you have the worse system
that I tried several times *Coda* (22–31)
then you have the Council. ...

In this example, which has been marked for basic narrative structure according to Labov and Waletzky's (1967) model, we see that this MEP's story is objectively about having a successful meeting with a Commissioner while acting as rapporteur on a report about Leonardo—an EU youth and education related program established in 1995 to provide financial support for professional development and job training. In lines 4–6, the complicating actions, she shows how she went to the commissioner with an interpreter, and because she and the commissioner had the same understanding of the issues involved, he was willing to support her, despite contrary advice by his 'civil servants' on the matters involved. The main point of the story, or evaluation, from the MEP's perspective, is to show that as an MEP, to get things done, you must be active and assertive, 'not sit behind your desk.' Thus, in this narrative, she positions herself as an MEP who is proactive and who will do what it takes, including arguing directly with Commissioners, to see that her voice is heard. She also orients to being a rapporteur (line 2), which carries some responsibility in a committee, and to being from the Netherlands (line 17), although this last identity is evoked only to characterize her style of work ('brutal' in Dutch, or 'assertive'). (For further details of the analysis see Wodak 2000a). This narrative also manifests the complexity of the European Parliament, the interwovenness of languages, cultures, positions, political affiliations, and gender roles, as well as professional experiences.

Another of the interviewees is oriented to a particularly wide range of identities (left, woman, Swedish, mother, political outsider, etc.) during her interview. Most striking is the way in which she repeatedly positions herself as being an 'atypical MEP.' Here we see one such occasion.

I figure here the most common—ah civil—job—for an MEP
is er to be a lawyer.
me myself *I'm far from that*
the job I had doesn't even *exist* outside *Scandinavia*.
so: it's a sort of a social teacher—so
so I'm/I'm very in/an/a very special bird in this er:
Interviewer mhm mhm so now you don't feel like you—fit
into sort of a *typical* MEP er
ME *no. no. no: I'm not. I'm left I'm a woman I'm Swedish*
and I'm also everything—/everything's wrong (laughs)

In this example, the MEP contrasts herself with what she considers to be a typical profile for an MEP (lawyer by profession), emphasizing the degree to which she feels different (I'm far from that ... I'm a very special bird ... everything's wrong'). She also points out many of the identities that she associates with and that she perceives as marking her as different (social teacher, left, female, Swedish). This MEP is very much concerned with her self-presentation as being different, different than other female MEPs and also male MEPs. Moreover, she explains her success as a female MEP by being exotic and narrates different strategies which she applies in 'doing politics and gender.'

Even such short examples illustrate the diversity and complexity of gender roles and make clear why an interdisciplinary, qualitative, and context-sensitive approach is necessary to access the whole domain of 'gender and language.' To be able to make more general observations, however, much more cross-cultural comparable research is urgently needed.

See also: Feminist Theory; Gender and Feminist Studies; Gender and Feminist Studies in Anthropology; Gender Differences in Personality and Social Behavior; Language and Gender; Sociolinguistics

Bibliography

- Baugh J 1999 *Out of the Mouth of Slaves*. University of Texas Press, Austin, TX
- Buchholtz M, Liang A C, Sutton L A (eds.) 1999 *Reinventing Identities. The Gendered Self in Discourse*. Oxford University Press, Oxford, UK
- Butler J 1990 *Gender Trouble*. Routledge, New York
- Cameron D 1997 Theoretical debates in feminist linguistics: Questions of sex and gender. In: Wodak R (ed.) *Gender and Discourse*. Sage, London, pp. 21–36
- Cameron D, Coates J 1990 Some problems in the sociolinguistic explanation of sex differences. In: Coates J, Cameron D (eds.) *Women in their Speech Communities: New Perspectives on Language and Sex*. Longman, London, pp. 13–26
- Chambers J 1992 Linguistic correlates and gender in sex. *English World Wide* 2: 173–218
- Coates J 1997 *Women's Friendships, Women's Talk*. In: Wodak R (ed.) *Gender and Discourse*. Sage, London, pp. 245–62
- Crawford M 1995 *Talking Difference: On Gender and Language*. Sage, London
- Eckert P, McConnell-Ginet S 1992 Communications of practice: Where language, gender, and power all live. In: Hall K, Buchholtz M, Moonwomon B (eds.) *Locating Power: Proceedings of the Second Berkeley Women and Language Conference*. Berkeley Women and Language Group, University of California, Berkeley, CA, pp. 89–99
- Essed P 1992 *Everyday Racism*. Sage, London
- Essed P 1997 Race, gender and academic leadership. In: Kotthoff H, Wodak R (eds.) *Communicating Gender in Context*. Benjamins, Amsterdam, pp. 315–34
- Fairclough N, Wodak R 1997 Critical discourse analysis. In: van Dijk T A (ed.) *Discourse as Social Interaction*. Sage, London, pp. 258–84
- Gal S 1989 Between speech and silence: The problematics of research on language and gender. *IPRA Papers in Pragmatics* 3(1): 1–38
- Günthner S, Kotthoff H 1991 Von fremden Stimmen: Weibliches und männliches Sprechen im Kulturvergleich. In: Günthner S, Kotthoff H (eds.) *Von fremden Stimmen. Weibliches und männliches Sprechen im Kulturvergleich*. Suhrkamp, Frankfurt, Germany, pp. 7–52
- Hall K, Buchholtz M (eds.) 1995 *Gender Articulated. Language and the Socially Constructed Self*. Routledge, London
- Harres A 1996 Tag questions and gender in medical consultations. PhD thesis, Monash University, Melbourne, Vic
- Hellinger M 1990 *Kontrastive Feministische Linguistik*. Hueber, Ismaning, Germany
- Henley N, Kramarae C 1991 Gender, power and miscommunication. In: Coupland N, Giles H, Wiemann J (eds.) *Miscommunication and Problematic Talk*. Sage, London, pp. 60–88
- Hirschauer S (ed.) 1992 *Die soziale Konstruktion der Transsexualität*. Suhrkamp, Frankfurt, Germany
- Holmes J 1996 Sex and language. In: Goebel H, Nelde P H, Stary Z, Wölk W (eds.) *Kontaktlinguistik—Contact Linguistics—Linguistique de contact*. de Gruyter, Berlin, pp. 126–46
- Kargl M, Wetschanow K, Wodak R, Perle N 1997 *Kreatives Formulieren. Anleitungen zum geschlechtergerechten Sprachgebrauch*. Schriftenreihe der Frauenministerin, Vol. 13, Vienna
- Kendall S, Tannen D 1997 Gender and language in the workplace. In: Wodak R (ed.) *Gender and Discourse*. Sage, London, pp. 81–105
- Kotthoff H 1992 Unruhe im Tabellenbild? Zur Interpretation weiblichen Sprechens in der Soziolinguistik. In: Günthner S, Kotthoff H (eds.) *Die Geschlechter im Gespräch: Kommunikation in Institutionen*. Metzler, Stuttgart, Germany, pp. 126–46
- Kotthoff H, Wodak R (eds.) 1997 *Communicating Gender in Context*. Benjamins, Amsterdam
- Labov W 1966 *The Social Stratification of English in New York City*. Center for Applied Linguistics, New York
- Labov W, Waletzky J 1967 Narrative analysis: oral versions of personal experience. In: Helm J (ed.) *Essays on the Verbal and Visual Arts*. University of Washington Press, Seattle, pp. 12–45
- Leodolter R 1975 *Das Sprachverhalten von Angeklagten bei Gericht*. Scriptor, Kronberg, Germany
- Martin-Rojo L 2000 Narratives at work: When women take on the role of managers. Workshop on Theory and Interdisciplinarity in CDA 6 and 7 July 2000, University of Vienna
- Milroy J, Milroy L 1978 Belfast: Change and variation in an urban vernacular. In: Trudgill P (ed.) *Sociolinguistic Patterns in British English*. Arnold, London, pp. 19–36
- Moosmueller S 1997 The relevance of fundamental frequency contour for interruptions: A case study of political discussion in Austria. In: Kotthoff H, Wodak R (eds.) *Communicating Gender in Context*. Benjamins, Amsterdam, pp. 401–21
- Nichols P C 1983 Linguistic options and choices for black women in the rural South. In: Thorne B, Kramarae C, Henley N (eds.) *Language, Gender and Society*. Newbury House, Cambridge, UK, pp. 54–68
- Postl G 1991 *Weibliches Sprechen*. Suhrkamp, Frankfurt, Germany
- Pusch L 1984 *Das Deutsche als Mönnersprache*. Suhrkamp, Frankfurt, Germany

- Saurer E 2000 *Liebe und Arbeit: Geschlechterbeziehungen in Europa 19 und 20 Jahrhundert*. Fischer-Taschenbuchverlag, Frankfurt am Main
- Schieffelin B 1990 *The Give and Take of Everyday Life: Language Socialization of Kaluli Children*. Cambridge University Press, Cambridge, UK
- Titscher S, Meyer M, Wodak R, Vetter E 2000 *Methods of Text and Discourse Analysis*. Sage, London
- Trudgill P 1972 Sex, covert prestige, and linguistic change in the urban British English of Norwich. *Language in Society* 1: 179–96
- Trudgill P 1974 *The Social Differentiation of English in Norwich*. Cambridge University Press, Cambridge, UK
- Wetschanow K 1995 Als wenns a Grammatikfehler wär: Splittingverhalten einer Gruppe mit links-alternativ feministischer Unisozialisation. MA thesis, University of Vienna
- Wilkinson S, Kitzinger C 1995 *Feminism and Discourse*. Sage, London
- Wittgenstein L 1967 *Philosophische Untersuchungen*. Suhrkamp, Frankfurt
- Wodak R 1984 *Hilflose Nähe? Mütter und Töchter erzählen*. Bundesverlag, Vienna
- Wodak R (ed.) 1997 *Gender and Discourse*. Sage, London
- Wodak R 2000 National and transnational identities: European and other identities oriented to in interviews with EU officials. Workshop on European Identity, European University, Florence, June 2000
- Wodak R, Benke G 1996 Gender as a sociolinguistic variable. In: Coulmas F (ed.) *The Handbook of Sociolinguistics*. Blackwell, Oxford, UK
- Wodak R, de Cillia R, Reisigl M, Liebhart K 1999 *The Discursive Construction of National Identity*. Edinburgh University Press, Edinburgh, UK
- Wodak R, Dressler W 1978 Phonological variation in colloquial Viennese. *Michigan Working Papers in Germanic Studies* 4(1): 30–67
- Wodak R, Pelikan J, Nowak P, Gruber H, de Cillia R, Mitten R 1990 'Wir sind alle unschuldige Täter!' *Diskurshistorische Studien zum Nachkriegsantisemitismus*. Suhrkamp, Frankfurt, Germany
- Wodak R, Rindler-Schjerve R 1985 Funktionen der Mutter beim Sprachwechsel—Konsequenzen für die Primärsozialisation und Identitätsentwicklung. In: (ed.) *Linguistische Berichte*. Westdeutscher Verlag, Wiesbaden, Germany, pp. 81–82
- Wodak R, Schultz M 1986 *The Language of Love and Guilt*. Benjamins, Amsterdam

R. Wodak

Gender and Physical Health

The health experiences of women and men differ, primarily reflecting gender differences in the cultural and socioeconomic circumstances of women's and men's lives. Gender is an important dimension of social life which intersects with others, especially social class and age. In this article gender differences in mortality and morbidity are examined, together with explanations. The divergent structural and role positions of women and men lead to gender differences in

the nature of inequalities in health, which vary over time and among societies.

1. The Distinction Between Sex and Gender

Expressed simply, sex refers to the biological differences between men and women, and gender refers to cultural differences relating to appropriate behavior of men and women. Gender roles are socially constructed and therefore change over time and across societies. The cultural and religious mores of a society influence the roles, attitudes, and behavior of men and women. Gender ideology prescribes and circumscribes the social behavior of men and women, and of different age or ethnic groups. Social myths and stereotypes about appropriate or supposedly 'natural' behavior for women and men impact on their health and well-being, for example, relating to body image and weight control (Lorber 1997).

1.1 Masculinity and Femininity

Over recent years the simple dichotomy of biological sex has been questioned, as has the cultural gender dichotomy of male/female. The distinction between masculinity and femininity is increasingly recognized as a continuous dimension, with each individual having both masculine and feminine attributes in varying degrees. Masculinity and femininity are inherently relational concepts, arising from gender relations in particular socio-historical contexts.

In Western societies, masculinity has traditionally been identified with physical courage, stoicism, hardness, endurance, aggressiveness, and taking risks. However, notions of masculinity vary with class and ethnicity, across the life course, over time, and among societies (Watson 2000). Structural factors and power relations among men in society may influence forms of masculinity which themselves have health consequences.

The 'risk' taking behavior of men is a key concern within health promotion and public health, but here masculinity is often seen as a unitary construct. Sabo and Gordon (1995) characterize masculinity as intrinsically damaging to men's health, in terms of predisposition to drinking, smoking, drug abuse, risky sports, and so on. However, this sees maleness as problematic, fixed, and immutable (Watson 2000), and fails to examine diversity among men.

In many cultures, illness is seen as a sign of weakness, closely allied to contemporary constructions of femininity, and thereby more easily accepted by women. In particular, chronic illnesses which undermine independence and self-control may threaten masculine self-identity. Masculinity and femininity may therefore influence the way symptoms are evaluated, willingness to consult healthcare providers

and to take medical remedies. This may help explain men's lower use of available health services.

2. Women's and Men's Health

2.1 Women's Health

In the 1970s and early 1980s, the primary focus of interest within medical sociology was women's health rather than gender and health. Biological differences were emphasized, linked to women's distinctive reproductive system, with social research on the specialized services for childbirth, contraception, abortion, and infertility (Doyal 1995) (see *Gender and Reproductive Health; Women's Health*).

Many women in developed countries expect pregnancy and birth to be a normal event, but their experience of maternity care is often a medicalized search for pathology. Medicalization in other areas of women's lives has received less attention from sociologists. There are rising rates of hysterectomy, while the menopause provides a new opportunity for medical intervention in mid-life, especially through the promotion of hormone replacement therapy (HRT).

The impact of gender roles on reproductive health needs to be considered. For example, women's capacity to control their fertility through contraception and abortion is influenced by partners, family members, governments, religious, and other social institutions. It is also important to examine the effects on women's reproductive health of poverty, class, race and ethnicity, and sexual orientation.

2.2 Women and the Male 'Standard' Patient

The focus on women and reproductive health has led to a lack of attention to women as patients within mainstream areas of health. The biomedical or scientific model assumes that disease is a deviation from 'normal' biological functioning. However, what is regarded as normal depends on who is being compared to whom. Men are treated as the 'norm' and women compared to a standard 'male' patient (Lorber 1997). Medical norms and research have primarily been based on white, male, middle class, working age bodies. This approach also disadvantages men, since they are seen as the normal, the 'control,' to which women are compared. This inhibits exploring diversity among men's health and recognizing that maleness itself is a cultural and social phenomenon.

Epidemiological research and clinical trials often exclude women or fail to include them in sufficient numbers, for example, in research on coronary heart disease (CHD) and AIDs (Sharp 1994, Doyal 1995). This lack of medical research on women means that less is known about the effects of drug or surgical treatment regimes on them.

Differences in men's and women's reported illness rates may result from diagnostic behavior and treatment by physicians, as well as gender differences in the process of seeking care—'illness behavior.' Reported gender differences in CHD may partly reflect doctors' behavior. US women are less likely than men to be diagnosed as having heart disease or to undergo coronary angiography or bypass surgery (McKinlay 1996). Thus, rates of recorded illness may differ from the 'true' prevalence in systematic and gender-related ways.

The image of CHD as predominantly affecting men is prevalent in the media, medical advertising, and among the medical profession. There is less publicity and discussion about the risk of CHD for women than men. Yet CHD is the most common cause of death for women. Similarly, there is high public concern with female-specific cancers, such as breast cancer and cervical cancer, and less publicity and research attention on other cancers such as stomach and lung cancer among women (see *Gender and Cancer; Gender and Cardiovascular Health; Gender and Health Care*).

2.3 The Emergence of Men's Health

A marked imbalance exists between the extensive literature on women's health and the paucity of research on men's health. Men's ill-health has remained a largely neglected area, with a lack of studies examining lay ideas about health and illness among men (Watson 2000), and men are often cast as a homogeneous group (Cameron and Bernardes 1998). While some conditions, such as CHD and lung cancer, have been highlighted, other areas, such as prostate problems or mental health, have been given little priority until recently.

Empirical work on ill-health conditions specific to men reveals a variety of problems stemming from the social invisibility of men's illness. In research on prostate problems, Cameron and Bernardes (1998) found that men regarded health as women's responsibility, knew little about men's health, and tended not to talk about any prostatic health problems they had experienced. Prostate problems appeared to represent a threat to the men's sense of masculine self.

In all societies, whether symptoms are perceived as illness by individuals is influenced by cultural values, gender roles, and family interactions. Health professionals also play a key role in defining presented symptoms as 'sickness,' sanctioning entry into the 'sick role,' and making decisions about treatment, all of which may be influenced by gendered norms and assumptions (see *Men's Health; Body Image and Gender*).

2.4 Gender Analysis of Health

A gender analysis of health sees both women's and men's health within the wider context of social and

economic relations between women and men. It recognizes the impact of gender inequalities in power and economic resources on both men's and women's health.

Many women perform the 'double shift' of housework, childcare, and paid labor, so it is important to assess how combining paid and unpaid work affects their health and varies with socioeconomic circumstances (Doyal 1995). Men's health may benefit from the unpaid labor of their wife in providing domestic services and health care; there is substantial evidence that married men have better health and health-related behavior than never married or previously married men.

Women provide the majority of care for elderly, frail, or disabled relatives (Arber and Ginn 1991). Caregiving places a major burden on carers, with consequences for their physical and mental health. The role of unpaid carers in the community is becoming increasingly significant within Western health care provision, as hospital stays become shorter, shifting the costs of care onto family members. The policy of promoting community care and de-institutionalization therefore has a gendered effect with a greater impact on women, who form the majority of carers.

Many women experience rape and domestic violence, which have adverse physical and mental health consequences, but the extent and nature of that violence varies among and within societies (Koblinsky et al. 1993, Doyal 1995). Violence is more common where men's masculinity is defined in terms of dominance, and women are relatively powerless economically and socially. Another form of culturally-sanctioned violence in some societies is female circumcision, which results in deaths, as well as pain, chronic urinary problems, and higher levels of maternal and child mortality.

3. Gender Differences in Mortality

In nearly all countries women outlive men. Four broad patterns of sex differences in mortality can be identified, each related to gender roles.

3.1 Women Outlive Men by 5–7 Years

This pattern is found in most Western countries. For example, in the UK in 1995, expectation of life for women was 79.4 years and for men 74.1, a 5.3 year difference; in the US, it was 78.9 for women and 72.5 for men, a 6.4 year difference; in Sweden, women outlive men by 5 years (expectation of life of 81.5 and 76.5, respectively), and in Japan by 6.6 years (83.6 and 77 years, respectively) (UN 1999).

The 'norm' of women having a 5–7 year longer life expectancy than men is the product of women's biological advantage and men's greater mortality from occupational hazards and risky behaviors such as

smoking, drinking, dangerous sports, fast driving, and violence. As women in Western countries have entered the workforce in larger numbers over the last 30 years and the male excess in smoking is now very small, there has been a reduction in the gender disparity of mortality (Hemstrom 1999). This trend is expected to continue as the lifestyles of women and men converge and women undertake more risky health behaviors.

3.2 Women Outlive Men by 2–4 Years

This pattern is exemplified by European countries in the mid-nineteenth century, and is currently typical of many developing societies in Africa and Asia. For example, in Kenya the expectation of life for women is 61.4 and men is 57.5 (a 3.9 year difference), and in China is 52.9 for women and 50.1 for men (UN 1999). This modest female mortality advantage is typical of transitional societies, where the previously high maternal mortality rate is falling, but where women occupy a disadvantageous structural position with adverse effects on their health. As countries move towards greater economic development and more education for women, the sex gap in mortality tends to increase.

3.3 No Gender Difference or Men Outlive Women

In some developing countries, there is little gender difference or men outlive women. In Bangladesh men outlive women, with an expectation of life of 58.6 for men and 58.2 for women. In Pakistan and India the sexes have almost equal life expectancy (in Pakistan, 59.0 for men and 59.2 for women, and in India, 59 for men and 59.7 for women) (UN 1999). This pattern occurs in countries where women's social status is very low. Women are more likely to have poor nutrition, have less access to healthcare, are subject to frequent births, and have high maternal mortality (Doyal 1995).

3.4 Women Outlive Men by Eight or More Years

This pattern is exemplified by countries in the former Soviet Union and among blacks in the US, partly reflecting a deterioration in men's life expectancy since 1980. In Russia, expectation of life is only 58.3 for men and 71.7 for women (a 13-year difference); the corresponding figures for Latvia are 63.9 and 75.6 (UN 1999). In some Eastern European countries and the former Soviet Union, mortality rates have risen for men over the past 20 years, but fallen or remained stable for women, showing how men and women can be affected differently by social and economic changes (Hemstrom 1999). There have been greater threats to the economic and psychosocial well-being of men in these societies, many of whom have lost a secure 'bread-winning' role. The fall in life expectancy of US

black men relates to socioeconomic disadvantage, unemployment, alienation, drug abuse, and violence (Lorber 1997).

4. Explaining Gender Differences

A range of types of explanations have been put forward to account for gender differences in mortality and in health. However, there can be no simple interpretation, since women have health advantages in some aspects while men are advantaged in others.

(a) *Biological differences* associated with genetic factors, including hormonal differences, tend to advantage women. However, the range of gender differences in mortality identified in Sect. 3 clearly demonstrates the importance of factors other than biology.

(b) *Psychosocial differences* between men and women may be related to socially-created differences in personality, coping behaviors, and self-efficacy, which influence the experience and reporting of symptoms. These include gender role orientations, associated with masculinity and femininity, discussed in Sects. 1.1 and 2.3. Emslie et al. (1999) show that high 'femininity' scores are associated with poorer reported health.

(c) *Risk behaviors.* Epidemiologists are particularly concerned with gender differences in health promoting and health damaging behaviors. In most societies, men have higher levels of smoking and drinking, including more 'binge drinking.' However, in Western countries women have a better diet and participate more in preventive health measures such as screening. The growing trend for women to engage in smoking, drinking, and more risky behaviors has been termed the 'female emancipation hypothesis' (Hemstrom 1999).

(d) *Occupational and work-related factors.* Health hazards are associated with both paid work and unpaid domestic work. Working class men have traditionally been employed in more dangerous occupations with greater industrial health hazards. Much of women's paid work is repetitive in nature and lacks autonomy, which may create stress and repetitive strain injuries (Doyal 1995). In addition, unpaid domestic or family work may involve hard physical labor, such as in developing countries carrying water long distances and heavy agricultural production.

(e) *Social roles and relationships* differ between women and men and may impact on health. Women in Western countries often have better social networks than men, and can more easily rely on contact with close friends and relatives in times of crisis and stress. Marriage may provide a major source of social support, but in a different way for men and women and across societies. In terms of social support, marriage seems to be good for men but less so for women. Men tend to rely more on their wife for social support and a confiding relationship, with divorced

and widowed men reporting particularly poor self-assessed and psychosocial health.

(f) *Power and resources within the home.* Patriarchal control within the family and domestic settings may lead to exploitation of women and violence against them. Women's lack of power and influence in the home and lack of access to valued resources (of food, opportunities for leisure, income) may have adverse health consequences.

(g) *Social structural differences within society.* Women in most societies are more likely than men to be poor, have less education and live in disadvantaged material circumstances, all of which are likely to be implicated in gender differences in health.

(h) *Health service related factors.* Women may have less access to health services in developing countries, but in Western countries women generally make greater use than men, both because of using health care associated with reproduction and contraception, and the gendered meanings of illness discussed in Sect. 1.1. There may also be gender differences in diagnosis and treatment, see Sect. 2.2.

5. Gender Differences in Morbidity

It has become accepted wisdom that 'women are sicker, but men die quicker' (Lahelma et al. 1999). The higher morbidity rate but lower death rate of women was characterized by Nathanson (1975) as a 'contradiction' which required explanation. Other authors in the 1970s and 1980s repeatedly demonstrated that females had higher rates of illness than males. A number of the explanations put forward related to gender-appropriate behavior, for example, that women tend to over-report morbidity more frequently than men.

There have been recent changes; in the 1990s in European countries, relatively modest gender differences in overall measures of health and well-being were found, which are less important than the previous orthodoxy suggested (Macintyre et al. 1996, Lahelma et al. 1999). Macintyre et al. (1996) examined different measures of health using various datasets and found no female excess morbidity on many health measures.

However, major gender differences in mental health remain substantial. Women are more likely to report a range of neurotic disorders, for example, anxiety and depressive disorders, although when men's higher rates of alcohol and drug dependence are included, there is a much smaller gender difference in the prevalence of psychiatric disorders. Moreover, older women are more likely than older men to suffer from physical conditions which are nonfatal but result in chronic and disabling illnesses (Lahelma et al. 1999, Annandale and Hunt 2000). In Britain, nearly twice as many women as men over 65 suffer from functional impairments sufficient to require help on a daily basis to remain living in the community (Arber and Ginn 1991). A key concern is the extent of gender difference

in 'healthy life expectancy,' that is, the period of life without disability or chronic ill-health. Because of women's higher level of disability, the gender difference in 'healthy life expectancy' is smaller than in life expectancy.

6. *Structural Inequalities in Health Among Women and Men*

In the late 1970s and 1980s, many researchers began to focus on differences in health *among* women. Early work examined to what extent additional roles, such as the parental role and paid employment, had beneficial or adverse consequences for women's health (Arber 1991). While a growing body of research, particularly in Britain, studied inequalities in men's health.

More recently, researchers have stressed the importance of examining women *both* in terms of their structural position within society and their family roles. This approach examines how socioeconomic circumstances, together with marital and parental roles, influence the health of women and men. The key structural variables influencing health are occupational class, employment status, and educational qualifications. The effect of women's family roles on health depends on the material resources of her household (Arber 1991, 1997).

Changes in family structure since the 1970s, particularly the growth in lone parenthood and cohabitation, are relevant to gender differences in health inequalities. The previous orthodoxy that married women have poorer health than single women no longer holds in Britain (Annandale and Hunt 2000), possibly reflecting changes in the meaning of marriage in the late twentieth century and the opening of career opportunities for married women which previously only existed for single women. Lone mothers report the poorest health, largely due to living in disadvantaged material circumstances.

A gender analysis of health requires consideration of differences in the structural position of women and men within societies, in terms of power, occupational status, income, and lone parenthood. It is therefore essential to examine the extent to which gender differences in health may be explained by women's social, income, and political disadvantages.

7. *Gender and Health Across the Life Course*

During the twentieth century there have been radical changes in actual and expected gender roles, as well as in norms about age-related behavior. Gender roles and responsibilities are socially constructed in ways which vary across the life course. Thus, gender differences in health are likely to reflect gender socialization as to behavior at different stages of the life course. Thus the nature of inequalities in health for

men and women may vary for different age groups (Annandale and Hunt 2000).

The health and other characteristics of women and men are influenced by their prior life course. This is vividly seen in later life, since the financial well-being of older adults is intimately tied to their previous role in the labor market and private pension acquisition. A life course approach emphasizes the linkage between phases of the life course, rather than seeing each phase in isolation.

Research often neglects gender differences in health in childhood and later life, while the dominant concern is working age adults. During the working ages, family structure has a greater effect on women's than men's health. Married men and women report the best health, irrespective of whether they have children, but lone mothers report the poorest health. Class continues to be a major factor determining the health of working age British men and women. However, other structural variables are likely to become increasingly salient in the twenty-first century. Whether or not a person is in paid employment has become a key marker of health status, while educational qualifications increasingly differentiate health, especially among women.

In later life, previous occupational class is the key determinant of health for older British men and women (Arber and Cooper 1999). The similarity in the pattern of class inequalities in health for older women and men is surprising given that the majority of the current generation of older women have spent less than half their adult life in paid employment.

The nature of health inequalities between women and men is likely to differ over time and across societies, reflecting gender roles and relationships historically and cross-nationally. Taking a life course approach provides added insights into the mechanisms underlying the creation and perpetuation of gender inequalities in health.

See also: Gender and Cardiovascular Health; Gender and Health Care; Gender and Reproductive Health; Gender, Economics of; Gender Role Stress and Health; Men's Health; Reproductive Rights in Developing Nations; Women's Health

Bibliography

- Annandale E, Hunt K (eds.) 2000 *Gender Inequalities in Health*. Open University Press, Buckingham, UK
- Arber S 1991 Class, paid employment and family roles: Making sense of structural disadvantage, gender and health status. *Social Science and Medicine* **32**: 425–36
- Arber S 1997 Comparing inequalities in women's and men's health: Britain in the 1990s. *Social Science and Medicine* **44**: 773–87
- Arber S, Cooper H 1999 Gender differences in health in later life: The new paradox? *Social Science and Medicine* **48**: 63–76
- Arber S, Ginn J 1991 *Gender and Later Life: A Sociological Analysis of Resources and Constraints*. Sage, London

- Cameron E, Bernardes J 1998 Gender and disadvantage in health: Men's health for a change. *Sociology of Health & Illness* 20: 673-93
- Doyal L 1995 *What Makes Women Sick: Gender and the Political Economy of Health*. Rutgers University Press, New Brunswick, NJ
- Emslie C, Hunt K, Macintyre S 1999 Problematizing gender, work and health. *Social Science and Medicine* 48: 33-48
- Hemstrom O 1999 Explaining differential rates of mortality decline for Swedish men and women. *Social Science and Medicine* 48: 1759-77
- Koblinsky M, Timyan J, Gay J (eds.) 1993 *The Health of Women: A Global Perspective*. Westview Press, Boulder, CO
- Lahelma E, Martikainen P, Rahkonen O, Silventoinen K 1999 Gender differences in ill health in Finland. *Social Science and Medicine* 48: 7-19
- Lorber J 1997 *Gender and the Social Construction of Illness*. Sage, Thousand Oaks, CA
- Macintyre S, Hunt K, Sweeting H 1996 Gender differences in health: Are things really as simple as they seem? *Social Science and Medicine* 42: 617-24
- McKinlay J 1996 Some contributions from the social system to gender inequalities in heart disease. *Journal of Health and Social Behavior* 37: 1-26
- Nathanson C A 1975 Illness and the feminine role: A theoretical review. *Social Science and Medicine* 9: 57-62
- Sabo D, Gordon D F (eds.) 1995 *Men's Health and Illness: Gender, Power and the Body*. Sage, Thousand Oaks, CA
- Sharp I 1994 *Coronary Heart Disease: Are Women Special?* National Heart Forum, London
- United Nations 1999 1997 *Demographic Yearbook*. United Nations, New York
- Watson J 2000 *Male Bodies: Health, Culture and Identity*. Open University Press, Philadelphia, PA

S. Arber

Gender and Place

'Gender and place' as an interdisciplinary field of study emerged during the 1970s. It attempts to identify ways in which gender stereotypes and spatial arrangements reinforce each other. Over its brief history, the field has become increasingly differentiated with respect to definitions of both place and gender. Spaces that acquire meaning through use or history are 'places' consisting of cities, neighborhoods, buildings, and even the human body. 'Gender' extends beyond the dichotomy of masculine and feminine to encompass gays and lesbians in a continuum of identities.

1. Introduction

'Gender and place' is a nuanced term for 'sex and space.' Both phrases refer to the ways in which gender or sex-role stereotypes are built into, and reinforced by, spatial arrangements. Take the adage that 'woman's place is in the home.' When ideology dictates

that women should occupy only the domestic sphere, they seldom learn skills that would prepare them to enter public spaces like colleges and the workplace. Conversely, men who spend their days in corporate boardrooms rarely are expected to excel at housework and childcare. Feminine and masculine activities both shape, and are shaped by, built environments. Homes, for example, are still more likely to evoke feminine connotations than offices and factories regardless of the number of women in the labor force.

Social scientists use 'gender' to represent socially constructed masculine and feminine characteristics, while 'sex' refers to the biologically determined categories of male and female. One's sex is determined at conception, but one's gender identity develops over the life course and can fluctuate across a wide continuum of masculine and feminine characteristics. Girls who work on the family car in the garage and boys who cook meals in the kitchen are exhibiting gender characteristics incongruent with those expected of their sex; they are also occupying spaces inconsistent with gender expectations.

Place is to space as gender is to sex. Actual spaces can be designated with commonly understood boundaries, but 'places' transcend simple attempts at measurement. Places are spaces imbued with significance, just as gender adds meaning to sex. The space of a neighborhood park may be a place to play for children, a place to sleep for the homeless, and a place to walk the dog for apartment dwellers. The park can be all of these places simultaneously or sequentially. Like gender, place is socially and subjectively created from objective elements. The park may occupy a *space* the size of a city block, but it functions as a *place* for fun, shelter, or exercise depending on one's construction of its purposes.

Space and place are quite different concepts. Space is abstract and limitless, while place is bounded and refers to a setting infused with identity and memory. Another way to think of space is as a location or site of activity. 'Location' is objective, as when real estate agents identify a house by its distance from an elementary school. Place is subjective. One neighborhood close to a school might be described as a 'good place to raise children,' while another neighborhood, equally close, fails to receive the same endorsement. Such different assessments could result from the type of housing or size of yards, but they all affect the 'feel' of the place. In short, space and location are analyzed and measured; place is experienced.

The analysis of gender and place is interdisciplinary in that the field seeks to combine the spatial orientation of geographers, architects, and urban planners with the social focus of anthropologists, sociologists, and historians to understand the ways space and gender mutually reinforce one another. A brief review of the literature reveals that its focus has grown steadily more differentiated. Places range from the geographic

scale of the city or neighborhood to the human scale of the body. Gender identities include masculine, feminine, gay, lesbian, and an ambiguous 'third gender' of queer that recognizes gender identity as variable both historically and individually over the life course (Herdt 1996, Ingram et al. 1997, Smith and Windes 1999).

2. Origins of the Field

One of the first social scientists to address the social construction of space was Goffman (1959), who used the theater analogy of 'front-stage' and 'back-stage' behavior to describe the spatial nature of formal and informal activities. Harvey (1973), a geographer, proposed that urban spaces and social relations are so intricately linked that a new language should be invented to capture the simultaneous occurrence of spatial form and social processes. Harvey's emphasis on the city (and the fact that the majority of American life occurs in metropolitan areas) may have directed subsequent investigations of gender and place toward its current urban focus.

Contemporary intellectual interest in the issues surrounding gendered aspects of space arose during the 1970s. When feminist political activism was at its height, women were, quite literally, entering spaces previously open only to men. Colleges, factories, and offices became increasingly integrated by sex as women's place expanded beyond the home. Young women attending, and teaching at, universities during that era were among those who challenged the boundaries of traditional scholarship. The literal and symbolic collision of place and gender soon became a conscious subject of study.

It is not surprising, then, that the first publications on gender and place were edited volumes generated from academic conferences about sex and space. Four of the most important volumes appeared almost simultaneously. *New Space for Women*, edited by Wekerle et al. (1980) emerged after the 1976 United Nations Conference on Human Settlements met in Vancouver. The editors organized a session addressing the role of women in the design and planning of urban settlements because they recognized that women's problems with housing, transportation, urban planning, and lack of access to environmental decisions transcended categories of developed and developing nations. The next year, British anthropologist Ardener published *Women and Space: Ground Rules and Social Maps* (1981) as part of a series based on lectures sponsored by the Oxford University Women's Studies Committee. *Women and Space* included ethnographies of sex differences in the use of domestic and public spaces in the Andes, Greece, Iran, Russia, Nigeria, and South Africa.

In the United States, the Lincoln Institute of Land Policy sponsored a conference that resulted in the publication of *Building for Women* (Heath 1981) edited

by sociologist Keller. It dealt with housing, community, and open space from a variety of academic disciplines; one of the chapters summarized the US Department of Housing and Urban Development's Women and Mortgage Credit Project to facilitate women's ability to buy their own homes (Shalala and McGeorge 1981). HUD played a role in producing the fourth volume on women and space when it awarded a research grant to *SIGNS: Journal of Women in Culture* to commission essays for a special issue titled *Women and the American City*. The editor of *SIGNS*, Catherine Stimpson, subsequently published the volume (with Elsa Dixler, Martha Nelson, and Kathryn Yatrakis) through University of Chicago Press in 1981. Contributors constituted the pantheon of researchers on women and space: Galen Cranz, Dolores Hayden, Jacqueline Leavitt, Susan Saegert, and Gerda Wekerle.

Donna Shalala, Secretary of Health and Human Services during President Clinton's administration, worked behind the scenes to support research on gender and place. Shalala was at HUD when the Mortgage Credit Project began and when *SIGNS* received funding for its special issue. As president of Hunter College, Shalala co-chaired the Ford Foundation-Hunter College Women and Housing Seminar with urban planning professor Eugenie Ladner Birch. Members of the seminar investigated the housing needs of single mothers, the elderly, and working women from the standpoint of design and affordability. The fruits of that pedagogical venture resulted in *The Unsheltered Woman: Women and Housing in the 1980s*, edited by Eugenie Ladner Birch (Center for Urban Policy Research 1985), one of the first books to alert urban planners to the gendered nature of the housing market.

Milroy and Andrews strengthened the tradition of edited volumes with *Life Spaces: Gender, Household, Employment* (1988), a volume written entirely by Canadians about Canada. Milroy and Andrews introduced American audiences to geographer Suzanne Mackenzie, a founder of the field of feminist geography in the 1970s. Mackenzie, in turn, linked Canadian and American authors to British geographers Jo Foord, Sophie Bowlby, and Linda McDowell. *Life Spaces* departed from the model of previous edited volumes in one important respect: Wekerle's chapter on women's housing cooperatives included floor plans and photographs. Wekerle was one of the few social scientists to illustrate spatial concepts on the written page. During the first wave of scholarship, only Hayden and Birch (an architect and an urban planner, respectively) provided images of their topics. Today it seems obvious that books about space should include drawings and photographs, yet at the time authors rarely incorporated them.

Among pioneering scholars of gender and place, Dolores Hayden deserves special mention because she approaches the topic from an architectural rather than

a social science perspective. Hayden's *The Grand Domestic Revolution: A History of Feminist Designs for American Homes, Neighborhoods, and Cities* (1981) unearthed the work of nineteenth century 'material feminists' who established collective solutions to household labor and demanded pay for their work. Her next book, *Redesigning the American Dream: The Future of Housing, Work, and Family Life* (1984), addressed the disparity between the massive housing stock constructed after World War II and changing household demographics. It opened debate among architects and planners about appropriate designs for new American families. Franck and Ahrentzen's (1991) edited volume on collective, single parent, and single-room occupancy housing was a direct response to Hayden's challenge.

3. *The Place Pyramid: Boroughs, Buildings, and Bodies*

The first stage of research on gender and place addressed sex and space, as the titles of books published in the early 1980s suggest. They all included the word 'woman' and reference to an actual space such as the house or the city. Research focused on women as active participants in, and shapers of, the built environment traditionally considered men's domain. For example, Cranz (1981) chronicled women as users and suppliers of urban park services. This is what some feminist scholars call the 'add women and stir' phase; it restored women's presence for the historical record. This stage also focused on a fairly large *geographic scale*: that of cities and neighborhoods. For alliterative purposes that will become clear, I call this the *borough* stage.

During the late 1980s and early 1990s, focus expanded from the geographic scale of boroughs to include the *architectural scale* of buildings. Urban and neighborhood analyses were not abandoned. Rather, they were seen as the larger context in which the design of buildings reinforced stereotypical gender associations. Weisman (1992), an architect, for example, analyzed the sexual symbolism associated with specific buildings—skyscrapers, department stores, public housing—and advocated the redesign of large swaths of urban architecture to take women's needs and preferences into account. Betsky (1995) produced an architectural critique of the gap between the masculine characteristics of the exterior built environment and the feminine traits of its interiors.

A significant British book illustrating the place pyramid was Moore's *Space, Text, and Gender: An Anthropological Study of the Marakwet of Kenya* (1986). Moore drew on the theories of Foucault (1972) to analyze household space as a 'cultural text' representing gender relations in Marakwet society. The study placed the Marakwet in their geographic context with a map of Kenya and a diagram of their Sibou

village, and in their building context with photographs of their dwellings. Moore's work progressed from a base of great geographical expanse to the smaller scale of the building. *Space, Text, and Gender* marked another important milestone. It was one of the first full-length ethnographies devoted to the topic of gender and place. Although edited volumes would still fill a niche, the field acquired stronger intellectual credentials with the publication of books posing theoretical arguments.

The author's own work sought to extend that tradition. Heavily influenced by previous research and theory, *Gendered Spaces* (Spain 1992) was a cross-cultural and historical study proposing that women's status is lower in societies in which men and women are separated spatially in homes, schools, and workplaces. An architectural perspective on gender and place also appeared in 1992 with *Sexuality and Space* edited by Colomina (1992). Chapters by Colomina, Lynn Spigel, and Mark Wigley dealt with the home, while one by Elizabeth Grosz examined the relationship between the body and the city. Drawing from Foucault's (1977, 1979) emphasis on bodily pleasure and pain, Grosz introduced a concept developed later by Sennett in *Flesh and Stone: The Body and the City in Western Civilization*. Sennett's (1994, p. 370) thesis, that 'urban spaces take form largely from the ways people experience their own bodies,' set the stage for subsequent emphasis on the 'body as place.'

Nowhere has the constriction of the definition of place—from the large city to the small body—been more evident than in geography. Initially, geographers worked on topics like the sexual division of space in cities and homes (McDowell 1983) and the public/private locations of production and reproduction (Mackenzie and Rose 1983, Massey 1984). More recently, however, geographers have adapted Foucault's (1977) perspective on the body as text. They write articles on *The Body and Geography* (Longhurst 1995) and on the male torso as landscape (Nash 1996). Urban spaces that were once the focus of geographic analysis have become the stages on which bodies display different degrees of fear of violence (Koskela 1997, Mehta and Bondi 1999). McDowell (1999) legitimized this move toward the body by including it among the themes currently qualifying as feminist geography (along with place, home, community, city, nation-state, public space, and travel).

4. *Current Theory and Research: The Expansion of Gender*

As the concept of place has been expanded to include a larger range of sizes, so too has the definition of gender been stretched beyond its original association with the dichotomy of biologically determined sex. This enlargement inevitably introduced homosexuality into descriptions of the relationships between

(heterosexual) men, women, and space. For example, the original four edited volumes reviewed in this essay contained not a single chapter on homosexual gender identities. In contrast, the current literature contains numerous references to gays, lesbians, queers, bisexuals, and transsexual and transgendered persons (those born one sex who identify themselves as the other sex through anatomical (transsexual) or behavioral (transgendered) modifications). Gender now reflects a continuum of identities.

Castells's mapping of San Francisco's gay male culture was one of the first explorations into the relationship between homosexual subculture and space. Castells proposed that spatial concentration is fundamental to the development of gay identity. He argued that gay men are more likely to establish urban neighborhoods than lesbian women because men's tendency toward dominance extends to territorial aspirations, while lesbians' emphasis on relationships transcends space, making lesbians 'placeless' and more politically radical (Castells 1983, p. 140, Castells and Murphy 1982; see Forsyth (1997) for a challenge to this interpretation).

Researchers have analyzed how gays and lesbians have shaped residential, commercial, and public spaces. Gay men, and to some extent lesbian women, have played significant roles in the revitalization of American urban neighborhoods (Forsyth 1997, Lauria and Knopp 1985). Commercial areas catering to gays and lesbians, such as San Francisco's Market Street and Chicago's Halsted Street, display highly visible rainbow flags to attract clientele and, increasingly, tourists. Gay and lesbian community centers, bookstores, and health clinics have become staples in many cities; gay and lesbian parades and festivals temporarily transform city streets (Forsyth 2000). Pressure is mounting to nominate prominent gay and lesbian sites as historic landmarks, such as the New York City gay bar (the Stonewall Inn) where a police raid sparked a riot (Dubrow 1998). Such places have become easier to identify as gay and lesbian histories with a spatial focus have emerged. Chauncey (1994), for example, has identified how gay men appropriated public and private spaces to provide settings for the expression of gay culture in New York City between 1890 and 1940, while Kennedy and Davis (1993) have mapped the landscape of lesbian meeting places in Buffalo, NY between the 1930s and 1960s.

What does the future hold for the study of gender and place? It will definitely involve cyberspace, as the literature already suggests (Boyer 1999, Light 1995). As the matrix of cyberspace supplements the grid of the city, women may experience the environment differently from men. Possibilities for gender ambiguities abound on the Internet in 'chat rooms' that allow more anonymity than face-to-face interactions. The digital landscape has the potential to change women's status if it changes the way knowledge is disseminated. The gendered spaces that once limited

women's access to knowledge, and thus power, may dissolve.

A fitting end for this article is reference to a feminist critique of historic preservation titled *Susan B. Anthony Slept Here: A Guide to American Women's Landmarks* by Sherr and Kazickas (1994). Everyone knows that George Washington slept (and owned slaves) at his Mt. Vernon plantation, but few people know where Susan B. Anthony, leader of the movement for women's right to vote, worked or slept. The ironic title captures the trend in the literature toward greater gender and place ambiguity. The first wave of scholarship would have highlighted the gendered nature of the town where Susan B. Anthony lived. The second wave would have analyzed the home she slept in, while the third wave would ask with whom, if anyone, she was sleeping. If Susan B. Anthony were alive today, the fourth wave might explore whether it matters if her body occupies a particular physical space at all.

See also: Economic Development and Women; Economic Globalization and Gender; Gender and Environment; Gender and Feminist Studies in Sociology; Land Rights and Gender; Place in Geography; Spatial Equity; Spatial Interaction; Spatial Labor Markets; Transition, Economics of

Bibliography

- Betsky A 1995 *Building Sex: Men, Women, Architecture, and the Construction of Sexuality*. William Morrow, New York
- Boyer C 1999 Crossing cybertowns: Urban regions and the cyberspace matrix. In: Beauregard R, Body-Gendrot S (eds.) *The Urban Moment: Cosmopolitan Essays on the Late-20th-Century City*. Sage, Thousand Oaks, CA
- Castells M 1983 *The City and the Grassroots*. University of California Press, Berkeley, CA
- Castells M, Murphy K 1982 Cultural identity and urban structure: The spatial organization of San Francisco's gay community. In: Fainstein N, Fainstein S (eds.) *Urban Policy under Capitalism*. Sage, Beverly Hills, CA
- Chauncey G 1994 *Gay New York: Gender, Urban Culture, and the Making of the Gay Male World, 1890-1940*. Basic Books, New York
- Cranz G 1981 Women and urban parks: Their roles as users and suppliers of park services. In: Keller S (ed.) *Building for Women*. Heath, Lexington, MA
- Dubrow G 1998 Blazing trails with pink triangles and rainbow flags: New directions in the preservation and interpretation of gay and lesbian heritage. *Historic Preservation Forum* 12: 31-44
- Forsyth A 1997 'Out' in the valley. *International Journal of Urban and Regional Research* 21: 36-60
- Forsyth A 2000 Sexuality and space: Gays, lesbians, queers, and planning practice. Unpublished
- Foucault M 1972 *The Archeology of Knowledge* (trans. Sheridan A). Tavistock, London
- Foucault M 1977 *Discipline and Punish* (trans. Sheridan A). Allen Lane, London
- Foucault M 1979 *The History of Sexuality*. Allen Lane, London

- Franck K, Ahrentzen S 1991 *New Households, New Housing*. Van Nostrand Reinhold, New York
- Goffman E 1959 *The Presentation of Self in Everyday Life*. Doubleday, New York
- Harvey D 1973 *Social Justice and the City*. Johns Hopkins University Press, Baltimore, MD
- Herdt G (ed.) 1996 *Third Sex, Third Gender: Beyond Sexual Dimorphism in Culture and History*. Zone, New York
- Ingram G B, Bouthillette A-M, Retter Y (eds.) 1997 *Queers in Space: Communities/Public Places/Sites of Resistance*. Bay Press, Seattle, WA
- Kennedy E, Davis M 1993 *Boots of Leather, Slippers of Gold: The History of a Lesbian Community*. Penguin, New York
- Koskela H 1997 'Bold walk and breakings': Women's spatial confidence. *Gender, Place and Culture* 4: 301-19
- Lauria M, Knopp L 1985 Toward an analysis of the role of gay communities in the urban renaissance. *Urban Geography* 6: 152-69
- Light JS 1995 The digital landscape: New space for women. *Gender, Place and Culture* 2: 133-46
- Longhurst R 1995 The body and geography. *Gender, Place and Culture* 2: 97-105
- Massey D 1984 *Spatial Division of Labor: Social Structures and the Geography of Production*. Macmillan, London
- McDowell L 1983 City and home: Urban housing and the sexual division of space. In: Evans M, Ungerson C (eds.) *Sexual Divisions: Patterns and Processes*. Tavistock, London
- McDowell L 1999 *Gender, Identity, and Place: Understanding Feminist Geographies*. Polity Press, Cambridge, MA
- MacKenzie S, Rose D 1983 Industrial change, the domestic economy, and home life. In: Anderson J, Duncan S, Hudson R (eds.) *Redundant Spaces in Cities and Regions?* Academic Press, London
- Mehta A, Bondi L 1999 Embodied discourse: On gender and fear of violence. *Gender, Place and Culture* 6: 67-84
- Nash C 1996 Reclaiming vision: Looking at landscape and the body. *Gender, Place and Culture* 3: 149-69
- Smith R R, Windes R R 1999 Identity in political context: Lesbian/gay representation in the public sphere. *Journal of Homosexuality* 37: 25-45
- Weisman L K 1992 *Discrimination by Design: A Feminist Critique of the Man-made Environment*. University of Illinois Press, Urbana, IL

D. Spain

Gender and Reproductive Health

With the world's population increasing at an alarming rate, reproductive health has become the focus of growing attention in recent years. While reproductive concerns and activities also influence men's health, both scientific studies and health policies primarily focus on the female gender. Because women become pregnant and can bear children, the risks of reproduction present a greater threat to their health.

1. Definition of Reproductive Health

Although research has long examined how human reproduction affects overall health, the concept of 'reproductive health' is relatively new. In many in-

dustrialized countries, a clinical view continues to dominate, with reproductive health seen as a specialty of obstetrics and gynecology. Characteristic of this clinical view is its focus on negative outcomes and the separation of services dealing with reproductive issues (Graham 1998). Recent years, however, have seen the rise of a more comprehensive concept of reproductive health, formulated in 1994 at the International Conference on Population and Development (ICPD) in Cairo as follows: 'Reproductive health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity, in all matters relating to the reproductive system and to its functions and processes. Reproductive health therefore implies that people are able to have a satisfying and safe sex life and that they have the capability to reproduce and the freedom to decide if, when and how often to do so' (Program of Action 1995, p. 202).

As this definition makes clear, sexual health is considered as a central feature of reproductive health. In line with this very broad definition, the scientific literature discusses a large range of reproductive concerns, with a preponderance of reproductive problems and risks (Graham 1998). The most well-known health risks associated with reproduction are sexually transmitted diseases (STDs), including HIV/AIDS, maternal mortality, unsafe abortion, adolescent pregnancy, and infertility. Over the last several years, the comprehensive ICPD definition of reproductive health has come to include additional aspects such as female genital mutilation and violence against women.

2. Major Reproductive Health Risks

Reproductive health risks or health crises are the source of high morbidity and mortality rates worldwide. Most are avoidable. The prevalence of reproductive risks or 'outcomes' varies enormously among populations throughout the world. The greatest differences in negative outcomes exist between developed and developing countries. At the same time, the clear improvement in the reproductive health of women in developed countries is a relatively recent achievement and can be traced back primarily to key medical advances (introduction of antiseptics, discovery of sulfonamide, modern obstetrics), as well as changes in gender roles. The significant reproductive risks women in Europe and the United States continued to face even until the turn of this century are described by Shorter in *A History of Women's Bodies* (1982).

The following section will outline the spread of some major reproductive health crises. Although not all concerns (such as infertility) can be presented in detail, attention will be given to the gravest dangers in terms of individual and public health (see Miller and Rosenfield 1996). Despite decreasing fertility rates in many parts of the world, it has not been possible to reduce *maternal mortality*. According to WHO statistics, every year 585,000 women die as a result of

pregnancy complications. More than 98 percent of these deaths occur in the developing world (Spicehandler 1997). In 1991, WHO calculated the incidence of maternal death for a woman in North America at 1:4,000, while her counterpart in West Africa faced a risk of 1:18. Both high overall fertility rates and the high rate of adolescent pregnancies in less developed countries contribute to the high incidence of maternal mortality. Compared with women aged 20–24, the risk of maternal death for women between the ages of 10 and 14 is five times greater, and between the ages of 15 and 19 twice as high (Miller and Rosenfield 1996).

According to WHO estimates, about every one in four pregnancies is intentionally terminated. Of an estimated 45 million abortions performed each year, almost half (ca. 20 million) are performed under unsafe and septic conditions. Most of these risky abortions, in turn, occur in developing countries. Illegal and unsafe abortions are responsible for another spectrum of morbidity (for example, cervical trauma, pelvic peritonitis, jaundice, chronic pelvic pain); if worst comes to worst, these cases end fatally (an estimated 50,000–100,000 deaths each year). Here one of the most important determinants for mortality risk is the illegality of the abortion (Miller and Rosenfield 1996). But also the high rate of legal pregnancy termination, for example in Central and Eastern European countries (until the end of the 1990s, abortion was the most common form of birth control), is problematic and has serious consequences for women's reproductive health. Short-term and longer-term repercussions of repeated abortions or abortions performed under unfavorable conditions in Eastern Europe include higher incidences of anemia, infertility, and premature delivery (Bruyniks 1994).

The prevalence of STDs has serious consequences for the health of women and their unborn children. The WHO estimates that more than 330 million cases of treatable STDs (such as chlamydia infection, gonorrhea, syphilis) occur each year, the vast majority in the developing world, particularly in sub-Saharan Africa (Drennan 1998). There are two reasons why STDs in general are understood as more serious for women than for men (Drennan 1998, Miller and Rosenfield 1996, Program of Action 1995): (a) during vaginal intercourse, the likelihood of infection from man to woman is higher than from woman to man, and (b) STDs in women are often not diagnosed and treated until the disease is in a later (and thus more serious) stage because in women the symptoms are often less conspicuous in their early stages (see Sexually transmitted diseases).

Among STDs, HIV and AIDS are particularly threatening; at the end of 2000, the HIV virus had infected almost 36.1 million people worldwide, about 18.3 million men and 16.4 million women as well as about 1.4 million children (UNAIDS/WHO 2000); the majority of these cases involve people in deve-

loping countries. In some parts of the world, the AIDS epidemic has already dramatically shortened the average life expectancy (Drennan 1998). While HIV and AIDS initially presented more of a threat to men than to women, there are now countries, including the United States and several sub-Saharan African nations, in which HIV is spreading faster among women than among men.

3. Determinants of Reproductive Risk Behavior

A range of factors influences reproductive health. Particularly important is the quality of medical care during pregnancy and birth. The main reason for the high reproductive morbidity and mortality rates in developing countries is found in inadequate health-care structures. Inadequate contraceptive behavior (or the lack thereof) is also responsible for a variety of risks to reproductive health (in particular unintended pregnancy and STDs); for this reason, special attention is paid to risky sexual behavior in connection with reproductive health (see *Sexual Risk Behaviors*).

Contraceptive behavior is influenced by a variety of internal and external factors, summarized in a model designed by Bruyniks (1994). Here knowledge of effective contraception and attitudes toward various methods of contraception are postulated as essential determinants of contraceptive behavior. Knowledge and attitudes are influenced by internal factors (age, intelligence, moral values) and external factors (access, social environment). It is precisely in less developed countries that access to contraceptive services is frequently blocked by financial, geographic, legal, or other barriers. Psychosocial acceptance or rejection of certain contraceptive methods in the social environment also plays a decisive role (Miller and Rosenfield 1996). Another issue is whether knowledge and positive attitudes toward effective contraception are in fact reflected in reproductive behavior, and depends on such personality variables as assertiveness or self-efficacy (Schwarzer and Fuchs 1995), as well as external barriers such as financial dependence (Wyatt 1994) or physical violence.

The social status of women is an essential factor in determining their contraceptive behavior. Social status influences not only the knowledge of efficient contraceptive methods and attitudes toward contraception, but also the possibilities for women to prevail in decisions for (or against) sexual activity and contraception, if need be against their male partner's wishes. Social status is a multidimensional concept; in terms of reproductive health, the educational level of women is the central and most frequently studied dimension. Also relevant are aspects such as political participation, employment and earnings, economic autonomy, as well as the reproductive rights of women (Kawachi et al. 1999). Williamson and Boehmer (1997) analyzed different status variables as predictors of

female life expectancy in 97 less developed countries. They found a positive correlation between the status of women and female life expectancy. Again, aspects of women's status apart from their educational level were important for determining female life expectancy. Of these other dimensions of women's status, reproductive autonomy had a particularly strong effect. This reproductive autonomy is operationalized via contraceptive prevalence as an indicator of women's control over their reproductive behavior.

The influence of social status on the (reproductive) health of women can also be found in developed countries. One recent analysis examined how the morbidity and mortality of women in 50 American states is related to different indicators of social status (Kawachi et al. 1999). Higher political participation was associated not only with lower overall female mortality but also with lower rates of death from different causes, including death due to reproductive risks such as cervical cancer or infant mortality. This analysis also examined an index for 'reproductive rights' ('reflecting reproductive well-being and autonomy') in connection with indicators of health. The index was operationalized in eight legal and political indicators reflecting the reproductive well-being and reproductive autonomy of women (for example, access to abortion services without a waiting period, public funding of infertility treatments, existence of a maternity stay law, whether gay or lesbian couples can adopt, etc.). For this indicator, however, no relationship to mortality was established, although a minor correlation to self-reported illness was found ('self-reported days of activity limitations').

Along with social status, *gender roles* have an important influence on reproductive decision-making and behavior. Gender roles refer to 'normative expectations about the division of labor between the sexes and to gender-related rules about social interactions that exist within a particular cultural-historical context' (Spence et al. 1985, p. 150). Traditional gender roles, which assign women the subordinate and dependent role in the partnership, are regarded as a particular risk for reproductive health. As Drennan (1998) shows in a detailed overview, in many developing countries men are the primary decision-makers about sexual activity, fertility, and contraceptive use. It is assumed that women who are not permitted to reject sexual intercourse or to decide about contraception have little means to influence their reproductive health. Cultural norms about femininity or masculinity might also harbor reproductive risks. In many Latin American or Caribbean countries, for example, promiscuous behavior among men is regarded as proof of their masculinity (Drennan 1998). The risky sexual behavior of their (marriage) partners frequently endangers women by increasing such risks as STD transmission. In many countries, recently married women continue to face strong pressure to 'prove' their fertility; contraception is rarely practiced,

and even adolescent or very young women become pregnant with the well-known attendant risks to the health of mother and child. In India, for example, 40 percent of all women between the ages of 15 and 19 are already married, and of these only 7 percent use contraceptives (Jejeebhoy 1998).

Gender inequality is expressed in various ways, including fewer educational opportunities for women, violence against women, female genital mutilation, and economic inequities (Miller and Rosenfield 1996). Some examples of gender inequality are concentrated on less developed countries, while others, such as violence against women and economic inequalities, are also found in developed countries. Gender roles are not independent of women's social status: improvements in social status, especially better education, generally go hand in hand with changes in traditional gender roles.

4. Promoting Reproductive Health

The ICPD Program of Action contains wide-ranging recommendations and calls for improving the reproductive health of women, especially in developing countries. What is innovative about this program in comparison to earlier approaches is that efforts to achieve this goal do not rely solely on isolated measures such as better access to contraceptive services. Rather, the program serves as a reminder that the essential requisite to bettering reproductive health is a fundamental improvement in the social status and rights of girls and women. The ICPD Program of Action lists the following points:

(a) Access to appropriate health care services: prenatal and postnatal care, obstetrics and emergency obstetrics care, treatment of STDs, abortion by qualified medical practitioners.

(b) Availability of contraceptive services: access to safe, effective, affordable, and acceptable methods of family planning.

(c) Sex education, including education for the adolescent population.

(d) Strengthening the legal, economic, and social position of girls and women.

The recommended measures of the ICPD conference continued to direct the greatest share of attention to women as the major agents of change. In the meantime, focus has shifted to include the male partner since the critical behavior unit is the sexually active couple (Becker and Robinson 1998). Overviews of research in this area (for example Becker 1996) show that reproductive health interventions (especially family planning) targeted at couples are more effective than those targeted at one or the other sex. The pair approach to reproductive health does, however, presume a certain level of equality between man and woman. In countries where very patriarchal structures continue to predominate, it is sometimes necessary to

provide initial separate services and consultations to men and women. In their report on an intervention program in Bangladesh, for example, Becker and Robinson relate how the male villagers first had to be convinced to 'allow' their wives to use contraceptive devices (Becker and Robinson 1998).

Improving the couple's communication is an essential way to increase men's participation and to promote reproductive health (Drennan 1998). In many less developed countries, only a minority of spouses talk about issues like sexuality and family planning. Demographic and Health Surveys in Nigeria, Senegal, and Pakistan conducted in the 1990s, for example, report that only one-quarter of men currently married had discussed family planning with their wives. The percentages in Cameroon (31 percent) and Ghana (37 percent) were not much higher. Drennan (1998) discusses social and cultural factors that obstruct couple communication. It is taboo in many societies to talk about sexuality. In addition, the inferior status of women often found in traditional societies continues to aggravate communication.

Many recent approaches to promoting reproductive health and preventing unintended pregnancies and STDs/HIV share a perceptible shift from models of behavioral change centered on the individual to attempts to include the partner and to consider the social network. This development can be traced back to the increasing awareness that sexual/reproductive behavior is strongly determined by social norms (Bond et al. 1999).

See also: Family and Gender; Fatherhood; Rape and Sexual Coercion; Reproductive Rights in Affluent Nations; Reproductive Rights in Developing Nations; Sexual Risk Behaviors; Sexuality and Gender; Sexually Transmitted Diseases: Psychosocial Aspects

Bibliography

- Becker S 1996 Couples and reproductive health: A review of couple studies. *Studies in Family Planning* **27**: 291–306
- Becker S, Robinson J C 1998 Reproductive health care: Services oriented to couples. *International Journal of Gynecology and Obstetrics* **61**: 275–81
- Bond K C, Valente T W, Kendall C 1999 Social network influences on reproductive health behaviors in urban northern Thailand. *Social Science and Medicine* **49**: 1599–614
- Bruyniks N P 1994 Reproductive health in Central and Eastern Europe: Priorities and needs. *Patient Education and Counseling* **23**: 203–15
- Drennan M 1998 Reproductive health. New perspectives on men's participation. *Population Reports Journal* **46**: 1–35
- Graham W J 1998 Outcomes and effectiveness in reproductive health. *Social Science & Medicine* **47**: 1925–936
- Jejeebhoy S J 1998 Adolescent sexual and reproductive behavior: A review of the evidence from India. *Social Science & Medicine* **46**: 1275–2190

- Kawachi I, Kennedy B P, Gupta V, Prothrow-Stith D 1999 Women's status and the health of women and men: A view from the States. *Social Science & Medicine* **48**(1): 21–32
- Miller K, Rosenfield A 1996 Population and women's reproductive health: An international perspective. *Annual Review of Public Health* **17**: 359–82
- Program of Action of the 1994 International Conference on Population and Development 1995. *Population and Development Review* **21**: 187–213
- Schwarzer R, Fuchs R 1995 Changing risk behaviors and adopting health behaviors: The role of self-efficacy beliefs. In: Bandura A (ed.) *Self-Efficacy in Changing Societies*. Cambridge University Press, New York
- Shorter E 1982 *A History of Women's Bodies*. Basic Books, New York
- Spence J T, Deaux K, Helmreich R L 1985 Sex roles in contemporary American society. In: Lindzey G, Aronson E (eds.) *The Handbook of Social Psychology*, Vol. II, 3rd edn. Random House, New York
- Spicehandler J 1997 Issues to consider in operationalizing reproductive health. *Advances in Contraception* **13**: 345–50
- UNAIDS/WHO 2000 *AIDS Epidemic Update: December 2000*. UNAIDS/00.44E-WHO/CDS/CSR/EDC/2000.9. <http://www.unaids.org>
- Williamson J B, Boehmer U 1997 Female life expectancy, gender stratification, health status, and level of economic development: A cross-national study of less developed countries. *Social Science and Medicine* **45**: 305–17
- Wyatt G E 1994 The sociocultural relevance of sex research: Challenges for the 1990s and beyond. *American Psychologist* **49**: 748–54

M. Sieverding

Gender and School Learning: Mathematics and Science

Scholarship related to gender, mathematics, and science has been extensive since the early 1970s and has accounted for a sizeable percentage of the published research in the fields of mathematics and science education. In response to concerns about the small numbers of females in upper level mathematics and science courses, the research started with the documentation of gender differences in participation and achievement, moved to the identification of related variables, and more recently has included the use of feminist and cognitive science research paradigms. A brief synthesis of the literature will set the stage for a consideration of the latest research in the field.

1. Documenting Gender Differences

1.1 Differences in Learning

It is impossible to draw firm conclusions about anything as global as overall gender differences in

mathematics and science learning. Data have been reported about a variety of populations, and measured in a variety of ways in diverse situations. The magnitude of gender differences reported varies depending on the content and/or form of the particular instrument used to assess learning. When open-ended forms of assessment are used, females tend to get higher marks while males tend to be more successful on multiple choice tests where time is a factor. For example, in Australia, gender differences in favor of males have often been reported on a multiple choice test of physics. But when the Australian state of Victoria changed the secondary school physics assessment to include extended response items and social context questions, females outperformed males (Hildebrand 1995). Females outperform males when course grades given by teachers are used as an indicator of learning (Kimball 1989).

The type of content measured also has an influence on gender differences reported. Females appear to perform as well as males do on overall measures of mathematical knowledge, but males have consistently outperformed girls on measures of complex problem solving (Hyde et al. 1990). In the Third International Mathematics and Science Study (TIMSS) study (Beaton et al. 1996), middle school females and males scored approximately the same on their test of mathematics achievement, but females scored significantly lower on the test of scientific achievement. In science, girls tend to do as well as boys in biology but far less well in physics and chemistry.

The diversity of the population being measured affects the findings. Some have suggested that gender differences are more prevalent in white populations, citing examples of populations where females outperform males at all levels. While there tended to be gender differences in favor of males in most of the participating countries of the TIMSS at the secondary level, the size of gender differences varied tremendously from country to country. It should be noted that for this portion of the study only European and American countries, predominantly composed of white citizens, participated.

To increase the complexity, the magnitude of gender differences over time may have changed. In the United States, there was some evidence of pervasive gender differences in favor of males in the 1970s (Fennema and Sherman 1977). By the 1990s, Kahle and Meece (1994) suggest that there was some evidence that gender differences in overall mathematics achievement had declined in the USA while in science they had increased. To summarize, in general males and females perform differently on measures of achievement in mathematics and science. But, while males tend to score higher on more complex measures of mathematics and science achievement, it is impossible to generalize about overall gender differences in mathematics and science learning across type of test, content, and population.

1.2 Differences in Participation

The data about gender differences in participation in mathematics and science is clearer than that regarding student learning. Females constitute diminishing proportions of the students in mathematics and science classes when these subjects become elective. In all of the countries participating in the TIMSS, with the exception of the Czech Republic and Austria, fewer females reported taking physics and advanced mathematics in secondary school than did males. The problem becomes more severe in postsecondary institutions where females constitute a small percentage of majors in engineering, physics, chemistry, and computer sciences. The differences are even more accentuated in adult life where women are dramatically underrepresented in the professional fields of engineering and information technology.

1.3 Gender Differences and Attitudes

Many studies have investigated several well-defined attitudes as they relate to mathematics and science as well as the relationship of these attitudes to learning and participation in mathematics/science. When compared to males, females have consistently reported lower levels of self-confidence in doing mathematics and science, as well as lower levels of perceived usefulness of mathematics and science. Results also have indicated consistently that males tend to report that mathematics is a male domain more than females have, while both males and females consider science to be a male domain. To illustrate this perception, students around the world typically depict a male when asked to draw a scientist. Girls tend to attribute their success to hard work rather than ability. They view their failures as evidence of their lack of ability in contrast to boys who attribute their success to high ability and their failures to bad luck or lack of effort. While a great deal of research has documented these differences in attitude and attributional style, little has been done to demonstrate if and how this affects the learning of mathematics and science (Leder 1992).

2. Understanding the Development of Gender Differences: Studies of Classroom Environments and Teachers

Although females and males sit in the same classrooms, they have different experiences in those classrooms. Researchers have observed that teachers, almost without exception, spend less time with girls than with boys. This finding has been replicated around the world in a variety of studies in both mathematics and science classrooms. Researchers have gone beyond just looking for the number of interactions with girls and boys and have studied the

types of teacher/pupil interactions. They have reported that teachers tend to ask females more routine questions for which they only have a short time to reply, while males are asked more complex questions and are given more time and encouragement to respond. Males tend to dominate the use of laboratory equipment in science classrooms and to take control of cooperative groups (Leder 1992, Kahle and Meece 1994).

While the classroom experiences of boys and girls differ, researchers have not been successful at measuring the correlation of these differences with learning. One classroom variable that has accounted for gender differences involves competition. Girls appear to learn more when classrooms emphasize cooperative activities and boys appear to learn more when classrooms emphasize competitive activities. Concern about the disparate experiences of boys and girls in coeducational settings has led some to advocate single-sex classrooms for mathematics and science. There is no evidence to indicate that girls' achievement in mathematics is significantly improved by being placed in single-sex classrooms (Leder and Forgasz 1998).

The attitudes and beliefs of teachers about their male and female students have also been studied. Teachers appear to be more aware of their male students than their female students. Teachers attribute the success of their male students to ability and the success of their female students to hard work, and hold many of the same perceptions as their students by viewing mathematics and science as masculine subjects. Based on the findings of the research on differences in attitude and classroom experiences, many educational intervention programs have been developed and implemented in various levels of schooling. While these have resulted in local change, the persistence of gender differences indicates that new perspectives on research in the field are needed to affect wide scale change.

3. Scholarship in the 1990s

During the 1990s, many changes have affected scholarship in the field. Feminist perspectives have become prominent. The field of cognitive science has matured, and mathematics and science pedagogy is changing. The impact of each of these developments will be explored in this section.

3.1 Feminist Perspectives

Many feminist scholars are critical of past research and have argued convincingly that there is a flaw in the assumption that before equity in education can be achieved, females must change. For example, many have interpreted past research to indicate that since

females lack confidence in their ability, the way to decrease gender differences in mathematics is to help females develop increased confidence, assuming females are dysfunctional because they lack confidence. Feminist scholars have suggested diverse but complementary interpretations of how females in mathematics and science education can be viewed. One feminist view critiques the subject matter, pointing out that females have not been engaged in the construction of mathematics and science. These subject matters must be redefined so that they are not alienating to females (Burton 1994, Harding 1991). It follows from this, that mathematics and science must become viewed as constructed by humans rather than as fixed and objective.

Another view suggests that instruction that has been predominately organized around a male perspective should be reorganized to include a female perspective. Several specific reorganizations have been suggested. One such reorganization has been to have single-sex classrooms where females will not be competing with males, and in which instructional activities which capitalize on 'women's ways of knowing' will be used (Rossi-Becker 1995). Another reorganization includes examination of the ways in which mathematics and science contribute to reproducing social inequality and the marginalization of females. Socially relevant problems form the basis of the curriculum providing females with the knowledge of the use of mathematics and science in political arenas (Mayberry 1998).

Recently feminist scholars have recognized that gender issues affect males as well as females. Males feel pressure from their peers to misbehave in class and find that high academic achievers are treated as social outcasts. These masculine stereotypes interfere with males' learning in all subject areas and have become the focus of a great deal of concern. (See Epstein et al. 1998.)

3.2 Perspectives from Cognitive Science

Cognitive science is a rich, complex field that defies definition. At its simplest, it has to do with the study of covert mental activities and the construction of models of mental functioning. Within mathematics and science education, cognitive science research has focused on the study of the mental strategies that are done as an individual solves problems or develops understanding of phenomena. Most work has been done with students in a specifically defined domain such as addition/subtraction or physics. Paths of the development of understanding within these domains have been identified that reflect development in many diverse cultures. However, there has been little work related to gender because it has been assumed that these paths were universal to all humans. Recently this assumption has been called into question in mathematics.

Until a few years ago, it was assumed that although children's solution strategies for addition and subtraction matured at different rates, the developmental paths for girls and boys were similar. However a recent longitudinal study called this assumption into question (Fennema et al. 1998). Results indicated that as children moved from first grade to third grade, girls tended to use immature strategies (counting and modeling) longer while boys tended to use more abstract invented strategies earlier. A significant number of girls skipped the use of these abstract strategies almost entirely, moving to standard computational algorithms before understanding of the concepts underlying the procedures had developed.

In spite of the speed with which they learned to perform the routine procedures, this developmental sequence appeared to interfere with the girls' work on difficult problems, and by third grade, boys were better able to solve more difficult problems than the girls. The researchers suggested that the girls were failing to develop their understanding in mathematics and this might presage the gender differences that are found in complex problem solving at about adolescence. These studies have alerted researchers that gender differences in mathematics may emerge at an earlier age than previously shown. This focus on how students think about problems shows promise for uncovering one of the sources of gender differences on solving complex problems. The study does not offer any evidence about why these differences developed. However, the reporting of the study was accompanied by interpretative comments by researchers with different scholarly backgrounds. They hypothesized that the differences were due to differential learning environments that existed in the classrooms.

3.3 Changes in the Teaching of Mathematics and Science

Educational scholars have been encouraging school-teachers to change the focus of their instruction to conceptual understanding rather than memorization of facts, rules, and procedures. Class discussion has become an integral part of this new form of instruction, where children have opportunities to share their problem-solving strategies and scientific observations. The curriculum includes ill-defined problems and complicated tasks which can be approached in a variety of ways.

Many educators believed that there would be fewer gender differences in classrooms with an emphasis on understanding. An examination of the research that has been done in nontraditional classrooms focusing on gender shows that this has not always been the case. It seems that when there is an emphasis on understanding, girls' performance improves but often the boys' performance improves as well, failing to ameliorate differences. New curricula do not always appeal

to girls any more than old curricula. When girls actively engage with science equipment doing experiments their performance tends to improve, but girls are often denied this opportunity by their male peers. It seems that to insure gender equity is achieved in these nontraditional classrooms, teachers must pay particular attention to insuring full participation of all children.

4. Conclusion

Few dispute that gender differences in participation in mathematics and science careers have existed for a long time and still exist. However, there is debate about the size and importance of the differences found in learning. Some believe that gender differences in mathematics and science learning indicate a pervasive educational inequity that demand major interventions in the educational system. Others believe the differences are inconsequential, and that attending to these small differences might exacerbate the problem by contributing to an erroneous perception that mathematics and science are male domains. Still others argue that mathematics and science tend to be privileged subject matters and are accorded more respect than they deserve because of their masculine/abstract qualities. They assert that society (and scholars in the area) should not be concerned about gender differences in mathematics and science but instead should celebrate interests that are more typically feminine. But, a majority of scholars in this field (including the authors of this article) believe that the gender differences found in school learning are both important and significant. They contribute to limiting the choices available to females as they progress through school, into adult life and careers. As long as this is the case gender differences are worthy of study so that equity in mathematics and science education can be achieved.

See also: Coeducation and Single-sex Schooling; Education and Gender: Historical Perspectives; Education (Higher) and Gender; Education (Primary and Secondary Schools) and Gender; Gender and School Learning: Mathematics and Science; Mathematical and Logical Abilities, Neural Basis of; Mathematical Education; Rationality and Feminist Thought; Science Education; Sex-role Development and Education; Spatial Cognition

Bibliography

- Beaton A, Mullis I, Martin M, Gonzalez E, Kelly D, Smith T
1996 *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study*. Center for the Study of Testing and Evaluation and Educational Policy, Boston College, Boston

- Burton, L 1994 Clashing epistemologies of mathematics education: Can we see the wood for the trees? *Curriculum Studies* 2(2): 203–19
- Epstein D, Elwood J, Hey V, Maw J 1998 *Failing Boys? Issues in Gender and Achievement*. Open University Press, Buckingham, UK
- Fennema E, Carpenter T, Jacobs V, Franke M, Levi L 1998 A longitudinal study of gender differences in young children's mathematical thinking. *Educational Researcher* 27: 6–11
- Fennema E, Sherman J 1977 Sex-related differences in mathematics achievement, spatial visualization and sociocultural factors. *American Educational Research Journal* 14: 51–71
- Harding S 1991 *Whose science? Whose knowledge? Thinking for women's lives*. Open University Press, Buckingham, UK
- Hildebrand G 1995 Assessment interacts with gender: The case of girls and physics in Victoria, Australia. Paper presented at the UNESCO/Institute of Education Colloquium: Is there a Pedagogy for Girls?, University of London
- Hyde J, Fennema E, Lamon S 1990 Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin* 107(2): 139–55
- Kahle J B, Meece J 1994 Research on gender issues in the classroom. In: Gabel D (ed.) *Handbook on Science Teaching and Learning*. Macmillan, New York, pp. 542–57
- Kimball M M 1989 A new perspective on women's math achievement. *Psychological Bulletin* 105(2): 198–214
- Leder G 1992 Mathematics and gender: Changing perspectives. In: D. Grouws (ed.) *Handbook of Research on Mathematical Teaching and Learning*. Macmillan, New York, pp. 597–622
- Leder G C, Forgasz H 1998 Single-sex groupings for mathematics: An equitable solution? In: Keitel C (ed.) *Social Justice and Mathematics Education*. Freie Universität Press, Berlin, pp. 162–79
- Mayberry M 1998 Reproductive and resistant pedagogies: The comparative roles of collaborative learning and feminist pedagogy in science education. *Journal of Research in Science Teaching* 35(4): 443–59
- Rossi-Becker J 1995 Women's ways of knowing mathematics. In: Rogers P, Kaiser G (eds.) *Equity in Mathematics Education: Influences of Feminism and Culture*. The Falmer Press, London, pp. 163–74

R. Ambrose and E. Fennema

Gender and Technology

Since the 1980s an exciting new field of study has emerged, concerned to develop a gender perspective on technology. Feminists have identified men's monopoly of technology as an important source of their power; and women's lack of technological skills as an important element in their dependence on men. There has been a tension between the view that most new technologies are oppressive and controlling of women, and the obverse view that technology would liberate women from drudgery. A key issue here is whether the problem lies in men's domination of technology, or

whether the technology is in some sense inherently patriarchal. This debate resonates with, but is more recent than, the gender questions articulated in relation to science (Harding 1986).

1. Definition of Technology is Gendered

The definition of technology is cast very much in terms of male activities. Technology tends to be thought about in terms of industrial machinery and cars, for example, ignoring other technologies that affect most aspects of everyday life. The history of technology still represents the prototype inventor as male. As a result, women's contributions have by and large been left out of technological history. So, as in the history of science, an initial task of feminists has been to uncover and recover the contribution of women technologists who have been hidden from history. There is now evidence that during the industrial era, women invented or contributed to the invention of such crucial machines as the cotton gin, the sewing machine, the small electric motor, the McCormick reaper, and the Jacquard loom (Stanley 1995).

A greater emphasis on women's activities suggests immediately that women, and in particular black women, were amongst the first technologists. After all, women were the main gatherers, processors, and storers of plant food from earliest human times. It was therefore logical that they should be the ones to have invented the tools and methods involved in this work, such as the digging stick, the carrying sling, the reaping knife and sickle, pestles, and pounders. If it were not for the male orientation of most technological research, the significance of these inventions would be acknowledged. By diminishing the significance of 'women's sphere' technologies, the cultural stereotype of technology as an activity appropriate for men is reproduced.

2. Technological Change and the Sexual Division of Labor

A major concern of feminists has been the impact of technological change on women's lives, particularly in the context of paid work. The introduction of computer-based technologies into offices is a prime site of this research, mainly because the majority of clerical and secretarial workers almost everywhere are women. This research examines the effects of technological change on women's employment opportunities, their experience of work, and their skills (Hartmann et al. 1986, 1987). The exploitation of Third World women as a source of cheap labor for the manufacture of computer components has also been scrutinized (Mitter and Rowbotham 1995).

Much of this research has been influenced by a theoretical perspective, mainly Marxist in orientation, which identifies class conflict as a major factor affecting the technologies of production (Braverman 1975). In order to control the labor force and maximize profitability, capitalism continuously applies new technology designed to fragment and deskill labor, so that labor becomes cheaper and subject to greater control. Rather than technical inventions developing inexorably, machinery is understood to be the result of capitalist social relations. Building on this analysis, feminists questioned the notion that control over the labor process operates independently of the gender of the workers who are being controlled (Cockburn 1983). They pointed out that the division of labor characterizing paid jobs is a sexual hierarchy, and that its gendered nature is not incidental. Both employers as employers, and men as men, have an interest in creating and sustaining occupational sex segregation. Gender is shown time and time again to be an important factor in shaping the organization of work that results from technological change.

From this historical perspective, women's exclusion from technology may be seen as a consequence of the male domination of skilled trades that developed during the Industrial Revolution. Male craft workers actively resisted the entry of women to areas of technological work to protect their own conditions. This rigid gender division of labor within the factory meant that the machinery was designed with men in mind, either by the capitalist inventor or by skilled craftsmen. Industrial technology from its origins thus reflects male power, as well as capitalist domination. The continuing underrepresentation of women in engineering, and indeed all scientific and technical institutions, is a legacy of this history.

In the domestic sphere, by contrast, women are the main users of technology. Considerable optimism has attached to the view that technology may provide the solution to the drudgery of housework. Women's unpaid work in the home, servicing men, children, and other dependants, has for a long time been seen by feminists as a key to women's subordination. Since the 1980s, with the recognition of housework as work, research on domestic technology has grown rapidly, especially among historians (Cowan 1983).

Such research has challenged the main orientation within the sociology of technology towards paid, productive labor in the public domain. Issues that have been central here are the relationship between domestic technologies and time spent on household labor; whether technology has affected the degree of gender specialization of housework; and gender bias in the use of new technologies. Dominating the debates is the knowledge that the amount of time women spend on household tasks has not decreased with 'mechanization of the home.'

This is the central theme of Cowan's (1983) study on the failure of the 'industrial revolution in the home' to

ease or eliminate household tasks. She points to the contradictions inherent in attempts to mechanize the home and standardize domestic production. Such attempts have foundered on the nature of housework—privatized, decentralized, and labor-intensive. The result is a completely 'irrational' use of technology and labor within the home because of the dominance of single-family residences and the private ownership of correspondingly small-scale amenities. Domestic technologies clearly reflect the sexual division of domestic activities and the social organization of the family.

3. *Technology Based on Women's Values?*

By the late 1980s, attention was tending to shift away from the focus on women and technology. Instead it began to examine the very processes by which technology is developed and used, and those by which gender is constituted. In other words, feminists were exploring the gendered character of technology itself. This approach has taken two broad directions: that which is influenced by radical feminism, and that which is more in tune with the social studies of technology.

According to the former view, Western technology itself embodies patriarchal values and its project is the domination and control of women and nature (Merchant 1980, Mies and Shiva 1993). This approach has been particularly influential in relation to the technologies of human biological reproduction. It is fueled by the perception that the processes of pregnancy and childbirth are directed and controlled by ever more sophisticated and intrusive technologies. Strong opposition to the development of the new reproductive technologies has been voiced by radical feminists, who see these technologies as a form of patriarchal exploitation of women's bodies (Corea et al. 1985). Central to this analysis is a concept of reproduction as a natural process, inherent in women alone, and a theory of technology as intrinsically patriarchal.

In a similar vein, ecofeminists have analyzed military technology and the ecological effects of other modern technologies as products of a violent patriarchal culture. Technology, like science, is seen as an instrument of male domination of women and nature. Just as many feminists have argued for a science based on women's values (see *Standpoint Theory, in Science*), so too there has been a call for a technology based on women's values. From this perspective, a new feminist technology would be based on a nonexploitative relationship between nature and humanity, and would embrace feminine intuition and subjectivity.

The representation of women as inherently nurturing and pacifist has been subjected to a variety of critiques (see *Feminist Epistemology*). The idea of a technology based on women's values has thus lost

much of its salience. Ecofeminism relies on an idealized femininity and overlooks the role of culture and history in shaping women's needs and priorities in different contexts. It also negates the benefits of modernization in favor of a mythical unchanging tradition.

4. Gender Relations Shape Technology

Current theory and research on gender and technology emphasize the specific characteristics of each technology and the differing positions of women in relation to them. While maintaining that male interests have structured profoundly the form of technologies that have become available, neither men nor the technologies are treated as a homogeneous group.

This approach explores the effects of gender relations on invention and innovation, the design and development processes of technologies, as a counter to the previous stress on the impact of technology on women. It developed in response to the social studies of technology (see *Technological Determinism*; *Actor Network Theory*; *Technology, Social Construction of*). Feminists have demonstrated that the almost complete exclusion of women from the technological community has had a profound influence on the design and technical content of artifacts (Wajcman 1991). Indeed, technological systems implicitly place men's experiences and men's investments at the center, without acknowledging their specificity. The corollary is the simultaneous denial of other realities, such as women's. This gender analysis of technology does not view technology as a patriarchal conspiracy. Rather, it shows how preferences for different technologies are shaped by a set of social arrangements that reflect men's power and resources in the wider society.

As few women actually design technology, we must look elsewhere for their imprint on it. The increasing recognition of the pivotal role of users in technological development has therefore been important for feminist technology studies. Although technologies are designed with specific purposes in mind, the originators of a technology cannot predict or control completely its final use. In practice, the user or consumer of an artifact interacts with it, and can challenge and renegotiate its meanings and uses. By widening the research lens from a focus on innovation to include the processes of manufacturing, marketing, and consumption, women immediately come into view as active participants in the making of technology.

A good illustration is provided by the diffusion of the telephone (Martin 1991). There was a generation-long mismatch between how customers used the telephone and how its inventors envisaged it being used. The telephone was designed by telegraph men who intended that it would be used in the same way as its parent technology, the telegraph, as a business tool.

Women's use of the telephone for sociability instead eventually persuaded a reluctant telephone industry to conceive of its networks as a residential service for personal conversation. There are numerous other examples of women's capacity to subvert the intended purposes of technology and turn it to their collective advantage.

Indeed, largely as a result of actor network theory (see *Actor Network Theory*), technology is conceptualized as a fluid, relational process and the symbolic meanings attaching to it as being negotiated and reinvented continually. Contemporary feminist theory likewise stresses that gender is constructed in social interaction. Early studies of gender and technology tended to theorize gender as a fixed and unitary phenomenon, existing prior to and independently of technology, and then becoming embedded within it. Rather than construing technology and gender relations as separate spheres influencing each other, it is now understood that technologies and gender relations are mutually constitutive.

5. Technology as Culture

During the 1990s there was an explosion of feminist writing on technology, much of it being carried out at some conceptual distance from the sociology of science and technology which developed during the same period. This writing is informed explicitly by a combination of cultural studies, anthropology, and postmodern philosophy. Reflecting the postmodern emphasis on discourse, technology is conceptualized as an object of consumption, as a text, and as a communication medium. The central concern is with how technology as culture is implicated in the construction of subjective gender identities.

The most influential feminist commentator writing in this vein is Haraway (1997), who argues that we should embrace the positive potential of technoscience. Sharply critical of those who reject technology, she prefers to be a 'cyborg'—a hybrid of organism and machine parts—rather than an ecofeminist 'goddess' (see *Cyborg*). Genetic engineering, reproductive technology, and the advent of virtual reality are all seen as affecting fundamentally the basic categories of 'self' and 'gender.' Haraway and her ilk positively revel in the very difficulty of predicting what technology's effects will be, and warn against any purist rejection of the 'unnatural,' hybrid entities produced by biotechnology.

Most of the recent literature is about biomedical technologies and information technologies. The increasing preoccupation in sociological theory with the body and sexuality has been paralleled by research on biomedical technologies—technologies for the body. There are numerous studies of childbirth and contraception, in-vitro fertilization, cosmetic surgery and

genetic engineering. Since the late 1980s, feminists have begun to argue that there is now no such thing as a natural, physiological body. One consequence of this work is that the conventional distinction between sex (natural) and gender (social) has been contested and deconstructed thoroughly. Technologies, like science, are now seen as contributing to the stabilization of meanings of the body. With the rise of modern science, bodies have become objects that can be transformed with an increasing number of tools and techniques. Modern bodies are made and remade through science and technology; they too are technological artifacts. Life itself (human, plant, and animal) has been biomedicalized (Clarke 1988).

These studies illustrate the connection between the gendered discourses of the biomedical sciences and the institutionalization of medical techniques applied to women's bodies. They remind us that the conceptualization of male and female bodies as being essentially different, rather than similar, is a modern one, dating only from the eighteenth century. Women were defined in terms of sex and reproduction, and this was reflected in the establishment of gynecology as a separate branch of medicine. With the emergence of sex endocrinology in the 1920s and 1930s, the essence of femininity came to be located in hormones (Oudshoorn 1994). The development of a contraceptive pill for women, rather than for men, is the product not of biological necessity but of twentieth-century biomedical practices.

In the same way as biomedical technologies are seen as transforming the relations between the self, the body, and machines, so too are information and communication technologies. There is a profusion of studies on the Internet, cyberspace, and virtual reality. Many authors celebrate the myriad ways in which modern technology is challenging traditional notions of gender identity and creating new meanings and new entities (Turkle 1995). They argue that gender swapping, or virtual cross-dressing, on the Internet encourages people to reflect on the social construction of gender, to acquire a sense of gender as a continuum. The idea that virtual people or selves can exist in cyberspace, with no necessary link to a physical body, is sure to be taken up in future research.

New communication technologies have certainly opened up novel possibilities for sociality in cyberspace. Excitement has been generated about the capacity they offer to create new feminist communities, and the new tools they offer for global information exchange and political organizing. The significance of this technological revolution for power relations between men and women is still very much an open question.

As a result of the sociological and feminist research carried out in the last decades of the twentieth century, there is now a much more complex understanding of gender, of technology, and of the symbiotic relationship between them. Increasingly, it is accepted that

masculinity, femininity, and technology are not fixed, unitary categories but that they contain multiple possibilities and are constructed in relation to each other. At the same time, there has been a rejection of the technophobia evident in earlier feminist writing in favor of a popular cyberfeminism that embraces new technology as a source of empowerment for women. Albeit reflecting diverse perspectives, the study of gender and technology is firmly established as an integral part of science and technology studies.

See also: Feminist Epistemology; Feminist Movements; Feminist Theory; Gender and Reproductive Health; Reproductive Rights in Affluent Nations; Science and Technology, Anthropology of; Women's Health

Bibliography

- Braverman H 1975 *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century*. Monthly Review Press, New York
- Clarke A E 1988 *Disciplining Reproduction: Modernity, American Life Sciences, and 'the Problems of Sex'*. University of California Press, Berkeley, CA
- Cockburn C 1983 *Brothers: Male Dominance and Technological Change*. Pluto Press, London
- Corea G, Klein R D, Hanmer J, Holmes H B, Hoskins B, Kishwar M, Raymond J, Rowland R, Steinbacher R 1985 *Man-Made Women: How New Reproductive Technologies Affect Women*. Hutchinson, London
- Cowan R S 1983 *More Work for Mother: The Ironies of Household Technology from the Open Hearth to the Microwave*. Basic Books, New York
- Haraway D 1997 *Modest Witness@Second Millennium*. Routledge, Ithaca, NY
- Harding S 1986 *The Science Question in Feminism*. Cornell University Press, New York
- Hartmann H, Kraut R, Tilly L (eds.) 1986, 1987 *Computer Chips and Paper Clips: Technology and Women's Employment*. National Academy Press, Washington, DC, Vols. 1 and 2
- Martin M 1991 *'Hello, Central?' Gender, Technology and Culture in the Formation of Telephone Systems*. McGill-Queens University Press, Montreal, PQ
- Merchant C 1980 *The Death of Nature: Women, Ecology and the Scientific Revolution*. Harper & Row, San Francisco
- Mies M, Shiva V 1993 *Ecofeminism*. Zed Books, London
- Mitter S, Rowbotham S (eds.) 1995 *Women Encounter Technology: Changing Patterns of Employment in the Third World*. Routledge, London
- Oudshoorn N 1994 *Beyond the Natural Body. An Archaeology of Sex Hormones*. Routledge, London
- Stanley A 1995 *Mothers and Daughters of Invention: Notes for a Revised History of Technology*. Rutgers University Press, New Brunswick, NJ
- Turkle S 1995 *Life on the Screen: Identity in the Age of the Internet*. Simon & Schuster, New York
- Wajcman J 1991 *Feminism Confronts Technology*. Pennsylvania State Press University, Philadelphia, PA

J. Wajcman

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

5979

Gender and the Law

In law, there are two different kinds of gender-related rules: those with gender-specific reference, and those with gender-specific meaning despite gender-neutral wording. These norms and their application in practice reinforce the social construction of gender. The norms depend, of course, on the historic and socio-political context in which they originated, and are subject to change. The development and extension of the equality principle as a human right has made gender inequality in law less and less justifiable, and gender discrimination has been reduced at least in the law on the books. Yet it was left to feminist theorizing to discover gender as a structural principle of modern legal orders in the Western world. This article deals with the questions why gender is used in legal norms, and how and why gender is constructed by the law.

1. Gender and Genealogy

All recorded legal orders contain at least some norms referring to the sex or gender of legal subjects, namely in the context of procreation where both women and men are unavoidably involved. Not everyone produces children, but everyone has a biological mother and a biological father, whether they are known or not. Yet the link between a child and its parents is established not only by biology, but also by law. In law, motherhood is in most cases linked to the mere fact of giving birth. The question of fatherhood is less easily settled, because the biological father is necessary for conception, but not for childbearing and giving birth. Therefore any sociopolitical order to which the paternal line is important has rules for establishing fatherhood legally.

The establishment of a genealogy is not necessarily a purpose in itself. However, in class societies and their respective political orders, two questions are of high importance: Who belongs to the ruling class? And who receives the property and other titles if an owner dies intestate? To avoid having to answer these questions by a legally binding act of will in every single case, a genealogy bestowing birth rights simplifies matters. Rules about linking the newborn to its elders then usually include rules about the distinction between rulers and subjects, locals and strangers, free persons and those who are not.

2. Marriage Law and Regulation of Sexuality

The legal mechanism of establishing fatherhood may be an act of will like recognition or adoption, but the most widespread mechanism is through marriage: all children born during an existing marriage are automatically considered to be their mother's husband's

children, regardless of whether this corresponds to the biological facts. Marrying then, is an anticipating legal act, but it has preconditions. Many rules exist that regulate the question of who may marry whom, or whether he or she may marry at all. And the implications of marriage make marriage a complex legal institution which regulates such issues as the rights and/or duties of the spouses (see *Family Law*). All these rules vary widely over time and space because marriage intersects with religion (see *Hindu Law*; *Islam and Gender*; *Jewish Law*; *Judaism and Gender*), class, race (see *Race and Gender Intersections*) and state interests in population control (encouraging or discouraging procreation). Moreover, many a marriage law treats men and women differently.

Therefore, marriage law is not only the central regulatory site for the establishment of a patrilinear genealogy, but also for the construction of gender. At the end of the Middle Ages, the only type of marriage acknowledged in Europe was, due to the influence of the Catholic Church, the monogamous and principally dissoluble marriage, and it was the only accepted way to procreate, i.e., to produce legitimate heirs. Whereas previously so-called bastards were accepted by noble families when the father was an aristocrat, now children born out of wedlock were always considered 'illegitimate' and were left with their mother's status. This was safeguarded by a criminal law penalizing all forms of sexual conduct save conjugal intercourse (see *Crime and Gender*; *Sexuality and Gender*; *Sexual Orientation and the Law*). Of course, actual behavior differs from legal norms, but nevertheless the impact of such laws worked in two directions: one was the state intervention in the sexuality of its subjects, the other was the gendered discrimination of those who did not comply. When an unmarried woman became pregnant, she could less easily evade prosecution than the man, whose begetting was harder to prove. At the same time, the gendered dependency for procreation was reversed: if a child was born out of wedlock, it was no longer the man who depended on the woman to confirm his fatherhood, but now the woman depended on the man to pass on the status of legitimacy to her children.

3. The Masculinization of Citizenship

The marriage laws had their importance for the reproduction of the medieval classes, and for the construction of gender, but gender itself did not yet function as a predominant structural principle for inclusion in the ruling classes (see *Class and Law*). The rulers in medieval Europe were mostly, but not exclusively, men. The laws of the aristocracy (see *Aristocracy/Nobility/Gentry, History of*) usually provided for heredity in the male line, but not all of them excluded women from ruling if there was no legitimate male heir. The rural population came under the

jurisdiction of the respective land owners, whereas the cities generally did not belong to the nobility, but were self-governed by their free citizens. Free citizens were persons who were listed as such, which usually required them to swear a civic oath and to pay civic taxes. Women, whether married or not, were also eligible for full citizenship in medieval European cities (Appelt 1999). This started to change when the feudal orders of the Middle Ages and absolutism were called into question.

Generally, the American and French revolutions are seen as the beginning of the formation of the civil state in the Western world, where estates and classes are abolished. These revolutions had been influenced by the writings of political philosophers of the seventeenth and eighteenth centuries, who had conceptualized a Natural Law in opposition to the Catholic order of the Middle Ages. The Natural Law conceives all human beings as God's creatures and thus became the basis for the equality principle (see *Natural Law*). However, when this concept was made the foundation of the modern state, only men who were free owners were judged capable of being equal citizens; slaves, serfs, Jews and women were, in the beginning, explicitly excluded from the citizenry. The contradiction between the equality of all human beings in Natural Law and the exclusivity of the class of citizens was circumvented by means of the legal mechanism of qualification for legal personality (see *Legal Personality*) which required, according to the same political philosophers, an equal footing in contracting, of which neither slaves and servants nor women were capable. (The exclusion of Jews followed another mechanism linked to the emerging concept of nation.)

It is well known that French women had a major part in the revolution of 1789. Female suffrage was even discussed, but was explicitly denied in 1793. The medieval concept of citizenship was gradually converted into a concept of gendered citizenship which contained also a new concept of gender and the family. Citizens were perceived as males, equals in the public sphere, participating in public affairs by exercising their right to vote and to be elected, and as heads of family in the private sphere. Women were perceived as inherently unequal to men, their 'natural' tasks being to please the men, marry, and care for the 'legitimate' children (Rousseau 1762). As wives they were subordinated to their husbands either by 'nature' directly (Rousseau 1762), or through the marriage contract by which the woman subordinated herself (Kant 1797). The exclusion of women from politics was followed by an exclusion of women from the market during the nineteenth century, the phase of the consolidation of Western nation-states. When finally Jews, slaves, and other nonowners had been enfranchised, women belonging to these groups of the population stayed excluded like the mothers, wives, and daughters of the 'old' citizens. The marital status was the only status that was left after the abolition of statuses connected

with the estates, classes, and birth. However, it had a legal meaning only for women, leaving them in some sort of inferior class, with the result that the social and state order was run exclusively along the line of gender.

4. *The Gendered Structure of Modern Laws*

The conception of the citizen as a member of the new ruling class was mirrored not only in the constitutions describing the political person who participates in ruling, but also in the civil laws describing the civil person who participates in the market. Therefore it is no accident that the nineteenth century was, in the Civil Law world (see *Legal Systems, Classification of*), also the period of 'great codifications' (see *Civil Law*). These included norms on the legal personality, on property, contract, tort, and on marriage, family, and succession. Hegel (1821) presents a concise normative foundation for the (bourgeois) civil perception of the relation between the individual, the family, society, and the state.

Marriage and family were made the ideals for both men and women, but with diametrical consequences for their legal capacity. Marriage meant for women that they fell back to the status of minor children even in countries where relief from paternal custody was granted to women of age. Wives lost their property (if they had any) to their husbands (like in England), or their right to dispose of it (like in German states). They gained in turn a claim against their husbands for maintenance, but were obliged to have sexual intercourse whenever their husbands wanted it, as can be seen from the longstanding exemption of marital rape from prosecution. The women's obligation to conceive and give birth to children was reinforced by the abortion laws. Child custody was legally given to the husband, and remained there even after divorce, if a divorce was legally attainable at all. Unmarried and widowed women were likewise judged incapable of exercising child custody, and a male custodian was appointed. (For the status of 'illegitimate' children which hardly changed between the sixteenth century and the end of the twentieth, see *Children and the Law*).

Married women were made economically dependent on their husbands not only by losing their right to dispose of the property they had brought into the marriage, but also because they needed their husband's consent for taking up paid work. However, unmarried women were not free to choose their occupation either, because of the complete exclusion of women from office holding, and an increasing exclusion from crafts and trade; they were dependent on maintenance granted by kin or the community, or on earning a living by working in agriculture, households, or the newly emerging factories—all of which was poorly paid.

This nineteenth-century segregation of the labor market along the line of gender, partly induced and

partly reinforced by legal norms, still exists at the beginning of the twenty-first century. The resulting gender inequality is extended when and where the entitlement to social security is subject to previous paid work. For further treatment of these issues, see *Sex Segregation at Work*; *Sex Differences in Pay*; *Sexual Harassment: Legal Perspectives*; *Social Welfare Policies and Gender*.

Taken all together, the laws of the nineteenth century can be characterized as founding and substantiating civil patriarchy.

5. Law as a Means to Fight Gender Inequality

That women were deprived of equal rights has been under criticism from the beginning of the modern constitutional nation-state, starting with the *déclaration des droits de la femme et de la citoyenne* ('declaration of the woman and citizeness') by de Gouges (1791) and the *Vindication of the Rights of Women* by Wollstonecraft (1792). The women's movements that emerged in all Western states around the middle of the nineteenth century called for political participation and civil rights, especially women's suffrage and liberation of the marital subordination of married women (see *Feminist Movements*). It has been a long uphill road to achieve most of these goals in the Western world, at least in black-letter law. The tool that was and is being used in this fight is the invocation of equal rights. So the law, in spite of its function as a means to secure heterosexual male supremacy, could also be used to fight gender inequality (see *Law as an Instrument of Social Change*).

In 1948, the United Nations included gender equality in their Universal Declaration of Human Rights, and many national constitutions followed suit (see *Civil Liberties and Human Rights*; *Civil Rights*; *Fundamental Rights and Constitutional Guarantees*). Since then, it has been much harder to legitimize discrimination against women. In most countries, at least in the Western world, much has been achieved in the name of equality and equal rights to fight women's subordination (and where not, it is still on the agenda); suffrage (see *Women's Suffrage*); access to education and professions; the right to dispose of marital property; the right of married women to work; child custody for unmarried and divorced women (see *Lone Mothers in Affluent Nations*; *Lone Mothers in Nations of the South*); access to social security; reproductive rights (see *Reproductive Rights in Affluent Nations*). Affirmative action laws in the United States (see *Affirmative Action: Comparative Policies and Controversies*) and equal opportunities legislation and adjudication in the European Union (see Hervey and O'Keeffe 1996) have their share in the promotion of gender equality. Women now have, in many countries, access to all kinds of professions and jobs, including the formerly typically male military service.

Gender inequality in law was not only related to the different treatment of men and women, but also to sexuality. In the traditional patriarchal concept of the civil state, heterosexuality was the only accepted form of sexuality until, during the last third of the twentieth century, the gay and lesbian movements in the Western world began to invoke the equality principle on their own behalf (see *Gay/Lesbian Movements*; *Sexual Orientation and the Law*). The decriminalization of homosexuality is still an ongoing process, but sexual orientation is gradually being acknowledged as a human right—e.g., in article 21 of the Charter of Fundamental Rights of the European Union (declared at the Nice European Council on December 7, 2000).

Another demand made by the gay and lesbian movements in the name of equality is directed towards access to marriage. The reason is that many rights and entitlements (e.g., in tax laws, inheritance laws, immigration laws) are restricted to married couples or spouses. Barring homosexuals from marriage according to marriage law means denying them access to these rights and entitlements. This demand has been taken up by the legislators in a number of European countries (first in Denmark in 1989) and some US States, yet without extending the term 'marriage' to such relationships, using a phrase like 'registered partnership' instead (see ILGA-Europe 1998). Only the Dutch laws allow for a 'real' same-sex marriage as of April 1, 2001.

Nevertheless the effect of the laws is often limited (see, e.g., *Affirmative Action: Empirical Work on its Effectiveness*), and in no case guaranteed over time. Therefore the struggle against gendered law is still going on, both in countries with no or little achievement and in countries with comparatively high achievement.

6. Resistance to Gender Equality

The fight for women's equal rights has always met with considerable resistance. The resistance has three reasons.

The first is that the exclusion of women from the public sphere, including business and the market, and the ideology of marriage and family as the 'natural' profession of women led to a Catch-22 situation: women were subordinated to their husbands if they married, and ridiculed as spinsters if they stayed unmarried, with limited access to the better-paid positions in the labor market. And, most importantly, women had, at the beginning of the twentieth century, no access whatsoever to the lawmaking process—be it parliament and government, or the judiciary—because they were excluded from higher education and the professions irrespective of their marital status (see *Education (Higher) and Gender*; *Legal Education*). Although universities and law schools hesitantly began to open their doors for women from the 1860s

onwards, it was only in the 1970s and 1980s that women's share in the legal professions went beyond tokenism.

The second reason is that modern Western law is conceptualized in terms of individual rights. This is the case also for marriage and family law, although a married couple necessarily consists of two persons. The underlying image of a male head of household and breadwinner, and a female housewife and mother, proved resistant to the mere extension of rights to women—such as suffrage, and access to education and office. It was only in the last third of the twentieth century, and after the emergence of new feminist movements, that the laws were gradually adapted to the image of equal partners, even if the gendered division of tasks, namely breadwinning and family-caring, still resonates.

The third reason is that the patriarchal image of marriage and family, and a corresponding image of professions acceptable for women, had infected all other areas of law. However, the gendered character of modern law went unrecognized by mainstream legal theory until feminist lawyers and political scientists (mostly women, although the recognition of the gendered character of the law is not restricted to females) began to scrutinize the contradictions between equality as a human right, the laws reflecting the civil patriarchy, and women's lived experiences. They uncovered 'The Hidden Gender of Law' (Graycar and Morgan 1990), i.e., many contradictory and/or gender-biased traits of the law (see, e.g., Smart 1984, MacKinnon 1987, Stang Dahl 1987, Patemann 1988, Moller Okin 1989, Young 1990, Fineman 1995, Levit 1998; for further treatment see *Feminist Legal Theory*; *Feminist Theory*).

7. Towards the End of Gender as a Legal Category? An Outlook

One aspect of gender as a legal category still remains to be dealt with: the gender dichotomy which is not only taken up, but also created by law. This has been proven for ancient Roman law (see Thomas 1992). Hermaphrodites were even explicitly dealt with by the colified Prussian law of 1794. What is called 'the third sex' vanished, however, from the Western laws in the nineteenth and twentieth centuries when the registration of nationals and residents became increasingly regulated, and the registers attributed either the male or female sex. This official sex dimorphism made the attributed sex seem essential, and also gender became a category with seemingly two mutually exclusive expressions, fitting perfectly the gendered state and law which assigned rights and duties in a gender-specific way.

Sex was usually assigned for life until medical treatment and surgery made it possible to change the originally assigned sex/gender, and a change of sex

was finally legally acknowledged (see *Transsexuality, Transvestism, and Transgender*). Yet the permission for a change of sex does not mean that the cultural dimorphism of sex and gender has been overcome. The same medical treatment is used to uphold this dimorphism in the case of hermaphrodites (intersexed, intersexuals) who, in Western countries, already as infants are treated for becoming unambiguous females or males (see Kessler 1998). It seems hardly questionable that the legally secured sex dimorphism and gender dichotomy has contributed to this branch of surgery. Up till 2000, only the Constitutional Court of Columbia (SU-337/97 of May 12, 1999, and T-551/99 of August 2, 1999) recognized the problem of genital surgery imposed on infants as a legal problem conflicting with the personal integrity and dignity of those concerned.

Altogether sex, and consequently gender, as a legal category is, at the beginning of the twenty-first century, called into question from four different, but partly related, angles:

(a) separation of 'legitimate' procreation from marriage;

(b) separation of marriage (or a functional equivalent) from heterosexuality;

(c) extension of political, civil, and economic rights to all individuals, irrespective of their sex, sexuality, and gender; and

(d) the (re)discovery that nature offers a greater variability than the Western dimorphic conception of sex and gender of the nineteenth and twentieth centuries.

Under such conditions, it would be superfluous to register sex/gender or refer to sex/gender in any legal provision. Legal regulation could refer to a given situation instead of generalized males or females. Even in the context of procreation (to come back to the starting point), the distinction along the sex/gender line seems disposable when motherhood and fatherhood are replaced by parenthood. The function of law for linking a newborn to someone responsible will remain, but the factual situation to which legal parenthood is knitted could be actual care instead of biological generation (which has become problematic in the light of the recent reproductive technologies; see Knibiehler 1987). Parenthood could then be assigned to one, two, or even more persons, whether the biological parents be among them or not. Likewise, laws protecting motherhood, which have a long tradition in Europe, could be linked to the actual occurrence, and, where only aimed at female reproductive capacity, be extended to everyone.

However, as long as a gendered structure can still be found in the laws, complete disregard of gender might be counterproductive to those who still have unequal access to power and the exercise of rights because the demand for equal rights, the tried and tested tool to fight discrimination by law, requires the reference to differences.

See also: ; Aboriginal Rights; Beauvoir, Simone de (1908–86); Bernard, Jessie (1903–96); Boserup, Ester (1910–99); Civil Rights Movement, The; Conflict Sociology; Critical Race Theory; Economic Globalization and Gender; Ethnic and Racial Social Movements; Ethnicity, Sociology of; Family and Gender; Family Bargaining; Family Theory: Feminist–Economist Critique; Fatherhood; Female Genital Mutilation; Feminist Economics; Feminist Epistemology; Feminist Ethics; Feminist Legal Theory; Feminist Political Ecology; Feminist Theology; Feminist Theory: Psychoanalytic; Gay, Lesbian, and Bisexual Youth; Gender and Feminist Studies in Anthropology; Gender and Feminist Studies in Economics; Gender and Feminist Studies in Geography; Gender and Feminist Studies in History; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Gender and Place; Gender, Economics of; Gender, Feminism, and Sexuality in Archaeological Studies; Gender History; Historiography and Historical Thought: Southeast Asia; Household Production; Identity Movements; Klein, Melanie (1882–1960); Labor Movements, History of; Male Dominance; Mead, Margaret (1901–78); Mortality of the Oldest Old; Motherhood: Economic Aspects; Multicultural Feminism: Cultural Concerns; Multiculturalism: Sociological Aspects; Peace Movements, History of; Plagues and Diseases in History; Queer Theory; Race Identity; Racism, Sociology of; Rape and Sexual Coercion; Rationality and Feminist Thought; Regulation: Sexual Behavior; Reproductive Rights in Affluent Nations; Sex Differences in Pay; Sex Segregation at Work; Social Movements, Sociology of; Youth Movements, History of

Bibliography

- Appelt E 1999 *Geschlecht—Staatsbürgerschaft—Nation: Politische Konstruktionen des Geschlechterverhältnisses in Europa* [Gender—Citizenship—Nation: Political Constructions of Gender Relations in Europe]. Campus, Frankfurt-am-Main, Germany
- Fineman M A 1995 *The Neutered Mother, the Sexual Family and other Twentieth Century Tragedies*. Routledge, New York
- Gouges O de 1791 *Les droits de la femme*. Paris [1989 *The Rights of Woman*, trans. V Stevenson, Pythia, London]
- Graycar R, Morgan J 1990 *The Hidden Gender of Law*. Federation Press, Annandale, Australia
- Hegel G W F 1821 *Grundlinien der Philosophie des Rechts oder Naturrecht und Staatswissenschaften im Grundrisse*. Nicolaische Buchhandlung, Berlin [1991 *Elements of the Philosophy of Right*. A W Wood (ed.) H B Nisbet (trans.), Cambridge University Press, Cambridge, UK]
- Hervey T K, O'Keeffe D (eds.) 1996 *Sex Equality Law in the European Union*. Wiley, Chichester, UK
- ILGA-Europe (European Region of the International Lesbian and Gay Association) (ed.) 1998 *Equality for Lesbians and*

- Gay Men: A Relevant Issue in the Civil and Social Dialogue*. ILGA-Europe, Brussels, Belgium
- Kant I 1797 *Metaphysik der Sitten*. Nicolouius, Königsberg, Germany [1991 *The Metaphysics of Morals*, trans. M. Gregor, Cambridge University Press, Cambridge, UK]
- Kessler S J 1998 *Lessons from the Intersexed*. Rutgers University Press, New Brunswick, NJ
- Knibiehler Y 1987 *Les pères aussi ont une histoire*. Hachette, Paris
- Levit N 1998 *The Gender Line: Men, Women, and the Law*. New York University Press, New York
- MacKinnon C A 1987 *Feminism Unmodified: Discourses on Life and Law*. Harvard University Press, Cambridge, MA
- Moller Okin S 1989 *Justice, Gender, and the Family*. Basic Books, New York
- Pateman C 1988 *The Sexual Contract*. Stanford University Press, Stanford, CA
- Rousseau J-J 1762 *Émile ou de l'Éducation*. Néaulme, La Haye, France [1979 *Emile: or On Education*, trans. A Bloom, Basic Books, New York]
- Smart C 1984 *The Ties that Bind: Law, Marriage, and the Reproduction of Patriarchal Relations*. Routledge and Kegan Paul, London
- Stang Dahl T 1987 *Women's Law: An Introduction to Feminist Jurisprudence*. Norwegian University Press (Universitetsforlaget AS), Oslo, Norway
- Thomas Y 1992 The division of the sexes in Roman law. In: Schmitt Pantel P (ed.) *From Ancient Goddesses to Christian Saints: A History of Women in the West*, Belknap Press of Harvard University Press, Cambridge, MA, pp. 83–138
- Wollstonecraft M 1792 *A Vindication of the Rights of Women: with Strictures on Political and Moral Subjects*. J. Johnson, London
- Young I M 1990 *Justice and the Politics of Difference*. Princeton University Press, Princeton, NJ

K. Plett

Gender, Class, Race, and Ethnicity, Social Construction of

1. Introduction

The categories gender, class, race, and ethnicity, widely used in the social sciences, history, and literature, have long been treated as distinct categories, which supposedly describe real differences among groups of people. In this article this assumption is debated and various theories about the usefulness and the interaction and interrelationship of these categories are examined.

2. Background

Since the 1970s under the impact of postmodernism, deconstruction, feminist scholarship, and challenges to it by women of color, fundamental categories of

knowledge and systems of explanation have been disputed, problematized and, in some cases overthrown. Rigidly defined categories that had long been accepted as based on 'natural,' self-evident differences, and as being disparate from one another, were now subject to critique and redefinition. This process of re-evaluation took place over several decades and in many different fields. It will be discussed here category by category, although there was considerable overlap and crossover from one category to the other. Each of the categories represents some noticeable distinction from an imaginary norm. Each also is a powerful mental construct, which can be positively or negatively evaluated, according to who it is that does the defining.

3. Challenging the Categories

3.1 Race

The term 'race' first appeared in the sixteenth century; only in the nineteenth century did it become biologized on the basis of evolutionary thought. The 'races of mankind' were classified and hierarchically ordered by Europeans, with whites inevitably at the top of the ranking order. Modern science, however, gives no support to essentialist racial theories. There is no such thing as 'black blood' or 'pure blood' and the only distinctions of blood—bloodtypes—have nothing to do with race (see *Race: History of the Concept*). Biological race is, in fact, a scientifically nonexistent category. Of course, the physical characteristics (such as skin and hair color) that we associate with race are determined biologically. But which physical characteristics are seen to define race, and where lines are drawn between groups, as well as how groups are treated differently, is determined socially. For example, we note differences in eye coloring and attach no value judgments to them, but when it comes to differences in hair and skin coloring, all the contradictory and often negative value judgements accruing to race come into play.

Challenges to a naturalized concept of biological race began early in the twentieth century and found expression in the thought of African-American intellectuals. W. E. B. DuBois early in his life regarded race as a sociohistorical term, but still accepted the idea of 'a common blood.' In 1911 he wrote: 'It is not legitimate to argue from differences in physical characteristics to differences in mental characteristics' (DuBois 1911 but by 1940 he concluded: 'race is the fact that a Negro in Georgia has to ride in the Jim Crow car' DuBois 1940 p. 153.). In short, 'race' is nothing else but 'racism.'

The unscientific nature of the category race was recognized legally by the Supreme Court when it ruled miscegenation laws unconstitutional in 1967 (*Loving vs. Virginia*). Contemporary scholars, both black and

white, agree with historian Barbara Fields: '... the view that race is a biological fact, a physical attribute of individuals, is no longer tenable. Race ... is a purely ideological notion' (Fields 1982, p. 144; Henry Louis Gates and Kwame Anthony Appiah have expressed similar views). Many authors have searched for words which might more adequately describe the content of this complex concept, such as institutionalized racism, experiential racism, resistance formation. Literary critic Hazel Carby asked academics to stop using the term, but rather to speak of 'racial formations.' In Germany, in the aftermath of the Holocaust the term race is rarely used in academic discourse and has been replaced by the term 'ethnicity.'

However, race as a factor describing and determining people's life chances continues to exist. It is manifested in three forms: (a) as an ordering concept for dominance and discrimination, which has become institutionalized; (b) as a historic experience, a force shaping the lives and consciousness of the designated group; and (c) as a distinguishing marker for the group, transformed into a mark of pride, resistance, and a tool for liberation.

By exposing the category race as a social construct and showing its connection to power and to the unequal distribution of resources in society, scholars opened the way to a critique of the other categories. They also showed that the ability to define and categorize had developed into a powerful tool for social control. Considered from the point of view of philosophy, false universalizing (the situation where one category, such as men, is defined as the norm and a category seen as opposite, women, is defined as 'the Other' or the deviant) was found objectionable, and seen as a basic conceptual error. False universalizing was also implicated in explanatory systems pertaining to race, to sexuality, to ethnicity (Minnich 1990).

3.2 Sex/Gender

Simone de Beauvoir's *The Second Sex*, which appeared in the USA in 1953, explained woman's subordinate status in society as the result of her being defined as man's 'Other.' The author argued that 'one is not born, one is created a woman.' This insight led feminist scholars to new definitions, whereby Sex (the biological given) was contrasted with Gender (the socially constructed behavior, and the roles deemed appropriate to each of the sexes). She argued convincingly that the definition of woman as man's 'Other' had led to the systematic disadvantaging of women in all aspects of society, law, and culture. Gender would later be developed as a tool of analysis in many fields (Scott 1988; see *Gender and Feminist Studies*).

Comparisons between the way 'Negroes' and women were characterized, stereotyped, and discriminated against appeared in the sociological literature of the 1950s and became central to the arguments

of modern feminists (Hacker 1951). The dichotomy male–female was paralleled to white–black and other binary contrasts, like heterosexual–homosexual. Feminists criticized the binary system of thought and began to concern themselves with the parallel structures of ideas, prejudices, and social customs by which African–Americans and women were conceptualized as ‘deviants’ from a white, male norm. This led to other questions. How did the categories by which people were defined as deviants affect their lives? Definitions are mental constructions; discrimination in access to education, to resources, to political power is tangible and real. How did this work? Were the categories, gender, class, race, and ethnicity linked in some way and if so, how? Were they part of one or several systems? In the 1970s and 1980s the feminist debate postulated the existence of separate systems. Rubin (1975) described a single sex/gender system that linked economic oppression with gender. Marxist-feminists tried to fit a feminist analysis into Marxist categories, adding new dimensions and structures (Rowbotham 1973, Mitchell 1974, Davis 1981). O’Brien described a capitalist system of production and class oppression, which operated in conjunction with a separate system of reproduction and sex oppression. The Canadian sociologist Dorothy Smith defined gender and race in relational terms. She argued that relations of power manifest themselves in multiple, fluid structures of domination (Smith 1987).

3.3 Differences Reconsidered and Universals Challenged

In the early 1980s the concept of ‘sisterhood’ came under sharp attack by women of color, by post-colonialists and various ethnic groups. The emphasis of the debate now shifted to differences among women. Where previously race, class, and ethnicity had simply been treated as aspects of women’s identity, which might render their situation different from one another, these differences were now seen as so great that no valid generalizations about ‘women’ were possible. African–American women historians challenged the notion of a homogeneous womanhood and argued for the integral connection of race, gender, and class in the thought and lived experience of black women (see, e.g., Fields 1982, Higginbotham 1992). Patricia Collins argued that the experiences of African–American women were so unique as to qualify them singularly to interpret their past from an ‘Afro-centric standpoint’ (Collins 1989). ‘Standpoint theory’ originally developed by Nancy Hartsock, has also been used in discussions of ethnicity, class, and identity (Hartsock 1983). While it is helpful in drawing attention to the faults of false universalization, it not only affirmed but seemed to increase belief in the fixity of the assigned categories.

3.4 Binary Sexual Categories

The binary distinction between heterosexuals and homosexuals mirrors the power relations in other dichotomized categories. In her incisive essay Rich (1980) showed how the marginalization of gays and lesbians derived from their being defined as ‘deviant’ and essentially different from the heterosexual ‘norm.’ She traced the intellectual constructs, the legal constraints, and the everyday cultural practices that undergird the resulting system of discrimination against yet another minority group. The pattern of construction and ‘othering’ is remarkably similar to that in the other categories. In recent decades gay history and theory has flourished and led to the formation of another new field, Lesbian and Gay Studies (an introduction to the growing literature can be found in Abelow et al. 1993).

4. Class

Class is a term that sorts people according to their relation to resources and power. The term is a social construct, but it is based on actual economic differences between people. There is a large literature illustrating the way race and gender affect a person’s class position. Studies of a gendered labor market, of racial and gender discrimination in educational opportunities and in access to jobs, the continuing wage differential between men and women, and the income gap between black and white citizens argue for the significance of sex and race in defining class. Various ethnicities have, at different times, been subject to similar economic discrimination. Many labor historians now argue that the concept ‘class’ is gendered and racialized inherently (Moody and Kessler-Harris 1989). Class privileges and disadvantages vary by sex and race. For example, in nineteenth-century America the definition of male middle-class status required a man to be the sole family breadwinner. A ‘white lady’ by definition was someone sharply differentiated from lower-class women, from women defined as ‘not respectable,’ and from all nonwhite women. Another argument for the gendered nature of class is that it has been maintained historically by homogamy (the practice of marriage among persons of the same class) and by gendered inheritance laws that specified that sons were privileged over daughters in inheriting property. Both practices have given men a long-range economic advantage over women, by giving sons land and daughters dowries, when land was the most valuable property to have. Similarly, in modern times, gendered inheritance has given sons more family investment in higher education than daughters, when such education guaranteed upward mobility (Lerner 1997, pp. 146–98). There is a large literature concerning the nature of women’s unpaid housework and child rearing services and its effect on class. Despite considerable analytic

differences, there is agreement that their responsibility for unpaid housework generally has debased the wages of working women and their opportunities for economic advancement (Hartman 1981, Kessler-Harris 1990, Boydston 1990).

5. *Ethnicity*

This is an affiliation of people sorted by cultural and historical characteristics. These categories are obviously imprecise and arbitrary and inadequately deal with people of mixed ethnic origin. The ethnic designations in the US census have come under constant attack by people who consider themselves of 'mixed' ethnicity or race and by others who object to the definition of categories such as 'Hispanic' (Root 1996, Anzaldúa 1987, Zack 1993).

Some scholars have developed the view that ethnicity is a modern, constructed identity which is reinvented in each generation and reinforced by a variety of cultural means in the interest of nationalisms (Sollers 1989). An outgrowth of the debates on ethnicity and on race was the creation of a new racial designation and a new field of study, the creation of Whiteness. If race cannot be proven on a biological basis, neither can whiteness. White people are defined as the norm against the 'Otherness' of people of color. Whiteness bestows privileges on whites that are withheld from 'Others.' One of these privileges is that it is whites who do the defining of the social categories and who can remain unconscious of their own racial construction (Minnich 1990, Roediger 1991, Frankenberg 1993). The racialization of whiteness was used to argue both against racism and against affirmative action, by the claim that whites should be considered a group of ethnics who were disadvantaged and oppressed by race-specific legislation.

6. *The Relationship of the Categories to Each Other; The System as a Whole*

The relationship of the categories under discussion has often been described in geometric terms as 'intersection' (race and sex intersect to form class); it has been visualized as forces (sex acts upon class creating gendered class) or as a relationship approximating that of a mathematical curve. More and more, these ideas deriving from mechanical thought have given way to more holistic concepts, such as interacting, interpenetrating. The key to such an approach is to challenge the very existence of separate categories and identities and the recognition that people do not define themselves by a single identity, but by a number of identities which are interactive and mutually constitutive (Memmi 1968, Lorde 1984).

Postcolonial studies have illuminated this process and argued for its wider applicability. Stoler has shown

how colonial powers, in Indonesia, Indochina, and India, constructed a concept of the 'colonizer' vs. 'the colonized' in which race did not serve as the main marker of difference. Rules of conduct and custom enforced strict segregation between the two groups, reinforced by the myth that white women needed protection from assault by colonized males. Rules against miscegenation, abandonment of 'mixed race' children by their fathers, and strictures against white single women together with the encouragement of marriage among whites, encouraged the domesticity of white women and upheld colonial rule (Stoler 1992). Mohanty described 'the ideological construction and consolidation of white masculinity as normative and the corresponding racialization and sexualization of colonized people' in the case of Africa. Other studies have strengthened this argument (Mohanty 1991).

The connections between the categories appear ever more dense and complex. There has slowly developed agreement among many scholars that the terms must be treated not as separate entities, but as interrelated. Bunch (1990) argued that various forms of oppressive systems interact. Hooks (1989) stated that: 'Sexism, racism and class exploitation constitute interlocking systems of domination.' Sacks (1989) saw 'Class, race and gender oppression as part of a unitary system.' Morgen (1990) explained 'that these social relations so interpenetrate that they are better understood as constituting...one another.' Noticing the arbitrariness of the selection of groups that are targeted for definition as 'the Other,' a number of scholars have come to view the process of forming deviant outgroups by way of categorization as part of the system of ruling (e.g., Lorde 1984, Williams 1991, Minnich 1990, Lerner 1997). Gender, race, ethnicity, and class are processes through which hierarchical relations are created and maintained in such a way as to give some men and women power and privilege over others by their control of material resources, sexual and reproductive services, education, and knowledge. Such control is maintained by a complex weave of social relations among dependent groups which offers most groups some advantages over other groups, sufficient to keep each group within the dominance system.

The categories are not just interrelated, they are also mutually constitutive. Gender is constructed racially and through class and ethnicity. Race is constructed genderically. Thus, the gender ideal of 'womanhood' was long defined for white women as embodying purity and dependency on male protection, while women of color were seen as 'impure,' open to sexual access by white men, and economically self-supporting. The gender ideal turns out to be race-specific. By a similar mental contortion middle- and upper-class mothers are today urged to stay home with their preschool children, while mothers on welfare are forced to leave their infants to go to work. This ideal turns out to be class-specific. A similar case can be

made for all other social formations. A holistic redefinition of the social categories may lead to a more accurate description of society as a functioning organism.

7. Conclusion

Scholarship on diversity and multiculturalism has gone through a process which began with the traditional acceptance of rigid, separate categories into which groups of persons were divided according to noticeable or invented differences. Challenging the categories as neither natural nor separate, denaturalizing them, led to the insight that this ordering system was connected with power—the power by which difference is turned into dominance. Underlying the construction of categories of difference was a binary system of thought, which soon came under attack from many quarters. The search for alternate modes of thinking and for models more appropriate to diversity in democratic societies led to greater concern with inclusivity, and a new scholarly interest in the problems of borderlands, mixed identities, hybridity, and holistic models.

See also: Conflict Sociology; Critical Race Theory; Ethnic and Racial Social Movements; Ethnicity, Sociology of; Feminist Theory and Women of Color; Gender and Feminist Studies; Multicultural Feminism: Cultural Concerns; Multiculturalism: Sociological Aspects; Race and Gender Intersections; Race: History of the Concept; Race Identity; Racism, Sociology of; Social Class and Gender

Bibliography

- Abelove H, Barale M A, Halperin D (eds.) 1993 *The Lesbian and Gay Studies Reader*. Routledge & Kegan Paul, New York
- Anzaldúa G 1987 *Borderlands: La Frontera: The New Mestiza*. Spinsters/Aunt Lute, San Francisco
- Berger P L, Luckmann T 1966 *The Social Construction of Reality: A Treatise in The Sociology of Knowledge*. Anchor Books/Doubleday, New York
- Boydston J 1990 *Home and Work: Housework, Wages, and the Ideology of Labor in the early Republic*. Oxford University Press, New York
- Bunch C 1990 Making common cause: Diversity and coalitions. In: Albrecht L, Brewer R M (eds.) *Ridges of Power: Women's Multicultural Alliances*. New Society Publishers, Philadelphia, PA
- Collins P H 1989 Social construction of black feminist thought. *Signs* 14(4): 745–74
- DuBois W E B 1911 Races. *Crisis* Aug: 157–8
- DuBois W E B 1940 *Dusk of Dawn: An Essay Toward an Autobiography of a Race Concept*. Harcourt Brace, New York
- Davis A 1981 *Women, Race and Class*. Random House, New York
- Fields B 1982 Ideology and race in American history. In: Kinser J M, McPherson J M (eds.) *Region, Race and Reconstruction*. Oxford University Press, New York, pp. 143–77
- Frankenberg R 1993 *White Women, Race Matters: The Social Construction of Whiteness*. University of Minnesota Press, Minneapolis, MI
- Hacker H 1951 Women as a minority group. *Social Forces* 30: 60–9
- Hartman H 1981 The family as the locus of gender, class, and political struggle: The example of housework. *Signs* 6(1): 366–94
- Hartsock N 1983 The feminist standpoint: Developing the ground for a specifically feminist historical materialism. In: Harding S, Hintikka M B (eds.) *Discovering Reality: Feminist Perspectives on Epistemology, Metaphysics Methodology and Philosophy of Science*. Reidel, Boston
- Higginbotham E B 1992 African–American women's history and the metalanguage of race. *Signs* 17(2): 251–74
- Hooks b 1989 *Talking Back: Thinking Feminist, Thinking Black*. South End Press, Boston
- Kessler-Harris A 1990 *A Woman's Wage: Historical Meanings and Social Consequences*. University of Kentucky Press, Lexington, KY
- Lerner G 1997 Differences among women and rethinking the paradigm: Race; Class. In: Lerner G (ed.) *Why History Matters: Life and Thought*. Oxford University Press, New York
- Lorde A 1984 *Sister Outsider*. Crossing Press, Trumansburg, NY
- Memmi A 1968 *Dominated Man*. Orion Press, New York
- Minnich E K 1990 *Transforming Knowledge*. Temple University Press, Philadelphia, PA
- Mitchell J 1974 *Woman's Estate*. Pantheon, New York
- Mohanty C T 1991 Introduction: Cartographies of struggle, third world women and the politics of feminism. In: Mohanty C T, Russo A, Torres L (eds.) *Third World Women and the Politics of Feminism*. Indiana University Press, Bloomington, IN
- Moody J C, Kessler-Harris 1989 *Perspectives on American Labor History: The Problem of Synthesis*. Northern Illinois University Press, DeKalb, IL
- Morgen S 1990 Conceptualizing and changing consciousness. In: Hansen K, Philipson I (eds.) *Women, Class and the Feminist Imagination*. Temple University Press, Philadelphia, PA, p. 285
- Rich A 1980 Compulsory heterosexuality and lesbian existence. *Signs* 5: 657
- Rowbotham J 1973 *Woman's Consciousness, Man's World*. Penguin, Baltimore, MD
- Roediger D R 1991 *The Wages of Whiteness: Race and the Making of the American Working Class*. Verso, London
- Root M P P (ed.) 1996 *The Multiracial Experience: Racial Borders as the New Frontier*. Sage, Thousand Oaks, London, New Delhi
- Rubin G 1975 The traffic in women: Notes on the political economy of sex. In: Reiter R R (ed.) *Toward an Anthropology of Women*. Monthly Review Press, New York
- Sacks K B 1989 Toward a unified theory of class, race, and gender. *American Ethnologist* 16(3): 534–50
- Scott J W 1988 *Gender and the Politics of History*. Columbia University Press, New York
- Smith D 1987 *The Everyday World as Problematic: A Feminist Sociology*. Northeastern University Press, Boston
- Sollers W (ed.) 1989 *The Invention of Ethnicity*. Oxford University Press, New York

- Stoler A L 1992 Sexual affronts and racial frontiers: European communities and the cultural politics of exclusion in colonial Southeast Asia. *Comparative Studies in Society and History* 34(2): 514–51
- Williams P 1991 *The Alchemy of Race and Rights*. Harvard University Press, Cambridge, MA
- Winant H, Omi M 1986 *Racial Formation in the United States*. Routledge & Kegan Paul, New York
- Zack N 1993 *Race and Mixed Race*. Temple University Press, Philadelphia, PA

G. Lerner

Gender Differences in Personality and Social Behavior

Research on psychological gender differences began before the beginning of the twentieth century, as scientists tried to discern differences between women and men in intelligence, reaction time, and personality, using measures that seem bizarre by today's standards, such as handwriting analysis. Sophisticated research on gender differences in personality and social behavior has flourished since the 1970s, generating hundreds of empirical studies as well as a number of highly informative meta-analyses. A meta-analysis is a quantitative review of the existing studies on a particular question; the meta-analyst extracts statistics from the existing studies and averages the results across all studies. (When reporting effect sizes from meta-analyses here, I follow the convention that a positive effect size means that men scored higher on the measure in question and a negative effect size means that women scored higher. An effect size tells how far apart the two groups—in this case, men and women—are in their average scores, standardized by dividing by the standard deviation.)

1. Gender Differences in Personality

Definitions of the term 'personality' vary widely. In order to sharpen the focus of this review, I will consider personality as conceptualized in the Big Five model (McCrae and Costa 1987), which specifies five basic dimensions of personality: conscientiousness, extraversion, openness to experience, agreeableness, and neuroticism. Below I review research on gender differences in each of these dimensions. A meta-analysis of gender differences in personality (Feingold 1994) is particularly helpful in discerning the overall pattern of findings.

1.1 Conscientiousness

Feingold located seven norming studies measuring conscientiousness using standardized personality tests.

Of those, women scored higher on four, one found exactly no difference, and men scored higher on the remaining two. The unweighted mean effect size was $d = -.13$ (Feingold 1994). Therefore, the gender difference in conscientiousness appears quite inconsistently across studies and, if there is a gender difference in conscientiousness, it favors women but the difference is slight.

1.2 Extraversion

Most studies find no gender differences in extraversion as measured by the MMPI (Zonderman et al. 1993), or the Eysenck Personality Questionnaire (EPQ) (Martin and Kirkcaldy 1998).

In his meta-analysis, Feingold (1994) distinguished between two aspects of extraversion: gregariousness and assertiveness. Most test norming studies show women to be more gregarious, although the gender difference is small (unweighted mean $d = -.14$) and inconsistent across countries (Feingold 1994). Test norms show men scoring higher on assertiveness or dominance, ($d = .50$).

1.3 Openness to Experience

Openness to experience—flexibility and openness to new ideas—has been less studied. The mean effect size for gender differences was $d = .03$ (Feingold 1994), indicating no gender difference in openness to experience.

1.4 Agreeableness

Findings regarding gender differences in agreeableness are mixed. Feingold (1994) distinguished between two aspects of agreeableness: trust and tender-mindedness. Test norming data showed a consistent trend for women to score somewhat higher on trust ($d = -.25$) and for women to score considerably higher on tender-mindedness or nurturance ($d = -.97$). Higher scores for women on tender-mindedness or nurturance have been documented in Canada, Finland, Germany, Poland, and Russia, but not China (Feingold 1994).

1.5 Neuroticism

Neuroticism is the tendency to experience negative emotions such as sadness or anxiety, as well as mood swings (Costa and McCrae 1985). Those scoring high on neuroticism tend to worry or ruminate a great deal and are prone to having their feelings easily hurt. Women consistently display higher scores on neuroticism scales than men do. These differences are found on the MMPI Neuroticism scale (Zonderman et al. 1993), the Eysenck Personality Questionnaire (Martin and Kirkcaldy 1998), and the NEO (Costa

and McCrae 1985). The magnitude of the gender difference is generally around $d = -.25$ (Feingold 1994, Zonderman et al. 1993), that is, the difference is small. For comparison, the gender difference on the masculinity–femininity scale of the MMPI is approximately $d = 2.0$ (Zonderman et al. 1993).

The gender difference has been documented in a wide variety of countries including the USA, Singapore, and Bangladesh (Lynn and Martin 1997), and Northern Ireland (Martin and Kirkcaldy 1998).

The gender difference in neuroticism is a particular concern because of the association between neuroticism and depression and the well-established gender difference in depression (Roberts and Gotlib 1997). Watson and Clark (1984) have demonstrated that the personality trait labeled neuroticism is part of a broader cluster of traits including emotionality, trait anxiety, and ego strength; they label this broad personality construct ‘negative affectivity.’ Scales measuring traits in this cluster consistently show women scoring higher.

1.6 Self-esteem

Self-esteem, although not one of the Big Five dimensions, is an individual difference variable that has stimulated much research on gender differences. On the basis of his meta-analysis, Feingold (1994) concluded that males had slightly higher self-esteem than females, ($d = .16$).

Kling et al. (1999) conducted a more extensive meta-analysis of gender differences in self-esteem (for another meta-analysis with similar results, see Major 1999). Averaged over 216 effect sizes, ($d = .21$), indicating a small gender difference favoring males. Contrary to claims in the popular press that girls’ self-esteem plummets in early adolescence, creating a large gender difference, the largest gender difference appeared in the high school years and was small to moderate in magnitude, ($d = .33$). Results based on nationally representative data sets from the National Center for Education Statistics showed smaller gender differences in the high school years, with nearly all effect sizes less than .16. When the studies were analyzed according to the ethnicity of the sample, the results indicated that a small gender difference favoring males exists for Whites ($d = .20$) whereas there is no gender differences for Blacks ($d = -.04$).

In conclusion, the gender difference in self-esteem is small at most and, in the USA, is found in White samples but not in Black samples.

1.7 Summary

On the Big Five dimensions of personality, there appear to be no gender differences in openness to experience and only a slight gender difference favoring women in conscientiousness. Within the dimension of

extraversion, if one distinguishes between gregariousness and assertiveness or dominance, there is a small gender difference favoring women on the former and a moderate gender difference favoring men on the latter. For agreeableness, if one distinguishes between trust and tender-mindedness or nurturance, there is a small gender difference in the former and a large gender difference in the latter; in both cases, women score higher. A small gender difference in neuroticism, women scoring higher, appears consistently across studies and cross-nationally. Finally, the gender difference in self-esteem is small at most and, in the USA, is found in White samples but not in Black samples.

2. Gender Differences in Social Behavior

This section reviews research on gender differences in aggression, helping behavior and altruism, group behaviors, and communication and nonverbal behaviors.

2.1 Aggression

Two meta-analyses provide summaries of the existing data on gender differences in aggression. Hyde’s (1986) meta-analysis used a developmental approach in reviewing 143 studies. Averaged over all studies, the average effect size d was .50; that is, the magnitude of the gender difference was moderate. Naturalistic studies of aggression yielded larger gender differences ($d = .56$) than laboratory experimental studies ($d = .29$). Studies of physical aggression produced larger gender differences ($d = .60$) than studies of verbal aggression ($d = .43$). One should note, however, that both effect sizes are positive; that is, males were more aggressive for both physical and verbal aggression. Gender differences were larger for preschoolers ($d = .58$) than for studies of college students ($d = .27$).

Eagly and Steffen (1986) took a different approach in their meta-analysis, limiting their review to social psychological research on aggression in adults, mainly college students. They obtained a mean effect size of .29. The discrepancy between this overall effect size and the value of .50 reported by Hyde (1986) doubtless lies in the fact that Eagly and Steffen reviewed only studies of adults, among whom gender differences are smaller; for studies of college students, Hyde found an effect size of .27, nearly identical to the one obtained by Eagly and Steffen.

2.2 Helping Behavior and Altruism

Eagly and Crowley (1986) conducted a meta-analysis of studies of gender differences in helping behavior. They grounded their meta-analysis in social role theory and argued that it made no sense to examine an overall effect size for gender differences in helping

behavior. Instead, whether males help more or females help more depends on whether the situation calls for helping that is consistent with the male role or the female role. The results are therefore highly dependent on the situations in which social psychologists have measured helping. Men, for example, are more likely to help in situations in which they know they are being observed ($d = .74$) whereas there is no gender difference ($d = -0.02$) when participants believe they are not being observed. Heroism is part of the male role and heroism requires onlookers. Helping in the private context of long-term relationships is part of the female role, for example, a mother helping a child. This type of helping has been neglected in social psychological research.

Eisenberg and Fabes (1998) conducted a meta-analysis of studies of gender differences in children's prosocial behavior. The overall effect size was $d = -.18$, indicating a small gender difference favoring girls. Girls scored higher on all types of prosocial behavior, including instrumental help and comforting; the largest gender difference was for being kind or considerate, $d = -.42$.

2.3 Group Behaviors

2.3.1 Instrumental and expressive roles. Research in this area has been influenced heavily by the early research and theorizing of sociologist Talcott Parsons (1955), who argued that two roles develop in all small groups including the nuclear family, the instrumental role (task-oriented, getting things done), and the expressive role (providing emotional support to group members), and that women specialize in the expressive role and men in the instrumental role. Support for the notion of instrumental and expressive roles came from a study of 14 all-male groups (Bales and Slater 1955), although Parsons apparently did not see the inconsistency between that research and his arguments about gender-based specialization. A reanalysis of a classic study that is often cited as evidence for gender differences in instrumental and expressive behaviors found that, although men exhibited significantly more task-oriented behaviors than women did, 79 percent of men's behaviors were task-oriented and 63 percent of women's were; that is, for both women and men, the majority of behavior was instrumental (Aries 1996). Several decades later, then, it is clear that, although there is evidence of gender differences in instrumental and expressive behaviors in groups, gender similarities are perhaps as striking. In addition, many features of the task and the situation influence the results (Aries 1996).

2.3.2 Leadership. In mixed-sex groups, men are more likely than women to emerge as leaders, al-

though the magnitude of the difference is modest and the results are too complex to encompass in a single statement about gender differences. Eagly and Karau (1991) meta-analyzed studies of the emergence of leaders, most of which were laboratory studies done with college students. For task-oriented leadership, men emerged more frequently as leaders ($d = .41$). For general leadership measures, men also emerged as leaders more, ($d = .32$). Women, however, were more likely to emerge as leaders when social leadership was assessed, ($d = -.18$).

Research on leadership styles generally distinguishes between democratic and autocratic styles. Eagly and Johnson (1990) meta-analyzed research on gender differences in these styles. They found that women showed a more democratic style than men ($d = -.22$). To make matters more complex, the results were found in laboratory studies but not in field studies in actual organizations. The null findings in the real-world setting may have to do with the tendency for organizations to select leaders with particular styles.

2.4 Communication and Nonverbal Behavior

Two recent volumes have reviewed the existing research on gender differences in communication and nonverbal behavior (Aries 1996, Canary and Dindia 1998). Here I summarize the trends in findings for gender differences in language use, the content of conversations, the evaluation of speakers in relation to gender-stereotyped speech, interruptions, and nonverbal decoding.

2.4.1 Gender and language use. In 1973, Lakoff published an influential paper on gender differences in language use. In it, she argued that women use a number of speech forms—for example, tag questions, hedges, and polite forms—considerably more frequently than men, and that these forms reflect uncertainty and deference. Lakoff theorized, in turn, that this speech style represents women's adaptation to their lack of power in society. The style avoids conflict and appeases others precisely because it is not forceful. For example, 'That was a great game!' is a forceful statement, but when the tag questions is added, 'That was a great game, wasn't it?' it becomes considerably less forceful, according to Lakoff.

Lakoff's thesis contains two empirically testable hypotheses: (a) there are reliable gender differences in the use of tag questions, hedges, and polite forms, with women using more of all three; and (b) these forms have the function of signaling uncertainty and deference. I review the data for each hypothesis in turn.

In regard to tag questions, research findings are inconsistent. Some studies confirm Lakoff's belief that women use more tag questions than men (e.g.,

Carli 1990, Holmes 1984), some find that men use more tag questions (e.g., Cameron et al. 1988) some find no gender differences (Baumann 1976, cited in Aries 1996), and some find that the direction of the gender difference depends on the situation (Holmes 1984, Cameron et al. 1988, Kollock et al. 1985). In short, the evidence does not strongly support Lakoff's assertion about gender differences in the use of tag questions.

The second of Lakoff's hypotheses concerns the meaning and function of tag questions. Do they really signal uncertainty? Holmes's (1984) data indicate that tag questions can serve three functions: expression of uncertainty ('The plane is scheduled to depart at 9:35, isn't it?'), expressions of solidarity ('You're still limping a bit, aren't you?'), and politeness to soften an order or negative comment ('To get tenure, you'll have to improve your teaching, won't you?'). In her data, men actually used tags to express uncertainty more than women did, whereas women were more likely to use tags for the other two functions, which maintain and facilitate conversation. Therefore Lakoff's second hypothesis is not supported, but the data suggest that women's tag questions have different functions such as expressing solidarity and softening a difficult comment.

The pattern of findings for hedges is quite similar to that for tag questions (Aries 1996). A purported hedge such as 'I think' may be a hedge or a statement of confidence depending on several factors, including whether the emphasis is on 'think' or 'I.' For example, in the sentence 'Melbourne is the capital of New Zealand, I think,' the phrase serves as a hedge, indicating uncertainty. In contrast, 'I think Truman was the best president of the century,' with emphasis on 'I,' is a strong statement of confidence. Again, there is little support for Lakoff's assertions.

The theme that there are striking gender differences in language use was taken up again vigorously by Tannen in her books written for a popular audience, such as *You Just Don't Understand: Men and Women in Conversation* (1991). Tannen argued that men's and women's language use is so different that they essentially belong to different linguistic communities, and therefore that male-female communication is as difficult and error-prone as cross-cultural communication. Moreover, she argued that men and women use conversation for different purposes: men to express status and independence and women to establish connectedness and intimacy with the other person. She believes that these differences arise because boys and girls grow up in essentially different cultures because of the gender segregation of childhood play. Aries (1996), however, in a review of the research, found little evidence to support Tannen's propositions.

2.4.2 Self-disclosure. A large research literature exists on self-disclosure in conversations and, in par-

ticular, gender differences in self-disclosure. Dindia and Allen (1992) conducted a meta-analysis of these studies. Across 205 studies, women self-disclosed more, but the difference was small ($d = -.18$). Patterns of self-disclosure depend on the sex of the target. There is no gender difference in self-disclosure to a male target ($d = .00$), whereas the gender difference is more substantial for a female target ($d = -.35$). Patterns also depend on the relationship between the two people engaged in the conversation. The gender difference in self-disclosure is trivial when the conversation is between strangers ($d = -.07$) and is found mainly when the conversation is with a friend ($d = -.28$), a spouse ($d = -.22$), or a parent ($d = -.26$).

2.4.3 Evaluation of speakers. If college students are asked to prepare and deliver a speech and the speeches are transcribed so that they can be rated blind to the sex of the speaker, women's speeches tend to be rated higher on sociointellectual status and esthetic quality and men's speeches are rated higher on dynamism (e.g., Mulac and Lundell 1980). This has been termed the gender-linked language effect and has been found as early as fourth grade (Mulac et al. 1990).

A large body of research in social psychology has demonstrated that the stereotypes we hold influence the impression that we form of an individual. Research has shown repeatedly that people hold stereotypes about women's and men's speech (Aries 1996). For example, people expect men to interrupt more than women (Hawkins 1988). By sixth grade, children's stereotypes of male and female speech agree with adults' (Edelsky 1976).

The evidence indicates that gender stereotypes do influence raters' evaluations of men's and women's speech. In one study, participants listened to tape recordings of male and female speakers, who in fact used tag questions equally; participants perceived the women to have used more tag questions than the men (Newcombe and Arnkoff 1979). In short, people perceive women's and men's speech to confirm gender stereotypes even when the actual speech does not.

Are women and men evaluated differently, depending on whether they display male-stereotyped or female-stereotyped speech? Erickson et al. (1978) created two scripts, in the context of a courtroom trial setting; in one, the speaker used powerful or male-stereotyped speech and in the other, powerless or female-stereotyped speech. Raters were led to believe that the speaker was either male or female. The results indicated that, regardless of the speaker's sex, raters found those who used more powerful, male-stereotyped language to be more credible and more attractive. Research using such a paradigm has consistently found that, regardless of the sex of the speaker, powerless, female-stereotyped language is

viewed more negatively and powerful, male-stereotyped language more positively on most dimensions (Carli 1990).

2.4.4 Interruptions. Research in the 1970s found that, in mixed-sex conversations, men interrupted women considerably more than the reverse (Zimmerman and West 1975). This pattern was found both in natural conversations (Zimmerman and West 1975) and in controlled laboratory situations (West and Zimmerman 1983). These results stimulated much interest; they were widely interpreted as an indication of nonverbal dominance of females by males (Zimmerman and West 1975, Smith-Lovin and Brody 1989).

Recent reviews, however, have challenged both the existence of the gender difference and its interpretation (Aries 1996). First, a review of more recent research as well as a critical examination of the original studies indicates that there are no well-documented gender differences in interruptions. Second, understanding of interruptions has become more sophisticated and it is clear now that, although some interruptions are power-oriented (e.g., cutting off a speaker to change the topic), interruptions can serve many other functions, including expressing support, interest, or agreement.

2.4.5 Smiling. One nonverbal behavior that has consistently shown gender differences is smiling. In an early meta-analytic review, Hall (1984) found that females consistently smile more than males, ($r = .30$), which translates to $d = -.63$, a difference that is in the moderate-to-large range and is noticeably larger than most other gender differences reviewed here. More recent studies continue to document the same phenomenon (Hall 1998). The pattern, though, is more characteristic of White women than African-American women (Halberstadt and Saitta 1987).

Generally this difference is interpreted as indicating the extent to which smiling is a part of the female role. For example, in one study, participants were given a written description of a person, accompanied by a photograph of a man or woman who was smiling or not smiling (Deutsch et al. 1987). Women who were not smiling were given more negative evaluations; apparently people react negatively to this violation of the female role.

2.4.6 Nonverbal decoding. Nonverbal decoding (also termed nonverbal sensitivity) refers to a person's ability to read correctly the nonverbal behaviors or body language of another person. A gender difference favoring females has been documented over several decades. In an early meta-analysis, Rosen-

thal et al. (1979) located 133 samples that had been administered the Profile of Nonverbal Sensitivity (PONS, which bases scores on responses to short video clips); females scored higher than males in 80 percent of the samples ($d = -.41$), a moderate effect. Later meta-analyses based on a wider variety of measures obtained quite similar results; for example, Hall (1984) found $d = -.52$.

3. Future Directions

Research on gender differences in personality and social behavior must certainly move beyond questions of simple gender differences, to systematic analyses of situation $\times$ person and situation $\times$ gender interactions. Which situations call forth gender differences and which do not? How can these situational variations be drawn into a coherent pattern that is theoretically meaningful?

4. Summary

Over a wide range of personality and social variables, there is much evidence of gender similarities. For the Big Five personality dimensions, there is no gender difference in openness to experience ($d = .03$) and small differences in conscientiousness ($-.13$), extraversion or gregariousness ($-.14$), and neuroticism ($-.25$); there is, however, a large difference in one aspect of agreeableness, termed tendermindedness or nurturance ($-.97$). The gender difference in self-esteem is small ($.21$).

Likewise, there are many gender similarities in the realm of social behavior. The direction of the gender difference in helping behavior depends largely on the situation. In small groups, men and women are similar in their instrumental and expressive behaviors. The gender difference in democratic vs. autocratic leadership style is small ($-.22$ for democratic style). The research evidence does not support popular media claims about enormous gender differences in language use. The gender difference in self-disclosure is small ($-.18$). There are several exceptions to this pattern of gender similarities. Gender differences are moderately large for aggression ($.50$), smiling ($-.63$), and sensitivity to nonverbal cues ($-.52$).

See also: Gender and Feminist Studies in Psychology; Gender and Language: Cultural Concerns; Gender and School Learning: Mathematics and Science; Gender Identity Disorders; Gender-related Development; Language and Gender; Masculinities and Femininities; Small-group Interaction and Gender; Social Networks and Gender

Bibliography

- Aries E 1996 *Men and Women in Interaction*. Oxford University Press, New York
- Bales R F, Slater P E 1955 Role differentiation in small decision-making groups. In: Parsons T, Bales R F (eds.) *Family, Socialization, and Interaction Process*. Free Press, Glencoe, IL
- Cameron D, McAlinden F, O'Leary K 1988 Lakoff in context: The social and linguistic functions of tag questions. In: Coates J, Cameron D (eds.) *Women in Their Speech Communities: New Perspectives on Language and Sex*. Longman, New York
- Canary D J, Dindia K (eds.) 1998 *Sex Differences and Similarities in Communication*. Erlbaum, Mahwah, NJ
- Carli L L 1990 Gender, language, and influence. *Journal of Personality and Social Psychology* **59**: 941–51
- Costa P T, McCrae R R 1985 *The NEO personality inventory: Manual Form S and Form R*. Psychological Assessment Resources, Odessa, FL
- Deutsch F M, LeBaron D, Fryer M M 1987 What is in a smile? *Psychology of Women Quarterly* **11**: 341–51
- Dindia K, Allen M 1992 Sex differences in self-disclosure: A meta-analysis. *Psychological Bulletin* **112**: 106–24
- Eagly A H, Crowley M 1986 Gender and helping behavior: A meta-analytic review of the social psychological literature. *Psychological Bulletin* **100**: 283–308
- Eagly A H, Johnson B T 1990 Gender and leadership style: A meta-analysis. *Psychological Bulletin* **108**: 233–56
- Eagly A H, Karau S J 1991 Gender and the emergence of leaders: A meta-analysis. *Journal of Personality and Social Psychology* **60**: 685–710
- Eagly A H, Steffen V J 1986 Gender and aggressive behavior: A meta-analytic review of the social psychological literature. *Psychological Bulletin* **100**: 309–30
- Edelsky C 1976 The acquisition of communicative competence: Recognition of linguistic correlates of sex roles. *Merrill-Palmer Quarterly* **22**: 47–59
- Eisenberg N, Fabes R A 1998 Prosocial development. In: Damon W (ed.) *Handbook of Child Psychology*, 5th ed. J Wiley, New York, pp. 701–78
- Erickson B, Lind E A, Johnson B C, O'Barr W M 1978 Speech style and impression formation in a court setting: The effects of 'powerful' and 'powerless' speech. *Journal of Experimental Social Psychology* **14**: 266–79
- Feingold A 1994 Gender differences in personality: A meta-analysis. *Psychological Bulletin* **116**: 429–56
- Halberstadt A G, Saitta M B 1987 Gender, nonverbal behavior, and perceived dominance: A test of the theory. *Journal of Personality and Social Psychology* **53**: 257–72
- Hall J A 1984 *Nonverbal Sex Differences*. Johns Hopkins University Press, Baltimore, MD
- Hall J A 1998 How big are nonverbal sex differences? The case of smiling and sensitivity to nonverbal cues. In: Canary D J, Dindia K (eds.) *Sex Differences and Similarities in Communication*. Erlbaum, Mahwah, NJ, pp. 155–77
- Hawkins K 1988 Interruptions in task-oriented conversations: Effects of violations of expectations by males and females. *Women's Studies in Communication* **11**: 1–20
- Holmes J 1984 Women's language: A functional approach. *General Linguistics* **24**: 149–78
- Hyde J S 1986 Gender differences in aggression. In: Hyde J S, Linn M C (eds.) *The Psychology of Gender: Advances Through Meta-analysis*. Johns Hopkins University Press, Baltimore, MD, pp. 51–66
- Kling K C, Hyde J S, Showers C J, Buswell B N 1999 Gender differences in self-esteem: A meta-analysis. *Psychological Bulletin* **125**: 470–500
- Kollock P, Blumstein P, Schwartz P 1985 Sex and power in interaction: Conversational privileges and duties. *American Sociological Review* **50**: 34–46
- Lakoff R 1973 Language and woman's place. *Language in Society* **2**: 45–79
- Lynn R, Martin T 1997 Gender differences in extraversion, neuroticism, and psychoticism in 37 nations. *Journal of Social Psychology* **137**: 369–73
- Major B 1999 Gender and self-esteem: A meta-analysis. In: Swann Jr W B, Langlois J H, Gilbert L A (eds.) *Sexism and Stereotypes in Modern Society: The Gender Science of Janet Taylor Spence*. American Psychological Association, Washington, DC, pp. 223–53
- Martin T, Kirkcaldy B 1998 Gender differences on the EPQ-R and attitudes to work. *Personality and Individual Differences* **24**: 1–5
- McCrae R R, Costa P T 1987 Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology* **52**: 81–90
- Mulac A, Lundell T L 1980 Differences in perceptions created by syntactic-semantic productions of male and female speakers. *Communication Monographs* **47**: 111–18
- Mulac A, Studley L B, Blau S 1990 The gender-linked language effect in primary and secondary students' impromptu essays. *Sex Roles* **23**: 439–69
- Newcombe N, Arnkoff D B 1979 Effects of speech style and sex of speaker on person perception. *Journal of Personality and Social Psychology* **37**: 1293–303
- Parsons T 1955 The American family: Its relations to personality and to the social structure. In: Parsons T, Bales R F (eds.) *Family, Socialization, and Interaction Process*. Free Press, Glencoe, IL
- Roberts J E, Gotlib I H 1997 Lifetime episodes of dysphoria: Gender, early childhood loss and personality. *British Journal of Clinical Psychology* **36**(Pt. 2): 195–208
- Rosenthal R, Hall J A, DiMatteo M R, Rogers P L, Archer D 1979 *Sensitivity to nonverbal communication: The PONS test*. Johns Hopkins University Press, Baltimore, MD
- Smith-Lovin L, Brody C 1989 Interruptions in group discussions: The effects of gender and group composition. *American Sociological Review* **54**: 424–35
- Tannen D 1991 *You Just Don't Understand: Women and Men in Conversation*. Ballantine, New York
- Watson D, Clark L A 1984 Negative affectivity: The disposition to experience aversive emotional states. *Psychological Bulletin* **96**: 465–90
- West C, Zimmerman D H 1983 Small insults: A study of interruptions in cross-sex conversations between unacquainted person. In: Thorne B, Kramarae C, Henley N (eds.) *Language, Gender, and Society*. Newbury House, Rowley, MA
- Zimmerman D H, West C 1975 Sex roles, interruptions and silences in conversation. In: Thorne B, Henley N (eds.) *Language and Sex: Differences and Dominance*. Newbury House, Rowley, MA
- Zonderman A B, Siegler I C, Barefoot J C, Williams R B, Costa P T 1993 Age and gender differences in the content scales of the Minnesota Multiphasic Personality Inventory. *Experimental Aging Research* **19**: 241–57

J. S. Hyde

Gender, Economics of

1. *Gender and the Economics of Gender*

The term 'gender' traditionally has referred, as has sex, to the biological differences between men and women. In the 1980s, however, increasingly it came to be used to refer instead to the distinctions which society has erected on this biological base. This article describes the relationship of this concept of gender to economics, focusing on economic analyses of gender differences in behavior and economic outcomes. This focus may be differentiated from feminist economics that is a critique of the discipline and practice of economics from a feminist perspective (see *Feminist Economics*).

2. *Economists' Increasing Interest in Gender Issues*

Historically, gender has not been perceived to be a central concept in economic analysis. However, as the force of current events has thrust gender-related issues to the fore, economists have responded. The outcome has been not only a better understanding of the nature and causes of gender differences, but also an enrichment of the discipline itself.

One reason for the historical neglect of gender issues by the economics discipline is that, traditionally, economists focused on market transactions and, traditionally, most married women were primarily engaged in nonmarket activities. As women have come to comprise a larger share of the paid labor force and as market work has loomed larger in the typical woman's life, interest in the determinants of gender differences in behavior and labor market outcomes has grown. At the same time, the growing interest of economists in analyzing nonmarket work and demographic outcomes such as marriage, divorce and fertility also played an important role in increasing economists' attention to gender-related issues.

While early economic analyses paid scant attention to gender issues, the nineteenth century campaign for women's rights and female suffrage did focus some attention on gender inequality. With few married women employed outside the home, observers like American feminist, Gilman (1898/1966) and Marxist, Engels (1884/1972) (see *Marxian Economic Thought*) focused on the gender division of labor itself, espousing the emancipating effects of women's participation in market work. Classical economist Mill (1869/78) eloquently argued for women's 'admissibility to all the functions and occupations hitherto retained as the monopoly of the stronger sex' and expressed the belief that women's 'disabilities elsewhere are only clung to in order to maintain their subordination in domestic life.' These views highlight the notion that understanding women's status involves

an understanding of both the market and family spheres and of the interrelationship between the two. This insight remains applicable to the current day.

3. *Time Allocation in the Family Context*

Prompted in part by their desire to understand the causes of the rising labor force participation of married women in the post-World War II period, economists extended the traditional theory of labor supply (see *Labor Supply*) to consider household production.

3.1 *The Traditional Theory of Labor Supply*

The traditional theory of labor supply was a simple extension of consumer theory. Individuals maximize their utility, which is derived from market goods and leisure, subject to budget and time constraints. The labor supply decision is determined by the opportunity wage (the wage rate the individual could earn in the labor market), nonlabor income (the amount of income available to them from other sources), and the individual's tastes or preferences for work vs. leisure.

3.2 *Household Production and the Allocation of Time*

While this theory is sufficient for many purposes, it has substantial limitations for understanding the determinants of the gender division of labor in the family and the factors influencing women's labor force participation. The key to addressing these issues is a fuller understanding and analysis of the household production process.

An early pioneer in analyzing household production was Reid (1934), but it was the work of Becker and Mincer in the 1960s that was particularly influential in bringing these issues into the mainstream. In analyzing married women's labor force participation in the US, Mincer (1962) pointed to the importance for them of the three-way decision among market work, non-market work, and leisure. He attributed the growth in married women's participation during the early post-war period to women's rising real wages which induced them to substitute market work for housework. This response was large enough to more than offset the negative income effect on their labor supply associated with the growth in their husbands' real wages.

Becker (1965) advanced this process further by proposing a general theory of the allocation of time. In this and other work (summarized in Becker 1981), he laid the foundations of the 'new home economics,' including economic analyses of time allocation, marriage, divorce, and fertility. Interestingly, while Mincer opened a window on household production by distinguishing nonmarket work from leisure, Becker was

able to provide a further advance by again eliminating the distinction. However, while in the traditional labor supply model all nonmarket time is spent in leisure, in Becker's model all nonmarket time is spent in household production. Becker assumes that households derive utility from 'commodities' that are produced by inputs of market goods and nonmarket time.

These ideas may be illustrated by considering the relationship between children and women's labor force participation. Young children may be viewed as a 'time-intensive' commodity. Traditionally, it has been the mother who was the primary care giver. And, while it is possible to substitute market goods and services (e.g., babysitters, day care centers) for home time in caring for children, these alternative production techniques tend to be costly and may be difficult to implement. Thus, at any point in time, the probability that a woman will participate in the labor force is expected to be inversely related to the number of small children present. Over time, the increase in women's participation rates has been associated with decreasing birthrates, as well as increases in the availability of alternative child care arrangements.

3.3 Trends in Labor Supply

Continuing in the tradition of Mincer (1962), the labor supply model has been used to explain the trends in female labor force participation. One criticism that has been made of this type of economic modeling is that it abstracts from historical and institutional factors. Goldin (1990), however, turns this feature of economic models to her advantage in an especially innovative application. She finds that a model applied to decennial US Census data for the 1890–1980 period overpredicts married women's labor force participation for 1930, 1940, and 1950. She presents this as evidence that the process of change in participation was retarded for married women by institutional barriers like marriage bars (i.e., prohibitions against employing married women that were particularly prevalent in teaching and clerical work) and the lack of availability of part-time employment.

As wage inequality rose in many advanced countries in the last quarter of the twentieth century, increasing attention focused on the contribution of this trend to the continued rise in the labor force participation of wives and the related question of the consequences of the increased employment of wives for family income inequality. On one hand, the income effect suggests that wives of low skilled men would be drawn into the labor force to compensate for the deteriorating labor market position of their husbands. On the other hand, the substitution effect implies that the increasing rewards to labor market skills which underlie the rise in wage inequality, would draw more skilled (e.g., more highly educated) women into the labor market. Since more highly educated individuals tend to marry

each other, a larger role for the income effect in explaining the trends would tend to reduce the correlation between wives' and husbands' earnings, while a larger role for the substitution effect would tend to increase the correlation.

Juhn and Murphy (1997) found for the US that employment gains were in fact largest for wives of middle and high wage men. Their findings strikingly mirror Mincer's (1962) for the immediate post-World War II period: own wage effects dominate cross-effects between husband and wife in accounting for changes in female employment. The correlation between husbands' and wives' earnings has risen in the US, both as a consequence of these participation trends and also due to a rise in the correlation between the earnings of spouses when both are employed. A similar rise has been reported for Britain, which also experienced a substantial increase in wage inequality during this time. However, while this increased correlation contributed to rising inequality in family income in both the US and Britain, the rise in family income inequality among married couples in both countries was principally due to rising inequality of husbands' earnings (Cancian and Reed 1998, Machin and Waldfogel 1994).

3.4 The Gender Division of Labor

In the discussion of the allocation of time above, it was simply assumed that women tend to bear the primary responsibility for childcare and housework. However, the gender division of labor in the family is also an issue that the new home economics addresses. According to Becker (1981), the division of labor will be dictated by comparative advantage. To the extent that women have a comparative advantage in household and men in market production, it will be efficient for women to specialize at least to some extent in the former while men specialize in the latter. In Becker's view, the increased output resulting from this arrangement maximizes family utility and constitutes one of the primary economic benefits to marriage. Thus, women's rising labor force participation is expected to reduce the gains from marriage, thereby contributing to the trend towards higher divorce and lower marriage rates.

As women have been increasingly working throughout the life cycle, more attention has focused on their ability to combine work and family responsibilities successfully. Since men tend not to face the same dilemmas as women, or not to face them to the same extent, this is a potential source of gender differences in wellbeing. Waldfogel (1998) details the prevalence of a 'family penalty' in many advanced countries, that is, lower earnings for women who are mothers, all else equal. This likely reflects the loss of their tie to the firm and associated firm-specific skills when they withdraw from the labor force upon the birth of a child.

Some evidence suggests that policy intervention could reduce the size of the family penalty. For example, Waldfogel (1998) reports evidence from analyses of microdata for the US, Britain, and Japan indicating that women who were covered by maternity leave were considerably more likely to return to their prior employer after childbirth. And women who return to their employer after a birth have higher subsequent wages. Consistent with these results, Ruhm (1998) finds, using aggregate data on 16 countries during the 1969–88 period, a positive, but not significant, effect of 1 percent on women's wages relative to men's of short leave entitlements (1–13 weeks), although longer leave entitlements (over 26 weeks) had a significantly negative effect of 3 percent. The negative effect of longer leaves could indicate that such leaves encourage women to extend their time away from the firm. It is also possible that the higher costs of long mandated leaves provide greater incentives to employers to discriminate against women in wages or job assignments.

A consideration of the difficulties women face in combining work and family raises the possibility of conflicts of interest between husband and wife which in turn makes defining 'family utility' problematic (Ferber and Birnbaum 1977). Such concerns have been addressed by the growing application of bargaining models to family decision-making, led by Manser and Brown (1980) and McElroy and Horney (1981). These models assume that the preferences of husbands and wives may diverge and that outcomes for each spouse can be modeled using game theory (see *Game Theory: Cooperative Games*; *Game Theory: Noncooperative Games*). In the most popular application of these models to the family, the cooperative bargaining model of marriage, the utility received by each spouse in the Nash bargaining solution is related positively to the utility of each partner at the 'threat point,' most commonly taken to be the utility that each would receive if the marriage irretrievably broke down. If it is assumed that, in the event of divorce, each partner maintains ownership of income received separately within marriage, a woman who works outside the home will have a higher utility at the threat point and hence a more favorable distribution within marriage. Studies reviewed by Lundberg and Pollak (1996) provide empirical support for this model in their finding that earned or unearned income received by the wife has a different effect on demand patterns than income received by the husband.

As women anticipate spending increasingly more of their adult lives in the labor market, their investments in market-oriented human capital may be expected to continue to grow and their comparative advantage in home work to decline. Moreover, as the quality of the opportunities open to women in the labor market continues to improve, the disadvantages of specialization in home work in the form of forgone earnings and possibilities for career advancement will

also rise. Thus, greater sharing of household responsibilities between men and women is likely to become increasingly prevalent.

4. Gender Differences in Labor Market Outcomes

Women earn less than men, on average, in all advanced economies, though the extent of the gender pay gap varies considerably across countries and has been declining in many of them in recent years (e.g., Blau and Kahn 1996). Economists' understanding of gender inequality in such economic outcomes has been accommodated principally through the development of new and interesting applications of existing theoretical approaches. Such theories must explain the considerable employment segregation of men and women—by occupation, industry, and firm—as well as gender pay differentials. Segregation of men and women by occupation, for which summary data are most readily available, has been found to be prevalent in all advanced economies, though it has been declining in recent decades in most OECD countries, including the US (e.g., Blau et al. 1998, Anker 1998).

As in the case of the analysis of women's roles in the family, the catalyst for the development of analyses of pay and occupation differences was provided by external events. Some moderate degree of interest in the pay issue was generated in England by the World War I experience when there was some substitution of women into traditionally male civilian jobs, although not nearly to the degree that would later occur during World War II. Questions of the appropriate pay for women under these circumstances arose and stimulated economic analyses of the gender pay differential. These included the work of Fawcett (1918) and Edgeworth (1922), both of whom assigned a prominent causal role to occupational segregation, thus providing the antecedents for Bergmann's (1974) overcrowding model (discussed below). The analysis of gender differentials in the labor market received further impetus in the early 1960s, this time in the US, with the development of the women's liberation movement and the passage of antidiscrimination legislation.

Two broad approaches to the issue have since evolved. First is the human capital explanation which lays primary emphasis on men's and women's own choices in explaining occupation and pay differences. Second are a variety of models of labor market discrimination (see *Discrimination, Economics of*) which share the common characteristic of placing the onus for unequal outcomes on differential treatment of equally (or potentially equally) qualified men and women in the labor market. More recently, some advances have been made by considering the gender pay gap and other demographic differentials in the context of the overall structure of wages (see *Wage Differentials and Structure*).

4.1 The Human Capital Explanation

The human capital explanation for gender differences in occupations and earnings, developed by Mincer and Polachek (1974) and others, provides an explanation for gender differences in market outcomes on the basis of the traditional division of labor by gender in the family. Anticipating shorter and more discontinuous work lives as a consequence of their family responsibilities, women have lower incentives to invest in market-oriented formal education and acquire less on-the-job training than men. Their resulting smaller human capital investments will lower their earnings relative to those of men. Working in a similar direction is Becker's (1985) model in which the longer hours women spend on housework lowers the effort they put into their market jobs compared to men, controlling for hours worked, and hence reduces their wages.

These considerations are also expected to produce gender differences in occupations as women choose occupations for which human capital investments are less important and in which the skill depreciation that occurs during time spent out of the labor force is minimized. Women especially will avoid jobs requiring large investments in firm-specific skills (i.e., skills which are unique to a particular enterprise), because the returns to such investments are reaped only as long as one remains with a particular employer. At the same time, employers are expected to be reluctant to hire women for such jobs because they bear some of the costs of firm-specific training (see the discussion of statistical discrimination below).

To the extent that gender differences in outcomes are not accounted for fully by productivity differences derived from these and other sources, models of labor market discrimination offer an explanation.

4.2 Models of Labor Market Discrimination

Theoretical work in this area was initiated by Becker's (1957) model of racial discrimination. Becker conceptualized discrimination as a taste and analyzed three cases, those in which the discriminatory tastes were held by employers, co-workers, and customers or clients. Under certain circumstances, such discrimination will cause a wage differential between men and women. So discriminatory employers will only hire women at a sufficient wage discount that compensates them for the disutility of employing women. Discriminatory male workers will demand a wage premium to work with women thus raising men's relative wages, and the reluctance of discriminatory customers or clients to buy goods or services provided by women will make women less productive in terms of revenue brought in, thus depressing women's relative wages.

Models based on tastes for discrimination are also consistent with employment segregation, but do not necessarily predict it will occur. If wages are flexible, it is altogether possible that discrimination will result in

lower pay for women, but produce little or no segregation. However, if discriminatory tastes against women in traditionally male pursuits are both strong and prevalent, women may tend to be excluded from these areas. If such segregation does occur, it may or may not be associated with gender pay differentials. In the presence of sufficient employment opportunities in the female sector, equally qualified women may earn no less than men.

The relationship between occupational segregation and earnings differentials is clarified in Bergmann's (1974) overcrowding model. If, for whatever reason (i.e., labor market discrimination or their own choices), potentially equally qualified men and women are segregated by occupation, the wages in male and female jobs will be determined separately by the supply and demand for labor in each sector. Workers in male jobs will enjoy a relative wage advantage if the supply of labor is more abundant relative to demand for female than for male occupations. Such 'crowding' of female occupations can also widen differentials between male and female jobs that would otherwise exist due to women's smaller human capital investments or employers' reluctance to invest in women's human capital.

One question that has been raised about the Becker model is how employer discrimination can persist in the long run in the face of competitive forces. The least discriminatory firms, which hire more lower-priced female labor, would have lower costs of production. In the long run, they should expand and drive the more discriminatory firms out of business. One answer, suggested initially by Becker himself, is that discrimination will be located in sectors of the economy that are not competitive (see *Market Structure and Performance*). While Becker emphasized monopolistic elements in the product market, a related approach targets monopsonistic power on the part of the employer in the labor market (e.g., Madden 1973).

In addition, models of statistical discrimination were later developed, in part to explain the persistence of discrimination in the long run in the face of competitive forces. Such models assume a world of uncertainty and imperfect information; thus differences between groups in the expected value of productivity or in the reliability with which productivity may be predicted are expected to result in differences in the treatment of members of each group. So, for example, employers may believe that women are more likely to quit their jobs than men with similar observed characteristics (education, experience, etc.) or they may believe that women's turnover can be less reliably predicted by their measured characteristics than men's. As a consequence, they may pay women less, exclude them from jobs requiring substantial firm-specific training, or deny them promotions.

It has been argued that such statistical discrimination, making decisions on the basis of the average characteristics of the group, is consistent with profit

maximization and can thus persist in the face of competitive forces. However, Aigner and Cain (1977) contend that such models are no more convincing in explaining the persistence of discrimination than models based on tastes for discrimination. To the extent that employers' views are accurate, the lower expected productivity of women will reduce their wages but women as a group will be paid their expected productivity. This does not constitute labor market discrimination as economists define it, that is, pay differences that are not accounted for by productivity differences. Presumably they would make a similar argument regarding hiring and promotion decisions. Moreover, they note that when employer beliefs regarding average differences are erroneous, discrimination clearly exists but in their view discrimination based on such misperceptions is even less likely to persist in the long run than discrimination based on tastes. Finally, though they acknowledge that less reliable predictions of a group's productivity combined with risk aversion by employers could produce a discriminatory differential, a perfectly elastic supply of risk neutral entrepreneurs could be expected to erode discriminatory differentials based on this factor.

However, such models may have greater usefulness in explaining the persistence of gender differentials in pay than Aigner and Cain's analysis suggest. Consider first the situation where employer perceptions are correct. In this case, the concept of statistical discrimination provides a plausible reason for the existence and persistence of pay and occupational differences between men and women with similar measured characteristics; this is in itself a useful contribution even if it does not shed light on the persistence of economic discrimination in the long run. But if the employer views are correct, is it appropriate to consider this a form of 'discrimination' in any sense? From a normative perspective, the answer may be yes, to the extent that basing employment decisions on a characteristic like sex—a characteristic that the individual cannot change—could be viewed as inequitable. Indeed, the practice of judging an individual on the basis of group characteristics rather than upon his or her own merits seems the very essence of stereotyping or discrimination. Such behavior is certainly not legal, for example, under the antidiscrimination laws and regulations.

Now consider the situation where employer perceptions are incorrect. If statistical discrimination is accompanied by feedback effects, this may be a more credible source of persistent discriminatory pay differences between men and women than Aigner and Cain suggest. For example, if employers incorrectly expect that women are more likely to quit their jobs, they may respond by giving women less firm-specific training or assigning them to dead end jobs. Faced with few incentives to remain on the job, women may respond by exhibiting exactly the higher turnover that employers expect.

This criticism of the Becker model—its inability to explain the persistence of discrimination against women in the long run—has proved to be a double-edged sword in that it has led some economists to doubt that labor market discrimination is responsible, in whole or part, for observed gender differences in labor market outcomes. Yet it is important to recognize that discrimination is an intrinsically complex process. From this perspective, it is not surprising that no easy solution has been found to the question of why discrimination may persist in the face of competitive forces. Similarly, the various models of discrimination, each emphasizing different motivations and different sources of this behavior, need not be viewed as alternatives. Rather, each may serve to illuminate different aspects of this complex reality. When it is further recognized that, as our discussion above suggests, discrimination can adversely affect women's human capital investments and labor force attachment by lowering the market rewards to this behavior (e.g., Weiss and Gronau 1981), it is clear that the potential impact of discrimination on women's labor market outcomes may be large indeed.

4.3 Evidence on Human Capital, Discrimination, and the Gender Pay Gap

The typical approach to analyzing the sources of the gender pay gap is to estimate wage regressions specifying the relationship between wages and productivity-related characteristics for men and women (see *Econometric Software*). The gender pay gap may then be decomposed statistically into two components: the first, due to gender differences in measured characteristics and the second, which is 'unexplained' and potentially due to discrimination.

Such studies provide considerable empirical support for both human capital differences and labor market discrimination in explaining the gender pay gap. For example, Blau and Kahn (1997) report that, in 1988, US women's wages were 72 percent of men's, on average. Adjusting for gender differences in human capital (i.e., education and experience) would raise the ratio to 81 percent, while additionally adjusting for gender differences in industry, occupation and unionism would further increase it to 88 percent. Thus, while measured characteristics are important, women still earn considerably less than similar men, even when all measured characteristics are taken into account.

One problem with studies of this type is that discrimination may be overestimated to the extent that unexplained pay differentials between men and women are due in whole or part to gender differences in unmeasured qualifications. On the other hand, to the extent that the regression models include controls for characteristics like industry or occupation that could themselves be due to discrimination, discrimination may be underestimated. Recent studies that take a

different approach to the question provide additional evidence of discrimination that addresses the former concern.

First, studies have applied traditional econometric techniques to especially homogeneous groups and employed extensive controls for qualifications, thus minimizing the effect of gender differences in unmeasured characteristics; and still found evidence of an unexplained pay difference between men and women (see, e.g., Wood et al. 1993) which focused on graduates of the University of Michigan Law School).

A second set of studies uses an experimental approach. Neumark (1996) analyzed the results of a hiring 'audit' in which male and female pseudo-job seekers were given similar résumés and sent to apply for waiter or waitress jobs at the same set of Philadelphia restaurants. The results provided statistically significant evidence of discrimination against women in high-priced restaurants where earnings of workers are generally higher than in the other establishments. A second study, Goldin and Rouse (2000), examined the impact of the adoption of 'blind' auditions by symphony orchestras in which a screen is used to conceal the identity of the candidate. Using data from actual auditions, it was found that the screen substantially increased the probability that a woman would advance out of preliminary rounds and be the winner in the final round.

Finally, there is some evidence that gender differentials in wages or employment are larger in less competitive industries. For example, Black and Strahan (1999) report that with the deregulation of the banking industry beginning in the mid-1970s, the gender gap in pay in banking declined. This is consistent with the expectation that discrimination will be more severe in firms that are shielded to some extent from competitive pressures and thus indirectly provides support for the view that discrimination is a factor influencing gender wage differentials.

4.4 The Role of Wage Structure

Both the human capital and discrimination explanations for the gender pay gap might be considered 'gender specific' in nature, that is, focused on gender differences in either qualifications or treatment. Analyses of gender differentials traditionally have tended to emphasize these types of gender-specific factors. Important work by Juhn et al. (1991) on trends in the race pay gap over time in the US, has highlighted an additional factor, overall wage structure, which is also relevant to analyzing the gender pay gap. Wage structure is the array of prices determined for labor market skills and the rewards to employment in particular sectors.

The human capital model suggests that men and women tend to have different levels of labor market

qualifications (especially work experience) and to be employed in different occupations and perhaps in different industries. Discrimination models, too, suggest that women may be segregated into different sectors of the labor market. This implies that the returns to skills and the size of premiums for employment in particular sectors potentially play an important role in determining the pay gap. All else equal, the larger the returns to skills and the larger the rents received by individuals in favored sectors, the larger will be the gender pay gap.

Blau and Kahn (1996) have applied this insight to understanding international differences in the gender pay gap in order to address a paradox. The relative qualifications of American women are high compared to women in other countries, and the US has had a longer and often stronger commitment to antidiscrimination laws than most economically advanced nations. Yet the US traditionally has been among the countries with the largest gender pay gaps. The resolution to this paradox lies in the decentralized system of wage setting in the US that has resulted in a highly dispersed wage distribution and high returns to labor market skills and employment in favored sectors. In contrast, the greater role of unions and government in the labor markets of other economically advanced nations has resulted in higher wage floors and more compressed wage distributions from which employed women, as a relatively low wage group, have benefited disproportionately.

This raises the question of how, given the rising wage inequality that has occurred in many of the advanced industrialized countries, women have nonetheless been able to narrow the wage gap with men in many cases. A detailed study of the US by Blau and Kahn (1997) suggests that women have been able to compensate for the adverse effects of rising inequality and higher rewards to skills by improving their relative qualifications, particularly their average work experience and their occupational distribution. In addition, the unexplained portion of the pay gap declined, reflecting either an upgrading of women's unmeasured labor market skills or a decline in discrimination against them, or both.

Similar studies of the UK (Dolton et al. 1996) and Sweden (Edin and Richardson 1999) also find that gender specific factors were principally responsible for declines in the gender gap in those countries over the 1970s, although during a subperiod in which considerable wage compression occurred in Sweden due to its 'solidarity' wage policy (1968–71), wage structure played a major role.

5. Conclusion

This review suggests that women's economic roles in the family and their labor market outcomes are closely related. This has undoubtedly played a role in the

persistence of gender differences in both spheres. However, increasing opportunities for women in the labor market may be expected to reduce gender differences in family roles, while such a diminution of adherence to the traditional division of labor in the family is likely to enhance women's position in the labor market further. Thus, just as a fuller understanding of the interrelationships between women's roles in the family and their status in the labor market helps us to understand the persistence of gender inequality in economic outcomes, it also enables us to appreciate how changes in either one of these spheres, or both, can induce a mutually reinforcing process of change. Recent reductions in the pay gap in many of the economically advanced countries may well signal the operation of such a process.

See also: Comparable Worth in Gender Studies; Economic Development and Women; Feminist Economics; Gender and Feminist Studies in Economics; Gender and Technology; Sex Differences in Pay; Sex Segregation at Work; Wage Differentials and Structure

Bibliography

- Aigner D J, Cain G G 1977 Statistical theories of discrimination in labor markets. *Industrial and Labor Relations Review* **30**: 175–87
- Anker R 1998 *Gender and Jobs: Sex Segregation of Occupations in the World*. International Labour Office, Geneva
- Becker G S 1957 *The Economics of Discrimination*, 2nd edn. University of Chicago Press, Chicago
- Becker G S 1965 A theory of the allocation of time. *Economic Journal* **75**: 493–517
- Becker G S 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA
- Becker G S 1985 Human capital, effort, and the sexual division of labor. *Journal of Labor Economics* **3**: 533–58
- Bergmann B R 1974 Occupational segregation, wages and profits when employers discriminate by race or sex. *Eastern Economic Journal* **1**: 103–10
- Black S E, Strahan P E 1999 Rent-sharing and discrimination: the effects of deregulation on the labor market. Working paper, Federal Reserve Bank of New York
- Blau F D, Ferber M A, Winkler A E 1998 *The Economics of Women, Men, and Work*, 3rd edn. Prentice-Hall, NJ
- Blau F D, Kahn L M 1996 Wage structure and gender earnings differentials: An international comparison. *Economica* **63**: S29–S62
- Blau F D, Kahn L M 1997 Swimming upstream: Trends in the gender wage differential in the 1980s. *Journal of Labor Economics* **15**: 1–42
- Cancian M, Reed D 1998 Assessing the effects of wives' earnings on family income inequality. *The Review of Economics and Statistics* **80**: 73–9
- Dolton P, O'Neill D, Sweetman O 1996 Gender differences in the changing labor market: The role of legislation and inequality in changing the wage gap for qualified workers in the United Kingdom. *Journal of Human Resources* **31**: 549–65
- Edgeworth F Y 1922 Equal pay to men and women for equal work. *Economic Journal* **32**: 431–57
- Edin P-A, Richardson K 1999 Swimming with the tide: Solidarity wage policy and the gender earnings gap. Working Paper, IFAU-Office of Labour Market Policy Evaluation. Uppsala, Sweden
- Engels F 1884/1972 *The Origin of the Family. Private Property and The State*. International Publishers, New York
- Fawcett M G 1918 Equal pay for equal work. *Economic Journal* **28**: 1–6
- Ferber M A, Birnbaum B G 1977 The new home economics: Retrospects and prospects. *Journal of Consumer Research* **4**: 19–28
- Gilman C P 1898/1966 *Women and Economics: A Study of the Economic Relation Between Men and Women as a Factor of Social Evolution*. Small, Maynard, Boston
- Goldin C 1990 *Understanding the Gender Gap*. Oxford University Press, New York
- Goldin C, Rouse C 2000 Orchestrating impartiality: The impact of 'blind' auditions on female musicians. *American Economic Review* **90**: 715–41
- Juhn C H, Murphy K M 1997 Wage inequality and family labor supply. *Journal of Labor Economics* **15**: 72–97
- Juhn C H, Murphy K M, Pierce B 1991 Accounting for the slowdown in black-white wage convergence. In: Kosters M (ed.) *Workers and Their Wages*. AEI Press, Washington, DC
- Lundberg S, Pollak R A 1996 Bargaining and distribution and marriage. *Journal of Economic Perspectives* **10**: 139–58
- Machin S, Waldfogel J 1994 The decline of the male breadwinner: Changing shares of husbands' and wives' earnings in family income. London School of Economics Suntory-Toyota International Centre for Economics and Related Disciplines Working Paper: WSP/103
- Madden J F 1973 *The Economics of Sex Discrimination*. Heath and Co., Lexington, MA
- Manser M, Brown M 1980 Marriage and household decision-making. *International Economic Review* **21**: 31–44
- McElroy M B, Horney J M 1981 Nash-bargained household decisions: toward a generalization of the theory of demand. *International Economic Review* **22**: 333–49
- Mill J S 1869/1878 *The Subjection of Women*, 4th edn. Longmans, Green, Reader and Dyer London
- Mincer J 1962 In: *Aspects of Labor Economics*. National Bureau of Economic Research, Princeton University Press, NJ
- Mincer J, Polachek S 1974 Family investments in human capital: Earnings of women. *Journal of Political Economy* **82**: S76–S108
- Neumark D, Bank R J, Van Nort K O 1996 Sex discrimination in restaurant hiring: an audit study. *Quarterly Journal of Economics* **111**: 915–41
- Reid M G 1934 *Economics of Household Production*. Wiley, New York
- Ruhm C J 1998 The economic consequences of parental leave mandates: Lessons from Europe. *Quarterly Journal of Economics* **113**: 285–317
- Waldfogel J 1998 Understanding the 'family gap' in pay for women with children. *Journal of Economic Perspectives* **12**: 137–57
- Weiss Y, Gronau R 1981 Expected interruptions in labor force participation and sex-related differences in earnings growth. *Review of Economic Studies* **48**: 607–19

Wood R G, Corcoran M E, Courant P N 1993 Pay differences among the highly paid: The male-female earnings gap in lawyers' salaries. *Journal of Labor Economics* 11: 417-41

F. D. Blau

Gender, Feminism, and Sexuality in Archaeological Studies

An important new field within archaeology, often referred to as the archaeology of gender, archaeology and gender, or gender archaeology, combines a critical feminist study of gender, sex, and sexuality with a feminist analysis and examination of central concepts within archaeology, and the ways in which archaeological knowledge is produced, presented, and represented. Although there are many different kinds of feminism, it is taken here in its broadest sense, with its major goal being a critique of the determination of women's lives by nature/biology, reducing their multifaceted lives to an essentialist category 'woman.' Rather, gender, sex, and sexuality are not constrained by essentialist and narrowly defined categories, but are seen as being historically constituted and varied categories, changing over time. The aims of feminist archaeology are to both investigate and challenge the assumed content of, and relationships between, gender categories such as women and men, femininity and masculinity; and to study gender divisions and sexuality over time and in different social and cultural contexts in the past. Archaeology is a way of knowing and investigating the past that is based on a particular androcentric understanding of objectivity that is historically constituted Euro-American. Feminist archaeology, using feminist critiques of science, challenges this traditional construction of archaeological knowledge.

1. A Slow Start

Although tied to the goals of Second Wave feminism, which not only critiqued gender relations in contemporary society, but also began very specifically to investigate how assumptions about the 'naturalness' of sex and gender affected the production of knowledge, feminism was slow in coming to archaeology. In the mid-1970s the first feminist critiques of equity issues in archaeology started in Scandinavia, particularly Norway. This concern quickly expanded to critiques of analyses and presentations of women in the past, and resulted in the founding of the first, and as yet only, feminist archaeology journal, *KAN: Kvinner i Arkeologi i Norge* (KAN: Women in Archaeology in Norway) in 1985. The previous year, Margaret Conkey and Janet Spector's groundbreaking article critiquing androcentrism in archaeology and discussing the possibilities of studying gender in the past was pub-

lished in the United States. It was, however, at the beginning of the 1990s that the archaeology of gender and feminist critiques exploded on the archaeological scene.

During the course of the 1990s there were numerous publications on the archaeology of gender, particularly edited volumes of symposia and conference papers, as well as a few monographs. Few of these publications use the word 'feminist' in their titles, preferring various combinations of the words 'gender' and 'archaeology.' However, the majority are feminist in content and intention, and in their interpretations of the past.

Feminist archaeology in its beginnings was particularly an Anglo-American, Australian, and Scandinavian phenomenon, but expanded in the 1990s to other European countries, particularly Germany and Spain. The slow development of feminist archaeology in some countries can be explained partially by how archaeology is viewed within the different academic systems. Archaeology and anthropology in North America are closely linked, and feminist archaeology benefited from its close connection with feminist anthropology. In many European countries, however, archaeology is not a social science but a discipline in the humanities, and thus society and social relations were not in the forefront of many archaeological studies. At the same time archaeology is one of the few humanities that works with material culture, rather than texts, and has thus been relatively isolated from the intense feminist discussions within the humanities.

This basic disciplinary difference between the North American and the European archaeology of gender is probably one reason for the different perspectives these two archaeologies had initially. North American gender archaeology has been concerned more with social processes while European gender archaeology continued with traditional interests in cultural (now gender) identity and distinctions. However, as the archaeology of gender has developed and changed, this tendency is now minimized as not only anthropology but social theory, post-structuralism, post-processualism, and feminist theory have become important elements of archaeological thinking in all areas.

The feminist philosopher of science Alison Wylie (1991) has considered why it took so long for feminist critiques to reach archaeology and emphasizes the essentialist assumptions in archaeological theorizing tied to macroscale analyses. Since these essentialist assumptions were based on biological determinism it was vital in the beginning to make a distinction between sex, a biological construct, and gender, a social construct. Here one must note the difference between the English language separation of sex and gender, and the fact that many other languages have but one word for both. Recent work in feminist archaeology, both in English and non-English-speaking countries, resurrects sex and sexuality from the

merely biological, but while viewing sex/sexuality as a social construction associated with gender, this work does not neglect changing biological processes.

2. Back to Basics

Much feminist work in the 1980s and 1990s concentrated on gender rather than sex and sexuality. As with feminism in general, this was necessary in order to negate the importance of previous biological determinism and essentialism in placing women in a subordinate position in the past. Androcentric archaeology considered women essentially the lesser part of the male–female, masculine–feminine, man–woman dichotomies. Women were often reduced conceptually to their biological function of giving birth. This also placed women on the lesser side of a whole series of related hierarchical dichotomies, such as public–domestic, public–private, powerful–dominated, active–passive, etc., thus giving biological reasons for social differences and social formations. For example, the origins of the nuclear family were tied closely tied to stable binary sex and gender distinctions which were seen as beginning almost 4 million years ago. This pattern was established in the archaeological literature before feminist critiques turned this set of universal hierarchical dichotomies into a problem to be investigated rather than assumed.

However, the classic feminist concern with the universal subordination of women played a peripheral role in feminist studies of the past. It was obvious that women were given a subordinated role in archaeologists' textual and visual presentations, and feminists critiqued such androcentric and essentialist assumptions. Conkey and Spector (1984) noted that despite its professed objectivity and neutrality, archaeology was 'neither objective nor inclusive on the subject of gender.' This false notion of objectivity contributed to making women invisible. The gender paradigms implicitly treated women and men differently, consistently linked certain activities to each gender/sex, and made men central and active in all societies in the past, while women were passive and 'at home.'

Feminist interpretations of the same past critiqued the subordinate role given to women in archaeological texts, exhibits, and illustrations. Feminist archaeologists have worked from the premise that universal subordination was not necessarily the case in the past, that gender ideologies and women's roles varied and were given different statuses over time and space, again emphasizing that these concepts are something to be investigated rather than assumed. In the 1980s both feminist theoretical and empirical work assumed that gender categories were not necessarily exclusive and that women were an active part of culture and society in the past. Studies were concerned

with defining women's various roles, contributions, and activities. Task differentiation was developed as a framework and methodology by Janet Spector in order to study the various dimensions of women's and men's activities in past societies. Some form of this method is still often used in current analyses in the archaeology of gender, and is one of many contemporary efforts to put women back into our understanding of the past.

Following feminist critiques of science, feminist archaeologists examined critically the basic assumptions behind previous understandings of the past. The majority of articles on gender have as implicit but important goals the examination of what counts as data, what counts as interpretation of that data, and what is considered acceptable interpretation and theory in the male-dominated archaeological community. Thus they examine the selective processes which determine relevant data for interpretation, and how this over-determines interpretations and results in ignoring gendered processes and structures. Why are 'men's' artifacts simply assumed to be of primary interest? Why did important interpretational frameworks consider only men to be the prime movers of social and cultural change? In addition, they examine androcentric theoretical concepts which were previously, and often currently, considered to be the only reasonably legitimate and, indeed, the only 'natural' and 'rational,' theoretical frameworks. Engendering the past not only redefines and extends what is considered as data, interpretation and theory, but also challenges many taken-for-granted concepts and models.

3. Approaches to Gender and Sexuality

As with all feminist projects, the archaeology of gender is grounded in the analysis and understanding of women's lives. Women's lives are reflected in archaeological remains of the past, in the material culture (objects and structures) of past societies. Richly detailed empirical studies, which put women (and children) back in to our understanding of the past, abound. These are tied to differing theoretical views of gender as a structuring principle, and examine how material culture was used to express and construct gender roles, differences, ideologies and gendered identities in past societies. Foregrounding gender effects changes in traditional archaeological thinking and leads to an interest in social relations between genders and between gendered individuals of differing ages, classes, ethnicity, and race.

Engendered studies of Stone Age archaeology have gone further than a concern with ecology and technology by considering the wider perspective of social relations between individuals and groups which leads to a restructuring of traditional archaeological understandings, offering new insights into social relations,

ideologies, and sociocultural institutions. Margaret Conkey (1991) considers the wider contexts in which tools were made, used and reused at a Magdalenian site where several families lived ('aggregation sites'). By viewing tools not as single objects, but as objects representing a complex set of technological, productive, and engendered social relations, she is able to show the changing contexts of action and contexts for power in which people and objects were involved. Gender is always in process, and the sexual/gender division of labor is being changed and negotiated constantly. These changing contexts of action are associated with the renegotiation of power relations and structures. This type of analysis challenges traditional views of stable and exclusive gender categories and opens up possibilities for analyses of archaeological material that can reveal that there can be different models of gender in one society, that these models may change in relation to particular contexts and changes in the lives of individuals.

European archaeologists have made considerable use of burial material in discussions of the construction of gender. In her study of the Viking period Oseberg burial, a rich grave with the burial of two women in a ship, Elisabeth Arwill-Nordbladh (1998) uses M. Conkey's 'context of action,' here relating to textile production, farm economy, communication, and ritual performance. Artifacts in the grave were analyzed in order to see how gender could have been constructed through the various contexts of action of which these objects were a part. Gender construction is not seen as following a strict set of gender categories, but as a process that can be maintained, changed, negotiated, or resisted through practice or performative acts. In this manner gender is seen as a difference between, and as a difference within, gender categories, so that the complexity of gender dynamics comes to the forefront.

Marie Louise Stig Sørensen (1991), working with grave material from the North European Bronze Age, has done a series of studies on how objects are active in expressing gendered identity. She examines especially how gender is constructed through appearance by examining the composition and placement of grave goods in relation to the body (skeleton) and in relation to their wider sociocultural contexts of use and meaning. In this manner it is possible to analyze not only gender but also cross-cutting social categories such as age, ethnicity and status.

Concentration on gender leads to an initial neglect of sex and sexuality. However, as feminism and feminist theory developed and was increasingly concerned with difference, rather than solely with the binary distinctions of gender, and view sexuality and the body as both socially constructed and historically constituted, so too has feminist archaeology newly expanded its interest to these previously neglected areas. Sexuality as an expression of desire is not limited to the heterosexual body, but is examined in its variety of

expression, including homosexuality. Much of this work is on representations of the body in art, particularly in classical Mediterranean cultures, but also in studies of prehistoric rock art.

As can be seen from this short overview, there are two threads in feminist archaeology. One is the 'archaeology of gender,' while the other is 'gendered archaeology' (Roberts 1993), what I have referred to as feminist archaeology. The archaeology of gender, while often premised on a feminist perspective and concerned with putting women and children back into our understanding of the past, reiterates traditional archaeological concepts and principles of interpretation. A gendered archaeology is explicitly feminist and tied closely to feminist critiques of science in an effort to generate new engendered ways of thinking, doing, and presenting archaeology. A gendered archaeology/feminist archaeology does not simply add women's contribution to past societies, but is concerned with interrogating and reworking our unquestioned starting premises and taken-for-granted concepts.

4. Feminist Practice

Although often considered to be neutral and objective, archaeological practice is highly gendered, and often implicitly so. Feminist interests include archaeological practice, particularly fieldwork and analysis, workplace issues, pedagogy, cultural heritage management, including museums, new ways of writing and presenting archaeology, and concern for the local communities in which one works.

Joan Gero (1996) has examined how fieldwork, particularly excavations and artifact analysis, is gendered. Gero has shown that observation and documentation are not neutral, objective processes but always contain an element of interpretation. Her work on excavations in Argentina show that there are gender differences in documentation, and these are often combined with differences in class and age. In addition, the gender of the person who typically does particular kinds of archaeological work varies. In North America men often lead large fieldwork projects, while women analyze extant collections. In other countries these tendencies take other forms, as in Australia, where men dominate in the universities while women dominate the work of cultural heritage management.

The status of women and equity issues in the archaeological workplace are continuing major concerns of feminists. Although there were some prominent women archaeologists in the past, these had little status and recognition in the profession. There is still a chilly climate for women in archaeology which not only slows down the advancement of female archaeologists but also seems to be preventing the full acceptance of feminist archaeology.

While an important aspect of feminism in archaeology is a concern with the institutional structure and workplace dynamics which hinder women from participating fully in the discipline, feminist archaeologists are also concerned with archaeology's relationship to the communities within which they work. In Australia, Laurajane Smith (1995) has written about the dominance of women in work with cultural heritage, or cultural resource management. Feminists have been particularly open to the importance and meaning that local communities and aboriginal peoples place on their past.

Much standard archaeological methodology is used in feminist archaeology. It is, however, reworked, because feminists ask different questions of the data. Instead of an interest in macroprocesses and population groups, typical, for example, of the systems approach, they ask questions at a microlevel, and are concerned more with social relations and processes, ideologies, and symbolism. Thus methods are combined in alternative manners. For example, emphasis is placed on entire artifact assemblages, and site assemblages, rather than tightly-defined typological categories (where the 'important' and 'interesting' types are often tools presumed to have been used by men), and comparisons over and between large regions. Ambiguous objects, often considered to be merely 'noise' in strict typological terms, are included in analyses since they are one way of analyzing the variety of social relations found at a site. In addition, interest is not placed on type of artifacts alone, but on the relationship between different kinds of artifacts and the processes involved in their production and use. This in turn emphasizes the variety of social relations involved in these processes and de-emphasizes the need to attribute the production and use of tools to an individual or group of the same sex.

5. From Margin to Center

Women have been invisible in archaeological interpretations of the past, but recent work grounded in feminist critiques, principles and perspectives have not only made women visible but have also shown the diversity of their lives and relations to others. Gendered archaeology/feminist archaeology has developed and changed considerably during its short existence. Feminist archaeology is exceptionally dynamic, drawing on new theoretical insights from other feminist disciplines, and expanding the scope and direction of archaeological analyses of gendered relations and ideologies in the past.

In a review article, Conkey and Gero (1997) note the lack of explicit feminist theorizing and critique of basic archaeological concepts in much of gender archaeology. There is little archaeological theorizing on gender/sexuality and there is little use of feminist theory in the archaeology of gender. Detailed empirical analyses are, of course, a necessary complement

to theoretical work, but too much of the archaeology of gender continues within this arena. Feminist archaeology must move beyond this implicit essentialism. The lack of theory only marginalizes the archaeology of gender, not only in relation to mainstream archaeology but also in relation to other feminist disciplines. Future work must continue to include women in our construction of knowledge about the past, but it will also need to grapple with the fundamentals of archaeological interpretation and theorizing, and incorporate feminist theoretical insights on gender and sexuality. Feminist archaeology is dynamic and constantly expanding its areas of interest, and is well on its way to realizing its critical and reconstructive potential.

See also: Gender and Feminist Studies in Anthropology; Gender and Feminist Studies in Economics; Gender and Feminist Studies in Political Science; Gender and Feminist Studies in Psychology; Gender and Feminist Studies in Sociology; Heterosexism and Homophobia; Lesbians: Historical Perspectives; Lesbians: Social and Economic Situation; Male Dominance; Rationality and Feminist Thought

Bibliography

- Arwill-Nordbladh E 1998 *Genuskonstruksjoner i Nordisk vikingetid, förr og nu*. [Gender Constructions in the Nordic Viking Period, Before and Now]. GOTARC, Gothenburg Archaeological Theses, Series B No. 9
- Conkey M 1991 Contexts of action, contexts for power: Material culture and gender in the Magdalenian. In: Gero J, Conkey M 1991 (eds.) *Engendering Archaeology: Women and Prehistory*. Blackwell, Oxford, UK, pp. 57–92
- Conkey M, Gero J 1997 Programme to practice: Gender and feminism in archaeology. *Annual Review of Anthropology* 4: 11–37
- Conkey M, Spector J 1984 Archaeology and the study of gender. *Advances in Archaeological Theory and Method* 7: 1–38
- Diaz-Andreu M, Sørensen, M L S 1998 *Excavating Women: A History of Women in European Archaeology*. Routledge, London, UK
- Gero J 1996 The social world of prehistoric facts: Gender and power in Paleoindian research. In: Wright R (ed.) *Gender and Archaeology*. University of Pennsylvania Press, Philadelphia, PA
- Gero J, Conkey M 1991 *Engendering Archaeology: Women and Prehistory*. Blackwell, Oxford, UK
- Gilchrist R 1999 *Gender and Archaeology*. Routledge, London
- Hager L (ed.) 1997 *Women in Human Evolution*. Routledge, New York
- KAN: *Kvinner i arkeologi i Norge* (KAN: Women in Archaeology in Norway) 1984–present
- Moore J, Scott E 1997 *Invisible People and Processes: Writing Gender and Childhood into European Archaeology*. Leicester University Press, London, UK
- Roberts C 1993 A critical approach to gender as a category of analysis in archaeology. In: du Cros H, Smith L 1993 (eds.) *Women in Archaeology: A Feminist Critique*. Department of Prehistory, Occasional Papers in Prehistory No. 23, Australian National University, Canberra, Australia pp. 16–23

- Smith L 1995 Cultural heritage management and feminist expression in Australian archaeology. *Norwegian Archaeological Review* 28(1): 55–63
- Sørensen M L S 1991 Gender archaeology and Scandinavian Bronze Age studies. *Norwegian Archaeological Review* 25(1): 31–49
- Wylie A 1991 Gender theory and the archaeological record: Why is there no archaeology of gender? In: Gero J, Conkey M (eds.) *Engendering Archaeology: Women and Prehistory*. Blackwell, Oxford, UK, pp. 31–54

E. Englestad

Gender History

Gender is a category of social analysis which denotes the social and cultural, as opposed to natural or biological, relations of the sexes. When this term first gained currency among historians of women in the early 1980s, gender was defined as the social relations between the sexes. As gender became an increasingly important and visible site of scholarly research and debate across the disciplines, its meaning widened to include the symbolic system or signifier of relations of power in which men and women are positioned differently. The embrace of the concept gender as a system of languages, symbols, and social practices that defined sexual difference and created sexual inequality, signaled the departure of feminist social scientists from the unchanging and universal notions of biological sex (Haraway 1991). The connotation of the English word 'gender' is a culturally formed set of social relations, distinct from sex, which is usually understood as rooted in biological or bodily differences. The word gender, however, does not exist in some languages, while in others its meaning diverges significantly from the English usage. The German term *Geschlecht*, for example, encompasses both 'sex' and 'gender' and thus blurs the core distinctions of the English term. In French the term 'sex' or *sexuel* is more commonly used than *genre*, which can refer to either grammatical or literary genre or serve as a classifying category in natural history (Riot-Sarcey 1999). The use of gender as an analytical tool does not *per se* connote a primacy of gender relative to other forms of inequality, such as race, class, or ethnicity; rather it suggests the inextricable links between gender and other social identities and categories of difference.

1. From Women's History to Gender History

The historical study of women, which became a vital field of scholarly inquiry during the 1960s, first sought to recover the ways in which women participated in and were excluded from processes of social transformations and political change. The empirical investi-

gation of the private (familial, sexual) and public (political) inequality of women in past societies drove the first decade of feminist social science. From its inception, women's history had a dual character: its practitioners were historians, trained and employed in history faculties, but the arenas of debate that shaped their inquiry, paradigms and methodologies were most often housed in interdisciplinary programs, such as women's studies.

In this first phase of women's history studies, sex and class often figured as parallel forms of oppression: the female *sex* was viewed as a subordinate *class*, subjugated by a dominant *class* of men. In feminist scholarship the notion of *patriarchy* signaled a shifting understanding of sexual exploitation as the primary form of women's oppression (see, e.g., Rubin 1975, Eisenstein 1978). Taking patriarchy as a hierarchical sexual order, women's historians located its origins in 'the private family' and sought to examine the ways in which it was reproduced in social modes of production, divisions of labor, and property relations (Kelly 1976). The examination of the origins and practices of patriarchal structures and ideologies, and the recovery of the lost stories of female actors and women's agency, significantly expanded the scope of scholarship on women and family, women and labor, social policy and social movements, and elevated 'sex' to a keyword of social analysis. Yet in furnishing a more complete rendering of the contemporary and past societies, the pioneering work of women's history also revealed the limitations of the keyword 'sex.' Unitary analytical categories, such as those of 'woman'/'women' and 'man'/'men,' were based on fixed, rather than historically or socially variable, notions of biological sex, thus creating a cohesive history of the oppression of women that overlooked differences of race, class, ethnicity, and sexual preference. Further, the emphasis on the patriarchal oppression of women by men meant not only that women formed a homogeneous category, but also that they were most often studied in isolation from men, leading to a fragmented and partial understanding of the workings of sexual difference in society (Davis 1976).

As sex became a crucial category of social analysis in the 1970s, its impact on other keywords of social analysis, such as class or race, became a topic of intensive debate among social scientists. Labor historians and political theorists often attempted to fit women into prevalent notions of class, some by arguing that women formed a separate or parallel class of their own, others by examining the ways in which (female) gender presented an obstacle to the formation of cohesive classes. Still others circumvented the dilemma of 'sex and class' by ascribing to women the same class position, identity, and interests as those of their fathers, or husbands. Feminist historian Joan Gadol Kelly sought to transcend this theoretical impasse by formulating a 'doubled vision' of society,

one that emphasized the ways in which both men and women's identities of class were shaped by sex and class (Kelly 1979). In emphasizing the inextricable links between sex and class, women's history studies also indisputably widened the scope of the political to include family and household, bodies and sexualities—all once considered as belonging to the sphere of the 'private.'

Joan Kelly's formulation of this doubled vision of society marks the point at which the concept of gender came into currency as feminist scholarship turned its attention to the ways in which ideologies, norms, and symbolic systems shaped identities, social institutions, and social relations. The editors of the pioneering collection, *Sex and Class in Women's History* (Newton et al. 1983), for example, noted that they employed gender in order to understand the systematic ways in which sex differences cut through society and culture and conferred inequality upon women. This formulation makes it clear that even as 'gender' began to emerge from women's history studies in the early 1980s, 'women' and 'gender' continued to be used as virtually interchangeable terms and scholarship on women continued to provide a foundation for study of gender in most fields (e.g., Hausen and Wunder 1992).

The intellectual forces driving the shift from 'women' to 'gender' were multiple. Despite the achievements of nearly two decades of accumulated scholarship in women's history, the overarching narratives of progress, the chronological and conceptual framings of historical change, remained unshaken. While feminist historians had delivered powerful critiques of the public/private dualism, widening the scope of the political to include family and household, bodies and sexualities, and had successfully dissolved the sex/gender distinction as well, the binary pairings of home/factory, production/reproduction and production/consumption remained tenaciously in place. As feminist historians sought, empirically and theoretically, to overcome or dissolve these dualisms, they began to pose more fundamental challenges to established chronologies, categories and theories of social transformation (Kelly 1984).

The dissatisfaction with biological essentialism as an explanation of sexual inequality, for example, focused new attention on the power of language—of discourses—to construct these inequalities socially and to anchor them in social practices and institutions. The work of analytically distinguishing gender from sex also took place in a more public context, namely in the arenas of debate about 'identity politics' in the USA and Britain, in which once-cohesive ethnic, racial, national or sexual identities were increasingly understood as multiple, mutable, even contradictory. The context for a redefinition of gender became broader as race also came to be viewed as a social construction rather than a genotypic or phenotypic form of difference. At the same time, the unitary category 'woman' began to break down, as historians

of race, particularly in the USA, castigated feminists for the false homogenizing of women and patriarchal oppression. Emphasizing the inextricability of gender and racial identities, feminist scholars of race argued that this category could only be dissolved by bringing race more centrally into the analysis of power (Higginbotham 1992). Feminist scholars of Western empires approached the history of white women's feminism as implicated in imperial ideologies, institutions, and practices. They called upon historians of white women's oppression to confront the racist, nationalist, and imperialist legacies of Western feminism (Amos and Parmar 1984).

2. The Turn to Gender History

Although a shift from women to gender was already well underway by the mid-1980s, the publication of Joan Scott's essay *Gender: A Useful Category of Historical Analysis* (Scott 1986), marks the turning point at which the two terms became distinct in the field of history studies. In the Anglo-Saxon academy the turn to gender as a cultural, linguistic, or discursive construct took place within a broader paradigm shift known as 'the linguistic turn' in the social sciences. The linguistic turn, a catch-all phrase representing a new receptivity of social scientists towards post-structuralist literary criticism, linguistic theory, and philosophy, as well as cultural and symbolic anthropology, signaled a new view of language as constituting historical events and human consciousness rather than as simply reflecting social reality (Toews 1987, Canning 1994). Scott's article, together with her essay collection, *Gender and the Politics of History* (Scott 1988), made explicit the theoretical innovation of gender, summoning 'mainstream' historians to consider gender as an essential category of historical analysis and unleashing a torrent of controversy, not least in the fields of women's studies and women's history.

At the same time Denise Riley (1988) interrogated the category 'woman,' noting its inherent and historically founded instability and pointing toward feminism as precisely the site at which that instability was systematically contested, taken apart and reconstructed. In this sense the turn to gender dissolved some of the stalwart categories of women's history, including both 'woman'/'women' and feminism, with 'experience,' 'identity,' and 'agency,' soon to follow. Furthermore, the breakdown of the category 'woman' and the more explicit formulation of gender as *relational*, as a category and process by which femininity and masculinity are mutually constituted, spurred the opening of gender history towards a more serious study of masculinity.

In the USA and Britain, feminist scholars of race also intervened at this juncture with powerful critiques of the categories and practices of women's history.

Defining race as a 'metalanguage,' as the 'ultimate trope of difference,' African-American critics explicated the powerful effects of race on the construction and representation of gender, class, and sexuality. Understanding race as a 'double-voiced discourse' revealed the ways in which it shaped both the oppression and self-representation of minorities (Higginbotham 1992).

3. *Controversies and Debated in Gender History*

Gender gained an analytic status of its own first by severing the links between sex-gender and biology and then by freeing it from the remnants of Marxism that hinged the production of gender to changing economic structures (Scott 1986). This process necessitated not only the dismantling of familiar concepts, but also the exploration of new methodologies that would illuminate the workings of language and symbols and their dissemination throughout society and culture. Feminist historians also debated the implications of extending gender historical inquiry to masculinity and men. Some of the controversy about these questions took place as an integral part of the debates among social historians on the 'linguistic turn,' while others were more specifically framed by the previous practices and boundaries of women's history. Historians of women feared, for example, that studying gender as a cultural, discursive, or linguistic construction in which masculinity and femininity are relationally and mutually constituted, would render female actors and women's agency invisible and obscure male oppression of women, ultimately producing a gender-neutral historical discourse (Bock 1991). Others suspected that the study of masculinity, while discovering the ways in which ideals of manliness were imposed upon men, might shift attention away from men's participation in those institutions and practices that oppressed women.

3.1 *Gender and the Keywords of Social Analysis*

The turn to gender, like all theoretical and methodological innovations, was an uneven and protracted process, as different fields and national scholarly arenas manifested widely differing responses to this challenge. The extent to which methodologies, concepts, and theories were redefined to meet this challenge also varies across these divergent contexts. In some cases women's history was simply renamed 'gender history,' leading to a protracted phase of interchangeability of the two without significant changes in concepts or methodologies. In other national contexts where the work of recovering the absent female subjects began later, historical research on women continues to provide the foundation for a more explicitly theoretical engagement with gender. In still other settings, where the turn to gender history

coincided with the rediscovery of masculinity and men, gender history denotes a dual engagement with the histories of both masculinity and femininity.

While the study of women once sought to fit female subjects into existing categories of social analysis, the turn to gender prompted more fundamental critiques, expanded meanings, and redefinitions of the keywords of social analysis. The concept of 'class,' for example, was at the heart of debates about both gender and the linguistic turn, leading some feminist scholars to call for an emancipation from class as a 'privileged signifier of social relations and their political representations' (Alexander 1984), while others have analyzed both middle-class (Davidoff and Hall 1987) and working-class formation (Clark 1995) as processes of differentiation in which gender is always centrally implicated. Eschewing models of class formation that seek to link economic, social, cultural, and political levels or stages of development, others have analyzed class as political language, emphasizing the significance of gender in the process of assigning and contesting the contingent meanings of class (Scott 1988, Sewell 1990, Canning 1992).

Other social-scientific terms, like 'public sphere' (Fraser 1992, Ryan 1992), 'civil society' (Pateman 1988, Hull 1996), and more recently, 'citizenship' (Lister 1997) have also been the site of fruitful rethinking by scholars of gender. Similar to the location of class in socioeconomic structures, citizenship is embedded in legal and constitutional frameworks and theories of state and nation formation. Gender scholarship on citizenship has critiqued models of citizenship as a progressive process of acquisition of civil, political and social rights, as outlined, for example, in T. H. Marshall's classic, *Citizenship and Social Class* (1950). New considerations of the gender dimensions of this term have analyzed citizenship as a process, as a set of claims upon the state and public sphere rather than as a historically fixed outcome or legal status. The historical meanings of citizenship can thus be understood to encompass not only those who possess citizenship rights, but also those excluded from them.

3.2 *Gender, Sexuality, Body*

Other key terms in the vocabulary of interdisciplinary gender studies, which were redefined in the course of the paradigm shift from women to gender, are those of 'sexuality' and 'body.' In addition to critiquing and redefining keywords of social and historical analysis, such as class and citizenship, the turn to gender also generated new waves of scholarship on topics which had been submerged within the complex of women and sex and which came into their own once gender had been analytically distinguished from sexuality and body.

Influenced by the work of Michel Foucault, historians sought to analyze the historically specific

domains in which sexualities were produced, such as the systems of medical knowledge and state moral regulation that defined both normal and deviant sexuality in nineteenth-century Europe, dispersing these definitions throughout the institutions of state and social reform (Mort 1987). In this sense, it is possible to view sexuality as a more specific site of definition, intervention and experience than gender. A new emphasis on the importance of desire in the articulation of sexuality suggested that sexuality might be understood as a site of agency which was quite distinct from gender. Yet Judith Butler (1990) posited gender as lived, at least in part, through sexuality, as embodied and performed through the production and ascription of sexualities. In recent years, studies of homosexualities have raised new questions about the implications of transgressive sexualities for the study of gender, suggesting that they disrupt gender categories by breaking up the neat opposition of male/female, masculine/feminine and by unmasking 'the epistemic regime of presumptive heterosexuality' (Butler 1990, Puff 1998).

'Body,' the term in the vocabulary of gender that was most deeply embedded in biologism and essentialism, was inextricably linked to sex before the 'linguistic turn' began to disentangle these terms. During the first wave of women's studies, the body had been the foundation for unities and shared experiences of women. The 'discursivation' of the body, which took place in the course of the 'linguistic turn,' not only detached the body from unchanging notions of physical difference, but also examined bodies as sites of inscription—as instruments of changing cultural meanings. While the advent of the discursive body delivered the final blow to biological determinism, it also raised new questions about the particular embodiedness or bodily mediation of female experiences that remained central in feminist historical research (e.g., Duden 1991). In the debates of the early 1990s about the relationship between discourse and experience, for example, the body became a litmus test of this relationship (Butler 1993, Bynum 1995).

In the course of the 1990s, then, body history became a prolific offshoot of gender history. Both theoretical debate and empirical research in the field have shifted recent discussions towards exploration of the process of *embodiment*, which takes bodies as sites of mediation between constructions of the body and its materialized experiences (pain, birth, death, illness, dismemberment) and suggests new frameworks for examining the ways in which bodies are lived, experienced and acted upon by the subject and the social collectivity (Grosz 1994, Gatens 1996).

While the turn to the concept of gender spawned new waves of historical scholarship on sexualities and bodies, gender history later expanded the scope of its inquiry to two new fields of historical inquiry: the history of consumption and history of empires. The study of consumption in Western societies was in-

vigorated in part by the collapse of communism in Russia and Eastern Europe, which made clear the ways in which consumption, as well as production, structure capitalist societies. Although consumption has historically been identified with femininity, Davidoff and Hall's *Family Fortunes* (1987) argued convincingly that domesticity, the ideology of the British middle classes, was vested in complex and elaborate practices of consumption and display. With gender as a framing concept, the historical study of consumption has expanded in recent years to encompass analysis of a social and cultural formation—consumer society—through the lens of commodification, spectatorship, commercial exchange and welfare reforms, each of which can be analyzed in conjunction with the gender identities produced through these processes (DeGrazia and Furlough 1996). Throughout the 1990s gender and sexuality also had a central place in the history of empire and figured centrally in new understandings of nations, empires, and imperial social formation.

4. Current Issues and Future Directions

This brief survey of gender historical scholarship indicates that the concept of gender has successfully broadened and deepened the meanings of some of the keywords of historical analysis, forged new arenas of scholarly inquiry, and helped to refurbish epistemologies and methodologies of historical analysis. Gender history has moved from the periphery once occupied by the 'herstory' of early women's history to the center in some historical fields, while in others it remains distant from the problems that define the field. The historical narrative of the French Revolution has been significantly transformed by gender scholarship, which went far beyond analyzing the exclusion of women from revolutionary citizenship during Jacobin rule to revise broader understandings of the crisis of state (e.g., Lynn Hunt's notion of 'family romance'), the emergence of the body politic, civil society, and notions of citizenship. The history of labor and labor movements, began to diminish in the late 1980s after two decades of field-defining scholarship, only to be significantly revived and recast through the wave of gender and labor scholarship of the mid- to late 1980s (Frader and Rose 1996). The histories of slavery in the antebellum South, of colonialism, empires, and post-emancipation societies have been rewritten as scholars explore the ways in which gender inequalities underpinned racial and imperial dominance. Gender has become a vital part of new scholarship on the two world wars, including military history. Yet other fields, such as the histories of the Weimar Republic or the Third Reich, for example, remain immune to the impulses of gender, despite an impressive accumulation of research on both women and gender for both

periods. Few would dispute that contests over masculinities and femininities are a crucial part of the post-World War I crisis of 'bourgeois Europe.' Yet in the historiography of Weimar Germany gender is meaningful mainly in those realms remote from high politics—realms as diverse as popular culture and sexual reform organizations. Nor has gender figured in the dramas of state, democracy, or the fragile social body of the 1920s. Recent scholarship on the Third Reich offers abundant evidence that the Nazis pursued, relentlessly and violently, a gender project, which was at the heart of the racial state they envisioned and built. But this knowledge does not significantly inform interpretations of the collapse of the Weimar Republic or the Nazis' mobilization of German society and economy for war and genocide. This uneven reception of gender history in various fields will undoubtedly shape the dynamic of gender history in years to come.

The relationship of gender history to 'mainstream' history has been of growing concern to those who formulated its emergence from women's history in the early 1980s. Joan Kelly's three-pronged project of renovating categories, chronologies, and theories of social transformation explicitly aimed at rewriting historical metanarratives, as she sought to do in her own essays on the Renaissance. Given the accomplishments of feminist history in recasting categories of historical analysis, it is now important to assess the potential of gender history to reconstruct historical narratives and to produce new syntheses of historical events and processes (e.g., Hunt 1998). Several dilemmas remain regarding gender history and the historical metanarratives it set out to subvert: should, for example, the empirical and conceptual achievements of gender history be assimilated into existing metanarratives? Should they inspire the composition of new metanarratives or prompt the abandonment of metanarratives altogether? These questions can perhaps be resolved first by breaking these metanarratives into their composite parts and examining the singular moments of subversion and incorporation along the narrative thread. Either task—that of revising existing narratives or that of composing new ones—requires a certain kind of piecework in which historical changes are rearranged on time lines and brought into new relationship with one another.

Attention to the temporalities of gender history, and to how they fit or diverge from those of established metanarratives, has only recently become possible. The genealogy of gender might be traced back to the eighteenth century, when humanist notions of the individual were synonymous with masculinity. The embodiment and masculinization of the male citizen during the French Revolution also provided the stage upon which women staked their first claims for inclusion in the terms of both equality and difference. The extensive scholarship on women and gender in nineteenth-century Europe encompasses the history of a wide range of public feminisms and domestic

ideologies, the emergence of 'women' as social category, of 'women's labor' as a social question, and the dissemination of these ideologies and social categorizations, which redefined both femininity and masculinity, through the media of the emergent social and natural sciences. This suggests (a) that the history of nineteenth-century Europe could now be rewritten from the perspective of gender, and (b) that the temporality of gender history may not follow the political or social chronologies of existent metanarratives. A rethinking of twentieth-century history may uncover distinct temporalities of a different gender history, one in which gender norms and ideologies were ruptured at points of political and cultural crisis but not necessarily restored at moments of political reconstruction and stabilization. This trajectory of twentieth-century gender history may indeed help to explain its extraordinarily embodied violence. The work of aligning the temporalities of gender history with those inscribed in established historical metanarratives represents the first step towards a history that admits distinct temporalities and allows for dissonance within grand narratives.

See also: Body, History of; Family and Kinship, History of; Gender and Feminist Studies; Gender and Feminist Studies in Anthropology; Gender and Feminist Studies in History; Gender and Feminist Studies in Sociology; Gender Differences in Personality and Social Behavior; Gender Ideology: Cross-cultural Aspects; Historiography and Historical Thought; Current Trends; Household Production; Linguistic Turn and Discourse Analysis in History; Social History

Bibliography

- Alexander S 1984 Women, class and sexual differences in the 1830s and 1840s: Some reflections on the writing of a feminist history. *History Workshop* 17: 125–14
- Amos V, Parmar P 1984 Challenging imperial feminism. *Feminist Review* 17: 3–19
- Bock B 1991 Challenging dichotomies: Perspectives on women's history. In: Offen K, Pierson R R, Rendall J. *Writing Women's History*. Indiana University Press, Bloomington, IN
- Butler J 1990 *Gender Trouble. Feminism and the Subversion of Identity*. Routledge, New York/London
- Butler J 1993 *Bodies That Matter. On the Discursive Limits of 'Sex.'* Routledge, New York/London
- Bynum C 1995 Why all the fuss about the body? A medievalist's perspective. *Critical Inquiry* 22: 1–33
- Canning K 1992 Gender and the politics of class formation: Rethinking German labor history. *American History Review* 97(3): 736–68
- Canning K 1994 Feminist history after the linguistic turn: Historicizing discourse and experience. *Signs* 19(2): 368–404
- Clark A 1995 *The Struggle for the Breeches. Gender and the Making of the British Working Class*. University of California Press, Berkeley, CA

- Davidoff L, Hall C 1987 *Family Fortunes: Men and Women of the English Middle Class, 1780–1950*. University of Chicago Press, Chicago
- Davis N 1976 Women's history in transition: The European case. *Feminist Studies* 3(3/4): 83–103
- DeGrazia V, Furlough E (eds.) 1996 *The Sex of Things: Gender and Consumption in Historical Perspective*. University of California Press, Berkeley, CA
- Duden B 1991 *Woman Beneath the Skin: A Doctor's Patients in 18th Century Germany*. [Transl. Dunlap T] Harvard University Press, Cambridge, MA
- Eisenstein Z 1978 *Capitalist Patriarchy and the Case for Socialist Feminism*. Monthly Review Press, New York
- Frader L, Rose S (eds.) 1996 *Gender and Class in Modern Europe*. Cornell University Press, Ithaca, New York
- Fraser N 1992 Rethinking the public sphere: A contribution to the critique of actually existing democracy. In: Calhoun C (ed.) *Habermas and the Public Sphere*. MIT Press, Cambridge, MA
- Gatens M 1996 *Imaginary Bodies: Ethics, Power and Corporeality*. Routledge, London
- Grosz E 1994 *Volatile Bodies: Toward a Corporeal Feminism*. University of Indiana Press, Bloomington, IN
- Haraway D 1991 *Simians, Cyborgs, and Women: The Reinvention of Nature*. Routledge, New York/London
- Hausen K, Wunder H 1992 *Frauengeschichte/Geschlechtergeschichte*. Campus Verlag, Frankfurt/New York
- Higginbotham E B 1992 African-American women's history and the metalanguage of race. *Signs* 17(2): 251–74
- Hull I 1996 *Sexuality, State, and Civil Society in Germany, 1700–1815*. Cornell University Press, Ithaca, New York
- Hunt L 1998 The challenge of gender. Deconstruction of categories and reconstruction of narratives in gender history. In: Medick H, Trepp A-C (eds.) *Geschlechtergeschichte und Allgemeine Geschichte, Herausforderung und Perspektiven*. Wallstein, Göttingen, Germany
- Kelly J 1976 The social relation of the sexes: Methodological implications of women's history. *Signs* 1/4: 809–23
- Kelly J 1979 The doubled vision of feminist theory. *Feminist Studies* 5(1): 216–27
- Kelly J 1984 *Women, History, and Theory: The Essays of Joan Kelly*. University of Chicago Press, Chicago
- Lister R 1997 *Citizenship: Feminist Perspectives*. New York University Press, New York
- Mort F 1987 *Dangerous Sexualities: Medico-Moral Politics in England Since 1830*. Routledge, New York
- Newton J, Ryan M, Walkowitz J (eds.) 1983 *Sex and Class in Women's History*. Routledge, New York
- Pateman C 1988 *The Sexual Contract*. Stanford University Press, Stanford, CA
- Puff H 1998 *Männergeschichten/Frauengeschichten*. Über den Nutzen einer Geschichte der Homosexualitäten. In: Medick H, Trepp A-C (eds.) *Geschlechtergeschichte und Allgemeine Geschichte, Herausforderung und Perspektiven*. Wallstein, Göttingen, Germany
- Riley D 1988 *Am I That Name? Feminism and the Category of 'Women' in History*. University of Minnesota, Minneapolis, MN
- Riot-Sarcey M 1999 The difficulties of gender in France: Reflections on a concept. *Gender & History* 11(3): 489–98
- Ross E, Rapp R 1983 Sex and society: A research note from social history and anthropology. In: Snitow A, Starnell C, Thompson S (eds.) *Powers of Desire*. Monthly Review Press, New York
- Rubin G 1975 The traffic in women: Notes on the 'political economy' of sex. In: Reiter R (ed.) *Toward an Anthropology of Women*. Monthly Review Press, New York
- Ryan M 1992 Gender and public access: Women's politics in nineteenth-century America. In: Calhoun C (ed.) *Habermas and the Public Sphere*. MIT Press, Cambridge, MA
- Scott J 1986 Gender: A useful category of historical analysis. *American History Review* 91/5: 1053–75
- Scott J 1988 *Gender and the Politics of History*. Columbia University Press, New York
- Sewell W H Jr 1990 How classes are made: Critical reflections on E P Thompson's theory of working-class formation. In: Kaye H, McClelland K (eds.) *E. P. Thompson: Critical Perspectives*. Temple University Press, Philadelphia, PA
- Toews J 1987 Intellectual history after the linguistic turn: The autonomy of meaning and the irreducibility of experience. *American History Review* 92(4): 879–907

K. Canning

Gender Identity Disorders

1. Definitions, Prevalence, and Sex Ratio

The gender identity disorders are defined as disorders in which an individual exhibits marked and persistent identification with the opposite sex and persistent discomfort (dysphoria) with their own sex or sense of inappropriateness in the gender role of that sex. These characteristics may be manifest in statements such as wishing to be the other sex, dressing as, and taking on the roles of the other sex, engaging preferentially in activities usually enjoyed by the other sex, and expressing dislike for the roles and activities of their own sex, as well as a dislike for their own genitalia and other physical characteristics of their own sex. In order to distinguish the gender identity disorders from more minor cross-gender interests or role-taking these characteristics must be persistent and marked and result in distress or impairment (*Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition, (*DSM-IV*), 1994; *International Classification of Diseases*, Tenth Edition, (*ICD-10*), 1992).

The word transsexual was used in the earlier literature, as was inversion, to refer to individuals with marked gender dysphoria, most of whom perceived sex reassignment as the solution to their psychological distress. It remains as a category in the *ICD-10* classification. Also in the older literature transvestite was used to designate persons who cross-dressed and may have included those who identified as the other sex as well as those in whom the cross-dressing was sexually arousing (now referred to as Transvestic Fetishism). Transgendered is a more recent term used to encompass a broad group of individuals with varying degrees of cross-gender behavior. It is not a diagnosis and is often the self-label assumed by individuals with cross-gender feelings and behaviors, particularly the more politically active. Gender role

refers to behaviors, roles, and attitudes that the specific culture defines as more appropriate to males or females. Sexual orientation refers to the sex of person to whom an individual is attracted, predominantly the opposite (heterosexual), or the same (homosexual) sex. Sexual identity connotes the adoption of an identity as a hetero- or homosexual person and involves more than the narrow notion of sexual orientation.

Gender identity disorders (GID) are relatively rare compared to other psychiatric disorders. Although there are no large scale epidemiological studies to provide true estimates recent studies suggest roughly 1:10,000 to 1:30,000. Sex ratios of adults with GID (largely based on referrals to clinics) have fluctuated with more males than females in earlier studies to a more equal ratio in many recent reports. Childhood GID is more prevalent in males, roughly 6 to 1. In adolescence the ratio is more equal (Cohen-Kettenis and Gooren 1999, provide a recent review.)

2. Historical Background

Awareness of individuals who identified strongly with the other sex has a long history. Some cultures even created a privileged role for such individuals. However, the scientific study of individuals with cross-gender identities really began in the late part of the nineteenth century. Initially many such individuals were thought to be intersex individuals, possessing the physical characteristics (both internal and external) of the other sex. It was not until the middle of the twentieth century that it became clear that most such individuals were physically normal but suffering from a psychological disorder.

The modern era of gender identity disorder began with the much publicized case of Christine Jorgensen, an American ex-soldier, who underwent sex-change surgery in Denmark in 1952, under the care of Christian Hamburger, an endocrinologist. Although demasculinizing surgery and hormonal replacement had been used earlier, typically for cases thought to be intersex, the Jorgensen case stirred the public imagination and encouraged other like-minded individuals to seek treatment for their distress (Meyerowitz 1998). Sex change surgery had been available in Europe since the early 1900s, the most prominent center being the Institute for Sexual Science in Berlin under the direction of Magnus Hirschfeld. In the USA physicians were less willing to perform such surgery, particularly on those considered to be physically normal and in some jurisdictions such body altering was considered illegal. Prior to the Jorgensen case, David Cauldwell, an American psychiatrist and editor of *Sexology*, had distinguished cases which he labeled '*psychopathia transexualis*' from intersex cases. Harry Benjamin, an endocrinologist, born and trained in Germany, had begun to advocate on behalf

of individuals wishing sex change. This combination of public and professional interest set the stage for the scientific study of persons with gender identity disorders.

In 1955 John Money working with the Hampsons published his first follow-up study of individuals born with ambiguous genitalia. His conclusions that sex of rearing overrides the child's anatomy in determining gender identity was the first empirical evidence that psychological factors may be etiologically important in GID. The 1960s saw the publication of several important books and papers as clinicians working with individuals with GID began to describe their experience. Harry Benjamin, who had been diligent in his effort to bring understanding to the plight of those whose cross-gender wishes were often demeaned or dismissed, published *The Transsexual Phenomenon* in 1966. In the same year the first gender clinic in North America was established at Johns Hopkins Hospital in Baltimore. Green and Money (1969) co-edited the book *Transsexualism and Sex Reassignment*, a standard text at that time. A year earlier, Stoller (1968), a psychoanalyst, had published *Sex and Gender*, in which he articulated a theory that cross-gender identification arises from a blissful symbiosis of a son with his mother. Although Stoller's theoretical position has more recently been dismissed it sparked interest in understanding the psychodynamic origins of GID and led to more systematic study in several of the newly developed clinics for children and adolescents.

The 1970s and 1980s saw a growth of interest in treatment of adults with GID. Clinics developed in major centers in North America and Europe. Standards of Care were developed under the auspices of the Harry Benjamin International Gender Dysphoria Association in 1985 and provided some uniformity to patient management. Clinics for children were developed as clinicians realized that GID had its origins in childhood and might be prevented with early intervention. Green (1974) carried out the first systematic study of effeminate boys, published in *Sexual Identity Conflict in Children and Adults*. He compared boys displaying significant cross-gender behavior by parent report, with an equal number of age-matched masculine boys. His study provided the first systematic description of boys with GID and their parents. Green (1987) published the follow-up of these two groups of boys. The majority of the effeminate boys had a homosexual orientation when assessed in late adolescence. Only one boy continued to have significant cross-gender wishes. These findings were in marked contrast to the findings on the masculine boys who were almost entirely heterosexual in orientation and displayed no evidence of gender dysphoria.

The study of individuals with GID soon became a subject of religious, moral, and political debate. Some investigators urged treatment of GID to prevent homosexuality. In the 1990s funders of GID programs

began to debate the logic of a medical and surgical intervention for what is primarily a psychological disorder. This debate was fueled by the results of the first follow-up study of the patients from the John Hopkins clinic which suggested that surgically re-assigned patients were no better off than were those who had not received surgery. Although this study was methodologically flawed, its results, together with economic pressures, led to the closure of many clinics in the USA and the removal of sex reassignment surgery as an insured service. Generally the situation in Europe, particularly in the Netherlands, remained more supportive for those with GID. In the late 1980s the gay and lesbian community began to question the pathologizing of cross-gender behavior in childhood, regarding it simply as an early manifestation of homosexuality. In the early 1990s many individuals with cross-gender identification began to organize politically, assuming the label transgendered, and arguing that society should recognize a third or intermediate position between the more traditional male and female genders.

3. *Emphases in Current Theory and Research*

3.1 *Classification and Phenomenology*

Early work focused on dividing individuals requesting sex reassignment into 'true' and 'secondary' transsexuals. This effort was abandoned as GID began to be conceptualized as being more dimensional than purely categorical. However, there was a recognition that there are two types of GID in males, one originating from early or childhood GID and being predominantly homosexual in sexual orientation, and the other, arising from Transvestic Fetishism, and being hetero- or bisexual in orientation. Individuals with the latter present most typically in their thirties in contrast to the former who present in adolescence or early adulthood. Although usually described as bi- or heterosexual, Blanchard (1991) has established that many of the latter group display what he calls autogynephilia, an erotic attraction towards oneself dressed as a female. Mature females with GID are essentially homosexual in their erotic attraction.

3.2 *Etiology*

As mentioned above, the first theorizing about etiology arose from the belief that individuals with GID are in some way intermediate in their biology between males and females. Because of the link between cross-gender identification and homosexuality and because of the relatively small numbers of individuals with GID this led to extensive investigation of the biological differences between heterosexual males and females

and between these two groups and their homosexual counterparts.

Although many of these studies have provided interesting leads very few differences have been substantiated in a way that would suggest strong evidence for a significant predictor of gender pathology or of homosexuality. Rather the findings to date suggest that temperamental or other vulnerability factors may create a biological predisposition towards gender pathology which when combined with the appropriate psychodynamic factors produce gender dysphoria and cross-gender identification.

3.3 *Biological Studies*

Genetic factors are assumed to play a role in GID although there have been no large-scale studies of heritability in GID. This assumption is based on the studies of heritability in homosexuality and its putative relationship with GID. In monozygotic twins there is a greater likelihood of both twins being homosexual as compared to dizygotic twins (Bailey et al. 2000). This relationship has been shown most strongly in males. Despite a concordance rate as high as 50 percent there appears to be a strong nonshared environmental component (due to factors unique to the experience of the individual). Similarly, molecular genetics studies of homosexuality in men suggest the possibility of a marker specific to sexual orientation on the distal portion of the long arm of the X chromosome (Hamer et al. 1993) although the only other independent study failed to replicate this finding.

The prenatal hormone theory has provided the most coherent framework for an examination of factors affecting sex-dimorphic behavior in both animals and humans. This theory, now well supported by studies in animals, posits that the prenatal hormonal milieu affects the development of brain structures and organizes or shapes various sex-dimorphic behaviors, such as activity level, mounting behavior, and responsiveness to sexual partners. In humans, Congenital Adrenal Hyperplasia (CAH), a disorder in which females are exposed to high levels of circulating androgen *in utero*, provides a convenient natural experiment which gives some support to the theory. Girls with CAH have been shown to be more physically active and tomboyish than their nonaffected sisters and to report more homosexual attraction and less heterosexual involvement in adulthood. They do not, however, report more wishes to be the other sex. This latter finding may be partially a result of the small sample sizes in all studies, too small to detect a disorder such as GID, which is relatively rare.

Another approach to testing the prenatal hormone theory has been to examine hormone responses to a dose of exogenously administered estrogen. In the female such an injection produces an initial fall and then a rise in luteinizing hormone. Males do not

display the rise seen in females. Homosexual men show a pattern intermediate between that of heterosexual men and women. These studies have not been replicated consistently and have been criticized in terms of not being interpretable, as the endocrine systems of homosexual men may be compromised by such things as HIV infection.

Prenatal maternal stress is another factor thought to possibly influence the development of homosexuality (and by extension GID) through affecting the prenatal hormonal milieu and brain development. Early studies in rats showed that stressing pregnant females can produce a demasculinizing and feminizing effect on their male offspring. Although attempts to test this out on humans have been equivocal, Bailey et al. (1991) found that maternal 'stress proneness' was correlated with boyhood effeminacy. As there is an emerging literature that suggests that maternal stress may correlate with vulnerability to psychiatric disorder in offspring by affecting stress reactivity, these findings may be relevant in terms of vulnerability to psychopathology, as opposed specifically to homosexuality or GID (Bradley 2000).

Efforts to define and compare the brain areas responsible for sexually dimorphic behavior have been a part of testing the prenatal hormone theory. Once more this work has been carried out on samples of homosexual individuals, largely male. Although brain structures in rats appear more clearly sex-dimorphic, agreement about the comparable brain structures in humans has been more complicated. Swaab and Fliers (1985) reported differences between males and females in the preoptic area of the anterior hypothalamus but in a subsequent study could not define differences in volume between heterosexual and homosexual men. They did, however, find a difference in the suprachiasmatic area of the hypothalamus between homosexual and heterosexual men, an area previously not related to sex-dimorphic behavior (Swaab and Hofman 1990). Another cell group identified as possibly homologous with the sexually dimorphic area in lower mammals is the interstitial nucleus of the anterior hypothalamus, subdivided into four sections, INAH 1-4 (Allen et al. 1989). These authors found differences in size between males and females in INAH 2 and 3. LeVay (1991) subsequently reported a smaller INAH 3 in homosexual men (comparable to that of heterosexual women) in comparison to heterosexual men. This study has recently been partially replicated by Byne et al. (in press). Zhou et al. (1995) reported that the bed nucleus of the stria terminalis was smaller in six male-to-female transsexuals than in control males and was comparable to that of females.

Another area of potential difference between homosexual and heterosexual individuals is cerebral laterality. This has been examined through assessment of handedness, cognitive testing, dermatoglyphic asymmetry, and related brain volumes. Some studies have

suggested that homosexual men and women have a higher likelihood of being left-handed than heterosexual men and that transsexuals (both male and female) display more left-handedness than the general population (Lalumière et al., 2000). A preliminary study on boys with GID from the Toronto clinic examining handedness suggests a higher incidence of left-handedness (Zucker et al. 2000). Neuropsychological studies using visual hemifield dot detection and dichotic listening tasks suggest differences between homosexual and heterosexual men with the performance of homosexual men being more similar to that of heterosexual women. Studies of transsexual subjects also suggested less lateralization but again these results are preliminary. Study of fingerprint patterns, dermatoglyphy, shows sex differences with women having more ridges on the left than right hand and homosexual men showing a similar left-ward asymmetry. The anterior commissure, a fiber tract connecting the two cerebral hemispheres, is larger in women than men and has been shown to be larger in homosexual than heterosexual men. Although interpretation of these findings should remain open they suggest that early brain development and with it handedness and other indices of laterality may be affected in individuals who later become homosexual and possibly in those who develop GID.

Temperamental variables, specifically activity level and involvement in rough-and-tumble play, have been shown to differ in boys and girls. Boys with GID display a lower activity level and less interest in rough-and-tumble play than control boys. In contrast, girls with GID are reported to have a higher activity level and to be more interested in rough-and-tumble play than control girls. Although these findings suggest differences between GID children and controls, it is highly unlikely that these differences are more than predispositions that add a dimension to their general vulnerability to GID. Physical attractiveness is another factor that may act as a predisposing factor but is unlikely to be more specific to GID. Boys with GID are rated by independent raters as more attractive, beautiful, cute, handsome, and pretty than control boys. Conversely, girls with GID are rated as less attractive than control girls. Although physical attractiveness may be a social construction these findings are consistent with parental reports of their GID sons being beautiful babies and may be related to greater facial symmetry, a component of attractiveness.

Particularly feminine homosexual men and boys with GID have a greater number of brothers than controls. Being born later in the birth order also occurs more frequently in homosexual men (regardless of extent of cross-gender behavior) than in heterosexual controls. These findings have been interpreted from both biological and psychological frameworks with no evidence to suggest that either explanation is more valid.

Lastly, birthweight appears to be lower in boys with GID than in controls. Again this may reflect a neurodevelopmental factor which is related to a general vulnerability to psychopathology as low birth weight has also been found as a vulnerability factor to psychotic disorders.

3.4 Psychosocial Factors

As indicated in Sect. 2 above, Money et al. (1955) reported on their follow-up of children born with ambiguous genitalia. They concluded that sex of assignment and rearing, particularly when that was begun unambiguously prior to 24 months of age, was able to override a child's anatomy. These early studies by Money et al. led to the belief that gender identity was developed during a sensitive period and became relatively fixed at an early age. Despite challenges, this position has not been significantly refuted. Questions were raised by the finding that individuals with 5-alpha-reductase deficiency, apparently raised as girls, for the most part begin to self-identify as males at puberty when their bodies begin to masculinize. This and other similar findings appear less to have refuted Money et al.'s findings than to have suggested that in some instances gender identity may be less fixed than had originally been proposed. This same malleability of gender identity appears to be relevant in the case of males with Transvestic Fetishism who although originally identified as male begin, over a number of years, to self-identify as female. Other anomalous situations include males with accidental ablation of the penis at a young age. One such case has been used to argue that biology overrides rearing while another suggests the opposite. Generally, however, individuals with GID are physically normal and so psychosocial factors appear to be particularly relevant.

Parents have been shown to respond differently to male and female children, generally encouraging same-sex behavior more than opposite-sex behavior. Parents of children with GID tend to be neutral about if not encouraging of their child's cross-gender behavior. Further they appear to be less encouraging of same-sex behavior. Despite Stoller's premise of a 'blissful symbiosis' between a boy with GID and his mother, Green (1974) found that both parents spent significantly less time with their son with GID than did parents of controls. Further, both mothers and fathers of children with GID have elevated levels of psychopathology, which would suggest impairments in their relationships with each other and with their children. Lastly, clinical observations suggest high levels of conflict between children with GID and their parents, which these sensitive children interpret as rejection.

Etiological factors are reviewed more thoroughly in Zucker and Bradley (1995) where a complete set of references is also available. Clinical formulations of

these etiological factors are available in Coates (1990) and Zucker and Bradley (1995).

4. Management

There are few formal studies assessing management of children or adults with GID. Current management of children with GID involves addressing the familial and child issues through different combinations of parent counseling, parent groups, and child psychotherapy.

Adolescents with GID may be referred for psychotherapy to rule out factors that if addressed might change the wish for sex reassignment and to ensure the stability of the cross-gender wish. Some clinics use hormonal treatment to reduce the development of secondary sex characteristics and reduce sex drive when such adolescents have been considered treatment eligible (Cohen-Kettenis and Gooren 1999). Psychological support is considered very important but often rejected by the adolescent.

Management of the adult with GID is somewhat similar to that of the adolescent but can proceed to surgical reassignment after the individual has demonstrated their capacity to surmount the hurdles of living as the other sex for a period of time (1–2 years). Once more, psychological support is considered critical but is not always welcomed by the individual with GID. Hormonal therapy (estrogens to male-to-females and androgens to female-to-males) are begun in some clinics at the beginning of the real-life test and in other clinics somewhat later. Sex reassignment surgery typically is not undertaken until sometime in early adulthood although those individuals with a history of Transvestic Fetishism may not have surgery until their thirties or forties.

5. Outcome

Children with GID can relinquish their desire to be the opposite sex. This appears to be facilitated when parents address some of the familial issues thought to contribute to the child's gender dysphoria. Children presenting at a later age (10–12) or whose families cannot address the contributing familial issues appear to do less well and continue to display significant gender dysphoria. Adults who undergo sex reassignment are generally satisfied although the capacity of surgery to produce well-functioning genitals, especially in the female-to-male, is quite limited.

6. Methodological Issues

Given the relative rarity of this condition sample sizes adequate to test hypotheses are only available in one or two clinics in the world. Evaluating treatment is similarly complicated by small sample sizes in any one

geographic location. Multisite studies are needed to address some of these issues.

7. Future Directions

Comparison of children with GID vs. contrast groups such as children with anxiety disorders are needed to address general (psychopathology) vs. specific (to GID) etiological factors. Evaluation studies of interventions in children, adolescents, and adults are needed.

See also: Body Image and Gender; Gender, Class, Race, and Ethnicity, Social Construction of; Gender Differences in Personality and Social Behavior; Gender-related Development; Identity in Childhood and Adolescence; Sexuality and Gender

Bibliography

- Allen L S, Hines M, Shryne J E, Gorski R A 1989 Two sexually dimorphic cell groups in the human brain. *Journal of Neuroscience* **9**: 497–506
- Bailey J M, Dunne M P, Martin N G 2000 Genetic and environmental influences on sexual orientation and its correlates in an Australian twin sample. *Journal of Personality and Social Psychology* **78**: 524–36
- Bailey J M, Willerman L, Parks C 1991 A test of the maternal stress theory of human male homosexuality. *Archives of Sexual Behavior* **20**: 277–93
- Benjamin H 1966 *The Transsexual Phenomenon*. Julian Press, New York
- Blanchard R 1991 Clinical observations and systematic studies of autogynephilia. *Journal of Sex and Marital Therapy* **17**: 235–51
- Bradley S J 2000 *Affect Regulation and the Development of Psychopathology*. Guilford, New York
- Byne W, Tobet S, Mattiace L A, Lasco M S, Kemether E, Edgar M A, Morgello S, Buchsbaum M S, Jones L B (in press) The interstitial nuclei of the human anterior hypothalamus: An investigation of variation with sex, sexual orientation and HIV status. *Hormones and Behavior*
- Coates S 1990 Ontogenesis of boyhood gender identity disorder. *Journal of the American Academy of Psychoanalysis* **18**: 414–38
- Cohen-Kettenis P T, Gooren L J G 1999 Transsexualism: A review of etiology, diagnosis and treatment. *Journal of Psychosomatic Research* **46**: 315–33
- Green R 1974 *Sexual Identity Conflict in Children and Adults*. Basic Books, New York
- Green R 1987 *The 'Sissy Boy Syndrome' and the Development of Homosexuality*. Yale University Press, New Haven, CT
- Green R, Money J (eds.) 1969 *Transsexualism and Sex Reassignment*. Johns Hopkins Press, Baltimore, MD
- Hamer D H, Hu S, Magnuson V L, Hu N, Pattatucci A M L 1993 A linkage between DNA markers on the X chromosome and male sexual orientation. *Science* **262**: 321–27
- Lalumiere M L, Blanchard R, Zucker K J 2000 Sexual orientation and handedness in men and women: A meta-analysis. *Psychological Bulletin* **126**: 575–92
- LeVay S 1991 A difference in hypothalamic structure between heterosexual and homosexual men. *Science* **253**: 1034–7
- Meyerowitz J 1998 Sex change and the popular press: Historical

notes on transsexuality in the United States, 1930–1955. *Gay and Lesbian Quarterly* **4**: 159–87

- Money J, Hampson J G, Hampson J L 1955 An examination of some basic sexual concepts: The evidence of human hermaphroditism. *Bulletin of the Johns Hopkins Hospital* **97**: 301–19
- Stoller R 1968 *Sex and Gender*. Science House, New York
- Swaab D F, Fliers E 1985 A sexually dimorphic nucleus in the human brain. *Science* **228**: 1112–5
- Swaab D F, Hofman M 1990 An enlarged suprachiasmatic nucleus in homosexual men. *Brain Research* **537**: 141–8
- Zhou J N, Hofman M A, Gooren L J G, Swaab D F 1995 A sex difference in the human brain and its relation to transsexuality. *Nature* **378**: 68–70
- Zucker K J, Beaulieu N, Bradley S J, Grimshaw G M, Wilcox A 2000 Handedness in boys with gender identity disorder. *Journal of Psychology and Psychiatry*
- Zucker K J, Bradley S J 1995 *Gender Identity Disorder and Psychosocial Problems in Children and Adolescents*. Guilford, New York

S. J. Bradley and K. J. Zucker

Gender Ideology: Cross-cultural Aspects

Gender ideology is concerned with normative beliefs about the proper roles for and fundamental natures of women and men in human societies. The distinction between sex and gender is central to the concept of gender ideology. When a child is born, the biological differences between males and females, or their sex differences, are used as the basis for the assignment of gender and the cultural construction of gender identities. People are assigned a gender identity in all societies, but gender systems and the gender ideologies that are thought to help sustain them are culturally variable.

1. The Historical Context for Studies of Gender Ideology

The Women's Liberation Movement of the late 1960s and 1970s, which started in the United States and then spread to Europe and other parts of the world, was an important stimulus for cross-cultural research on gender ideologies, and the politics of the movement significantly influenced this research as it emerged in the early 1970s. The most general political position of the Women's Liberation Movement that shaped the study of gender ideologies was the view that women are not equal to men in American society. They do not have the same control over their own lives and the lives of others that men have. They are dominated by men in their family life, in the workplace, and in other social domains as well, particularly religion and politics. This domination, it was argued, is bolstered by patriarchal gender ideologies that provided justification for men's domination of women. The term 'patriarchal' was used to refer to ideologies that either

assumed or asserted that the correct social order was one in which men dominate women, have authority over them, and tell them what to do. The patriarchal ideology that received the greatest attention in the Women's Liberation Movement was the American view that women are biologically inferior to men—less intelligent, physically weaker, less aggressive, and more emotional—in ways ultimately explained by differences in their biological makeup. The political theorists of this social movement actively criticized this patriarchal ideology. They acknowledged that there were biological differences between women and men, but argued that these were exaggerated by cultural processes and were in themselves not sufficient to be a basis for men's political and social domination. Instead, differences in women's and men's socialization experiences explained cultural differences between women and men that were used to justify men's authority. Academic disciplines that had contributed to the development of biological patriarchal ideology, also came under attack from the women's movement. 'Science' in general was conceptualized as a form of ideology, that is as a framework promulgated by people in power to help them maintain their power, rather than as objective factual truth. There was evidence that scientists who had studied sex differences between humans had indeed been influenced by the patriarchal ideologies of European cultures. For example, nineteenth century physical anthropologists found that women's brains were smaller than men's and argued that this was the reason women were less intelligent. Yet by the early twentieth century, physical anthropologists were arguing that intelligence was a reflection of the size of an animal's brain in proportion to its body, and women and men did not differ in this proportional relationship.

Patriarchal gender ideology was also apparent in the way women were largely absent from scholarly representations of human activity in the social sciences and humanities of the 1960s. For example, in anthropological theory about early human societies, considerable attention had been given to the importance of men's collaborative hunting for human evolution. But women's gathering of plant foods, which provided the bulk of foodstuffs eaten by early humans, was almost completely ignored. Feminist theorists argued that both biological theories which denigrated women and representations of humans that rendered women invisible conveyed gender ideologies which were psychologically harmful to girls and young women. These scholarly visions of gender undermined young women's self-confidence, encouraging them to believe that they were incapable of much in life that they were in fact quite capable of, and providing perspectives that suggested male domination was indeed an appropriate state of affairs.

This vision of the role of gender ideology in US women's subordination was projected to the world as a whole in early feminist anthropological comparative

research on the topic. In the early 1970s, Rosaldo (1974) argued that the cross-cultural evidence on gender relations available at that time indicated that in all societies of the present and past men have had greater power and authority than women. In all societies, although both men and women have culturally legitimated authority in the domestic or household sphere, it is overwhelmingly men who have had culturally legitimated authority in public spheres, due to women's greater involvement in childcare. Rosaldo suggested that this arrangement was supported everywhere and reinforced by gender ideologies in which men were held to be superior to women by virtue of their greater involvement in culture and their lesser involvement in nature. Women were held to be more involved in nature than men because of their role in the natural process of giving birth to and caring for children. This argument was appealing in part because of its scope: it was a theory about women and men in all times and places. It was also appealing because it resonated with the views of feminists involved in the Women's Liberation Movement of that time.

In many ways, Rosaldo's position set the agenda for comparative research on gender ideology for the next several decades. Her position was challenged as well as embraced and the research it stimulated showed that there is far more diversity in gender ideologies than was initially envisioned. The rest of this discussion focuses on the nature of this diversity in gender ideologies. Consideration will first be given to how many gender categories there are in gender ideologies. Early work assumed two—woman and man—but this view has been questioned. Then the discussion will shift to the nature of cross-cultural similarities and differences in fundamental features of gender ideologies that are predominantly dualistic in organization. Finally, the extent to which patriarchal gender ideologies sustain male domination cross-culturally will be considered.

2. How Many Genders Are There?

One key source of cultural and historical variation in gender ideologies is the extent to which gender systems are thought of as fundamentally dualistic, or binary. Much contemporary Western scholarship assumes that binary gender systems are universal: sexual females become women and sexual males become men. However, research on sex and gender ideologies suggests that this is itself a culturally specific gender ideology. To begin with, binary biological sex is itself a cultural construct. Laqueur (1990) argues that until the nineteenth century, Europeans conceived of only one sex, that of male, with female as a lesser version of him. It was only in the late eighteenth century that a two-sex model of human biology began to emerge in which the differences between males and females were rhetorically elaborated. Today, although it is common to assume that all babies will be obviously biologically

male or female, sex is sometimes ambiguous. In such cases an infant will still be assigned to one category or the other and surgery may be performed to alter the body in a manner consistent with that assignment. Adults sometimes decide that they want to have the physical characteristics of the 'opposite' sex. To achieve this they engage in cultural activities that will either create this impression, as when women bind their breasts, or bring about sex changes as a physical reality through surgery and hormonal treatments.

There is no doubt that binary biological sex is partly a material reality. It is not something that we as humans have imagined completely. But it is also partly a cultural construct. Scholars agree that binary gender ideologies, as well as binary sex ideologies, are ideological constructs. They also agree that the gender ideologies articulated by people do not capture for any society the range of gender identities expressed in that society. But they disagree about how to conceptualize nonbinary models of gender. Some researchers assert the existence of societies in which there are three rather than two salient genders, where the third gender is most commonly a named category referring to a man who is womanly in dress, occupation, sexual preference, or some combination of these aspects of gender identity. For example, Europeans who first contacted North American Indian societies reported that many of them had a third gender category of men who dressed like women and carried out women's work roles, the name for which was sometimes glossed into English as 'manly hearted woman.' Other scholars argue that even in cultures with three or more named gender categories, gender as an ideological system is still and always about cultural transformations of a biological difference between males and females. In any case, gender dualism, or the idea of two genders, woman and man, is very strongly developed and salient in gender ideologies around the world and through time, and most scholars working on gender ideologies have focused on the nature, causes, and consequences of gender ideologies that posit woman and man as central categories.

3. Cross-cultural Similarities and Differences in Dualistic Gender Ideologies

Dualistic or two-category gender ideologies are concerned with how women and men are similar and different, with the emphasis on difference. The widespread presence of ideologies of difference indicates the importance of culture and ideology in creating women and men as unlike. Anthropologists have found that cross-culturally, people are concerned with similar aspects of human experience in their production of gender ideologies. People also engage in similar cognitive and symbolic processes as they talk about and otherwise represent women. At the same time, there are also differences from society to society

and within societies in the nature of dualistic gender ideologies, so that similarities and differences across cultures are intertwined.

There are at least four aspects of human experience that regularly enter into gender ideologies. These include work, appearance, sexuality, and reproduction. People constitute themselves and are socially constituted by others as women and men through their behavior in these areas and the meanings attached to those behaviors. At the same time, societies vary in the emphasis that is given to each of these four aspects and in the extent and kind of variation tolerated in behavior relative to canonical expectations about gender.

It is very common cross-culturally for women and men to be expected to do different kinds of work, and for girls and boys to be socialized into different kinds of labor. So, for example, in hunting and gathering societies, which no longer exist in isolation from other economic types of societies, it is common for women to be expected and trained to gather plants and for men to be expected and trained to hunt for game. Among European agriculturalists, large-scale cultivation has been associated strongly with men. Consequently, where women were largely responsible for cultivation in places that have been colonized by Europeans over the last 500 years, strong pressure was often exerted to get men to replace women in these activities as, for example, with the Seneca Indians in New York state.

After about the age of 3 years, girls and boys and women and men are also commonly expected to look different from each other in ways that are specific to individual cultures. They are expected to dress differently, to wear their hair differently, to differ in the presence and nature of body decor such as jewelry and tattoos, and to differ in patterns of nonverbal behavior such as gesture and facial expression.

Sexually, binary gender ideologies entail the expectation that men's sexual object choices will be women and women's sexual object choices will be men. But whether or not and how departure from these expectations leads to the view that a male is not a man or a female is not a woman varies greatly cross-culturally.

Finally, binary gender ideologies commonly entail expectations of women and men regarding the roles they will play in the reproduction of human communities, through giving birth to, taking care of, and socializing children. It is well documented that failure to meet cultural expectations in these areas can often have major consequences in the lives of the people involved. However, a worldwide decrease in emphasis on this aspect of gender identity may be ongoing as changes in global economic processes bring about changes in family size.

Cognitively and symbolically, it is common for these substantive areas of gender ideology to be elaborated around specific gender dyads (Mathews

1992). Typically these dyads are drawn from cross-gender relationships within the family, reflecting the fundamental centrality of the family for both the maintenance of and changes in gender ideologies. Some key nuclear family dyads that are frequently ideologically elaborated cross-culturally include wife–husband, sister–brother, mother–son, and father–daughter.

However, societies differ in which gender dyads get selected for ideological elaboration. For example, in many European societies, the husband–wife relationship is the primary relationship that is ideologically elaborated, and the mother–child relationship is also the focus of a great deal of cultural production. But in the South Pacific, in Polynesian cultures such as those of Samoa, Tonga, and Tokelau, it is the sister–brother relationship that is most celebrated and discussed, and the husband–wife relationship is simply not developed explicitly or represented to the same extent in public events. So, for example, stories and plays depict the kind of work the sister should do and the kind of work the brother should do, but not what husbands and wives do, though one may infer that the division of labor will be similar for them.

While a good deal of gender ideology develops around nuclear family dyads, it also receives expression in representations of culturally idealized individual men and women, who usually represent the good, but can also represent the bad. For example, in Mexico, there are three key female icons: the Virgin of Guadalupe, La Malinche, an Indian woman who betrayed her people by consorting with the Spanish conqueror, and La Llorona, a ghostly figure who wanders riverbanks weeping over the loss of her dead children and luring men to death by drowning (Limon 1986).

The use of simple cross-gendered dyads from the family and the use of individual icons to convey gender ideology are, then, common cross-cultural cognitive and symbolic strategies. An additional such strategy is to metaphorically project gendered family dyads and individual gendered prototypes outside the family onto other more encompassing levels of social organization. For example, George Washington is referred to as the father of his country. And entire nations may be gendered, often as women, as in the expression, ‘India is the mother of her people’ and the term ‘Mother Russia.’

The processes involved in cross-cultural similarities and differences in gender representation that have been described so far also provide resources for the production of intrasocietal as well as intersocietal similarities and differences in gender ideologies. Different segments of a society—e.g., classes, ethnic groups, or religious groups—may emphasize different aspects of gender identity in their gender ideologies. Diversity in gender ideologies within a society can consequently become the basis for criticism and conflict between groups over how women and men

should be expected to behave and should treat one another. This can lead to the imposition of the gender ideology of one group on another, and to changes over time in gender ideologies.

For all of these reasons it can be problematic to talk about ‘the’ gender ideology of a given society or culture. It is true that some ideas about gender may be ‘dominant’ in a given society, in that they are represented widely or represented by those in power. But diverse views on gender are held in all societies, if only by virtue of role differentiation within the family, so that what is expected of a sister is not what is expected of a wife or mother.

4. Gender Ideologies and Male Domination Over Women

The final important and controversial aspect of gender ideologies to be considered here is the extent to which they provide justification for men’s control of women and contribute to its perpetuation. Patriarchal gender ideologies are very widespread, but not all of them are alike. The political critique of the Women’s Liberation Movement concentrated on the American and European patriarchal ideology that gender differences are rooted in biologically based sex differences which render women inferior to men. But this was and is not the only patriarchal gender ideology in the United States or elsewhere. Biological differences between women and men are not always involved. Nor is women’s inferiority always asserted. Neither is necessary for a patriarchal gender ideology. What is necessary is that there be a cultural understanding that men should have power and authority over women that women should not have over themselves or men. Two important bases for patriarchal gender ideologies besides biology are tradition and religion.

Adhering to tradition means continuing to do something because it has been done in the past, often a highly valued past that itself defines who people are in the present. In religious ideologies, in contrast, patriarchal gender ideologies are often conceptualized as supernaturally ordained. This is true of gender ideology in the three world religions that originated in the Middle East and spread around the world: Judaism, Christianity, and Islam. In all these religions, gender roles and the greater authority of men is interpreted as the will of God.

Although patriarchal gender ideologies are very widespread, nonpatriarchal gender ideologies and gender ideologies that do not denigrate women but value them and represent them in positive ways are also very widespread. Quite often they co-exist with patriarchal gender ideologies. In Tonga, for example, gender ideologies reveal different attitudes toward and expectations for sisters and wives. Women are conceptualized as below their husbands, but above their brothers, so husbands can tell wives what to do and

sisters can tell brothers what to do, but brothers are not supposed to boss their sisters and wives are not supposed to boss their husbands. A good deal of comparative research on gender ideologies assumes or asserts a close relationship between the amount of power women have and the extent to which gender ideology portrays women positively in a given society. Both women's power and gender ideology about them are in turn affected by aspects of cultural systems other than the gender frameworks themselves. The idea that gender ideologies and power relations between women and men vary in ways shaped by other social forces is important. It raises the possibility that women's positions can be enhanced by human transformation of social systems through identification of processes in them that favor gender equality. Various aspects of social systems have been put forward as influencing gender ideologies and power relations. For example, many scholars have argued that women have more power and more positive images in societies where they make significant visible contribution to the economic well being of the groups they belong to. Others argue that although present day societies may all have patriarchal gender systems, this has not always been the case. Some argue that 'states' as a type of political system entail a concentration of power in men's hands that was not true of prestate political systems in the evolution of human societies. Others focus on European colonialism over the last several centuries as a cultural influence that has made many societies patriarchal which were not patriarchal before. However, these ideas have been criticized for romanticizing past societies whose gender ideologies and power relations may never be completely reliably retrieved. Since the 1990s, there have also been changes in the way the power of gender ideologies has come to be conceived. It is increasingly common for anthropologists and others to write of the 'gendering' of very large-scale social processes, particularly colonialism, the emergence of nation states, and international relationships. This work suggests that the ideological power of gender is located not so much in explicit representations of gender in myths and rituals as in a more implicit, pervasive, and diffuse semiotic gendering of all of social life. Still, the most important contribution to date from the study of gender ideologies has been the documentation of considerable cross-cultural diversity in gender ideologies that describe and explain how and why women and men are different from each other. The documentation of this diversity in itself undermines and deconstructs essentialist views that differences between women and men are natural, supernaturally ordained, or fixed for all time in any society.

See also: Belief, Anthropology of; Culture as Explanation: Cultural Concerns; Economic Development and Women; Feminist Theory; Fertility and Culture: Anthropological Insights; Folklore; Foucault, Michel

(1926–84); Keynes, John Maynard (1883–1946); Male Dominance; Masculinities and Femininities; Mead, Margaret (1901–78); Witchcraft

Bibliography

- Abu-Lughod L 1986 *Veiled Sentiments: Honor and Poetry in a Bedouin Society*. University of California Press, Berkeley, CA
- Bozzoli B, Nkotsie M 1991 *Women of Phokeng: Consciousness, Life Strategy, and Migrancy in South Africa 1900–1983*. Heinemann, Portsmouth, NH
- Kennedy E L, Davis M D 1993 *Boots of Leather, Slippers of Gold: The Story of a Lesbian Community*. Routledge, New York
- Laqueur T 1990 *Making Sex: Body and Gender from the Greeks to Freud*. Harvard University Press, Cambridge, MA
- Limon J 1986 La Llorona, the third legend of greater Mexico: Cultural symbols, women and the political unconscious. *Renato Rosaldo Lecture Series Monograph 2*: 59–93
- MacCormack C P, Strathern M (eds.) 1980 *Nature, Culture and Gender*. Cambridge University Press, New York
- Mani L 1998 *Contentious Traditions: The Debate on Sati in Colonial India*. University of California Press, Berkeley, CA
- Mathews H 1992 The directive force of morality tales in a Mexican community. In: D'Andrade R G, Strauss C (eds.) *Human Motives and Cultural Models*. Cambridge University Press, New York
- Ortner S B, Whitehead H (eds.) 1981 *Sexual Meanings: The Cultural Construction of Gender and Sexuality*. Cambridge University Press, New York
- Reiter R R (ed.) 1975 *Toward an Anthropology of Women*. Monthly Review Press, New York
- Rosaldo M Z 1974 Woman, culture, and society: A theoretical overview. In: Rosaldo M Z, Lamphere L (eds.) *Woman, Culture, and Society*. Stanford University Press, Stanford, CA
- Sanday P R 1981 *Female Power and Male Dominance: On the Origins of Sexual Inequality*. Cambridge University Press, New York
- Sanday P R, Goodenough R G (eds.) 1990 *Beyond the Second Sex: New Directions in the Anthropology of Gender*. University of Pennsylvania Press, Philadelphia
- Stoler A L 1989 Rethinking colonial categories. *Comparative Studies of Social History* 31(1): 134–61
- Yanagisako S, Delaney C (eds.) 1995 *Naturalizing Power: Essays in Feminist Cultural Analysis*. Routledge, New York

S. U. Philips

Gender-related Development

Being born a girl or boy sets a wide range of forces in motion. These forces go beyond the genetic, hormonal, and genital differences inherent in one's sex. Parents provide many opportunities for learning about what it means to be a boy or girl. Teachers and peers reinforce and encourage gender-typed behaviors in children. Children also become active participants in learning about gender. Gender development is a term that

encompasses all the many gendered aspects of children's development. To understand children's gender development, the entire range of social, cognitive, and biological influences must be considered.

1. Overview of Gender Development

In recent years, gender has come to be viewed as a multidimensional construct that covers many domains including biological sex, activities and interests, personal-social attributes, gender-based social relationships, gender-based styles and symbols (such as gestures, style of dress), and gender-related values (see Ruble and Martin 1998, Huston 1983). Each domain can be examined in terms of an individual's concepts, self-perceptions, preferences, and behaviors. In the past, there was an implicit assumption that a person who was gender typed in one domain also was gender typed in other domains. The multidimensional view of gender has been useful in questioning this assumption and replacing it with the challenge to researchers that interrelations among domains need to be assessed. Also, this view has the advantage of making clear the distinctions among the different domains.

Certain aspects of gender development have been investigated more than others, including gender roles, self-perceptions of gender, gender identity, and gender-related behaviors. Gender roles have been a popular topic of study, within both psychology and sociology. Gender role is a broad term referring to the behaviors, attitudes, and traits that are associated with being male or female. Many studies have been done to identify the roles that men and women fulfill in western cultures as well as in other countries and in our history. Gender stereotypes, which are culturally defined expectations about the sexes in terms of personality, appearance, occupations, abilities, and interests, have been a very central subject for social psychologists. Also, psychologists have been involved in studying self-perceptions of gender, especially individual's perceptions of masculinity and femininity. Gender identity, the sense of oneself as being a male or female, has been studied by clinical and developmental psychologists, whereas a typical gender identity has been investigated largely by medical researchers. Gender-related behaviors and characteristics have been the subject of numerous studies, including how children develop skills and social behaviors such as their relationship preferences.

2. Age Trends in Gender Development

2.1 Gender Identity

Gender identity is the sense of oneself as a male or female. Gender identity involves more than the anatomical—it involves the inner knowledge and outer expression of one's sex. The dominant view has been

that gender identity resides within the individual, although more recent views challenge this assumption by arguing that gender identity is created in the context of interactions, societal structures, and cultural expectations (Deaux and LaFrance 1998). Research on gender identity has also focused on the consequences of developing gender identity that is consistent with or different from one's genetic sex.

The predominant view has been that for typical children, the process of developing a sense of oneself as a girl or boy undergoes three stages. First, at about the age of 2–3 years, children learn to label themselves and others by sex. Second, 3–4-year-old children come to understand the stability of gender over time. Third, by the age of 4–5 years, children recognize that one's sex is a constant aspect of oneself that is not variable over situations (Slaby and Frey 1975). As children's sense of gender identity develops, they often show increased attention to people of the same sex, and are motivated to be like members of their own sex (Kohlberg 1966).

A very small group of children, about 2–5 percent of the population, develop a sense of gender identity that does not match their biological gender. These children, labeled as having gender identity disorder, express discomfort with their own sex, wish to be the other sex, and engage in behavior typical of the other sex (Zucker and Bradley 1995). The development of gender identity disorder may be related to biological factors, such as hormones and temperament, in addition to environmental factors, such as parental encouragement of cross-gender play.

An even smaller group of children are born with ambiguous genitalia or genitalia that do not match their genetic sex. These children may be assigned a sex that differs from their genetic endowment or may not be assigned a sex until later. Gender assignment, and also gender identity, become very complicated issues for these children. Although early research suggested that children who were assigned sex at an early age would develop according to their assignment regardless of their biological sex, more recent research has challenged this view. In some reports, both biological and social factors play a role in developing gender identity, whereas in others, only biological sex or only gender assignment is central in developing gender identity. Controversy continues over how to best support these children in developing their gender identities (see Ruble and Martin 1998).

2.2 Gender Differences in Interests, Behaviors, and Well-being

As early as 2 years of age, children begin to show sex-differentiated activities and interests (Ruble and Martin 1998). In the preschool years, children's preferences for stereotypic play become even more apparent. Girls and boys prefer different toys and prefer to play in different areas. The tendency to play with stereotypic

toys increases during early childhood. Both girls and boys show preferences for stereotypic activities, although boys seem to avoid other-sex activities more than girls.

Sex differences in activities persist into middle childhood and adolescence although the similarities in behavior are more striking than differences. Boys are more involved in sports, outdoor chores, and leisure activities; girls are more involved in indoor chores, shopping, and socializing. Through adolescence, girls spend increasing amounts of time in interpersonal activities, personal care, and household chores and less time in sports whereas boys' interest in sports remains stable or may increase.

Males are more active than females with the differences being more pronounced in older than in younger children. Girls and boys differ in many motor tasks, with boys outperforming girls in throwing and girls outperforming boys in tasks involving flexibility and fine-motor skills. These differences increase with age, possibly owing to differential participation of girls and boys in sports after puberty, although biological factors may also play a role in differences (see Ruble and Martin 1998).

Moderate and consistent sex differences are found in the area of physical aggression, with boys showing more aggression than girls. The exception to this is relational aggression—using behaviors to damage friendships—which girls engage in more than boys. Sex differences in aggression are larger in children than in adults, although men commit many more violent crimes than women.

The sexes differ to only a small degree in prosocial behavior and differences vary depending on the kind of behavior being assessed. Generally, girls help other people more than boys (Eisenberg and Fabes 1998). Females also are more accurate in decoding emotions from visual cues and they are more expressive and responsive than males.

Many varied indicators of well-being suggest that boys experience higher levels of well-being than girls. Beginning in adolescence, females are more likely to be clinically depressed or have depressive symptoms than males. Although the differences are small, boys report higher levels of self-esteem than girls and several studies suggest that girls experience a drop in self-esteem after the age of 9 years as compared with no change for boys. Girls have many more appearance and weight concerns than boys and these concerns increase during adolescence (see Ruble and Martin 1998). Girls' concern with their appearance may relate to their self-esteem and to eating disorders and disturbances.

2.3 Gendered Social Relationships

One of the most pervasive and consistent differences between the sexes is their choice of play partners (Maccoby 1990). Girls and boys are often found

playing with members of their own sex. The phenomenon of sex segregation appears to be virtually universal across cultures. Sex segregation varies in differing situations, however. Children's tendency to interact with same-sex children is common in situations in which children have choices about playmates, such as in school, and in unstructured situations, and is less common when fewer choices are available, such as playing in their neighborhoods, or when children are in structured situations in which they need to cooperate.

Sex segregation begins very early in life. Girls tend to segregate before boys, usually starting between 2 and 3 years of age. Boys who are more socially disruptive and girls who are more socially sensitive segregate earlier than other children (Moller and Serbin 1996). Sex segregation increases as children grow older. Sex segregation continues until middle adolescence when children begin to form heterosexual dating couples. Children's friendship patterns mimic their play partners in that most often, children's friends are of the same sex as themselves.

The kinds of play typical of girls' groups and boys' groups are different. Boys tend to play more actively, encompass more space, and they often play outdoors. Girls tend to play indoors, more quietly, and use less space. Girls are more likely to encourage cooperation and interaction among play partners; boys are more often interested in issues of dominance and competition. From their experiences with same-sex partners, girls and boys learn different methods of interacting with others and may be considered to be socialized in 'different cultures' (Maccoby 1990).

The origins of sex segregation are unclear. One contributing factor to sex segregation is similar interests. Children tend to play with same-sex children regardless of activity, so similarity in interests is not the only explanation for sex segregation. Compatibility in play styles is also important. Children show higher levels of social interaction when playing with their own sex, which suggests that same-sex play is more entertaining than other-sex play. Activity level may also be a factor. Boys tend to play actively and that may be compelling to other boys but not to girls. Finally, children's expectations about the sexes also likely contribute to sex segregation (Ruble and Martin 1998).

The developmental course of social relationships in later adolescence and adulthood, especially the development of sexual orientation, has been difficult to chart. Although both sexes show very strong propensities to select other-sex romantic partners, about 5 percent of the population selects same-sex partners. The origins and developmental pathways of sexual orientation have been hotly debated. The central issues appear to be the extent to which sexual orientation is influenced by biological factors, such as genes, hormones, and brain structures, or by social factors, such as one's upbringing.

2.4 Gender Differences in Cognitive Skills

Many studies have been conducted to examine whether the sexes differ in cognitive abilities. In many areas, no sex differences are found, e.g., in overall intelligence or in analytic reasoning (Halpern 1992). When the sexes differ in a cognitive ability, the differences tend to be fairly small and hence are unlikely to account for differences in career choices or trajectories. Nonetheless, the topic of sex differences in cognitive abilities remains controversial.

Most of the research on sex differences in cognitive abilities has focused on three areas—mathematical, spatial, and verbal abilities. In the domain of mathematical abilities, the patterns of similarities and differences vary depending on many factors including the particular skill being assessed, the age and skill level of the sample, and assessment procedures. Differences tend to be larger and favor males when studies are conducted on the most mathematically skilled children and when using certain standardized tests, such as the SAT. Also, sex differences vary for different skills, with girls receiving higher grades than boys in mathematics courses, and a small difference in favor of girls for computational skills in elementary school. Males show better performance in problem-solving tasks beginning in junior high or high school. Some evidence suggests that sex differences in mathematical abilities are lessening over time.

The sexes differ in some aspects of spatial abilities but the size of the difference depends on the particular ability being assessed and the age of the participants. Few differences are found between young girls and boys. Most sex differences begin to appear in adolescence and differences increase with age. A moderate-size difference favoring adolescent and adult males is found in tasks involving mental rotation, which involves the rapid manipulation of three-dimensional objects. A moderate to small difference favoring males occurs in tasks involving recognition of the horizontal or vertical. A small difference (or no difference) is found when assessing performance on spatial visualization tasks, which involves finding hidden spatial figures (Halpern 1992).

Very small sex differences are found in verbal tasks, with females outperforming males. The differences are found most often in the areas of spelling and language and tend to be stable across age. Verbal differences may be lessening over time although not all researchers concur with this conclusion.

In a large-scale assessment of adolescents' academic abilities, researchers have found that males outperformed females in tests of science, mathematics, social studies, and vocational aptitude (mechanical reasoning and electronics), whereas females outperformed males in reading comprehension, perceptual speed, associative memory, and writing (see Ruble and Martin 1998).

2.5 Gender Stereotypes

Gender stereotypes are the beliefs that people have about the characteristics of males and females. The content of stereotypes varies over cultures and over time. These expectations are often related to the roles that the sexes fulfill in the culture.

Children learn some aspects of stereotypes at a very young age. By the age of $2\frac{1}{2}$ –3-years, children show evidence of having some rudimentary knowledge of the activities and objects associated with each sex. Children's gender stereotypes of activities and occupations develop quickly during the preschool years, reaching a very high level by kindergarten. During the elementary school years, gender stereotypes broaden to include sports, school subjects, and personality traits. With age, children become increasingly knowledgeable about gender stereotypes and yet the rigidity of their stereotypes declines as they increasingly recognize the cultural relativity of these norms (Huston 1983). Some evidence suggests that boys hold more rigid gender stereotypes than girls and are held to more rigid ideals than girls. In adolescence, flexibility in stereotypes fluctuates in response to two opposing forces—increasing cognitive flexibility tends to increase adolescents' flexibility in applying stereotypes whereas increasing pressure to conform to stereotypes in preparation for sexual roles and adult status increases adherence to stereotypes (Ruble and Martin 1998).

Children use gender stereotypes to make inferences about others at a young age. When making judgments of other people, children and adults will apply their gender stereotypic expectations to them. Even more than adults, children will rely on a person's sex to make judgments and they are less likely to consider other relevant information about the person than adults are (see Ruble and Martin 1998).

3. Theories of Gender Development

3.1 Psychoanalytic Theory

The earliest and most influential theory of gender development was proposed by Sigmund Freud in the 1930s. Freud contended that boys and girls acquire gender identity by learning to negotiate desires for the opposite sex parent and by eventually identifying with the same-sex parent. Although little empirical support has been found for his theoretical framework, his ideas have been very influential in encouraging research on gender development. Over the years, additions and amendments have been made to Freud's original psychoanalytic theory. Many of these changes take into consideration the role of societal influences and include a feminist perspective.

3.2 Social Learning Theory

Social learning theorists (Bandura 1977, Mischel 1966) propose that gender development is accounted for by social factors. Girls and boys learn different roles through the effect of direct reinforcement, differential socialization, and imitation of models.

Direct reinforcement, which refers to the reaction children receive in response to their behavior, leads to children learning to discriminate between the behaviors typical of boys and girls. Children learn that there are different consequences for engaging in behaviors depending upon whether one is a girl or a boy. Behaviors considered to be stereotypic for their sex are more likely to be rewarded. Children perform activities that are consistent with their reinforcement histories (Mischel 1966).

Children also experience differential socialization of girls into feminine roles and boys into masculine roles as they are exposed to distinct environments in which gender differences are encouraged by socializing agents, such as parents, teachers, peers, and media. Parents fill boys' rooms with educational, interactive toys that promote exploration, whereas they fill girls' rooms with toys that promote nurturance and domesticity, such as doll houses and tea sets. Teachers respond differently to boys and girls: boys are more often called on to answer questions when their hands are raised than girls. Differential treatment by social agents provides boys and girls with separate sets of rules, expectations, and opportunities which, in turn, encourages traditional gender development.

Children also learn gender roles through observing others. Social learning theorists contend that individuals learn and reproduce the actions and responses of the models they encounter in their families, schools, and neighborhoods, and in books, television, and films. Gender-typed behavior develops as girls imitate the behavior of females and boys imitate the behavior of males. Children's imitation also is affected by their relationships with the model, in addition to the amount of power they perceive the model to possess (Mischel 1966, Bussey and Bandura 1984). Witnessing people's reactions to others as they display a behavior transmits information about the appropriateness of the behavior (Mischel 1966).

3.3 Cognitive Social Learning Theory

In the late 1970s and 1980s, social learning theory was adapted to emphasize the cognitive aspects involved in learning from the social environment (Bandura 1986). In 'cognitive social learning theory,' emphasis was placed on the processes involved in creating expectancies of other's behaviors and how these expectancies guide one's own actions. Specifically, it was proposed that observational learning required four processes: paying attention, being motivated to enact the behavior, having the ability to reproduce the action, and

having the ability to remember the behavior (Bandura 1986).

Social learning approaches have encouraged an expanded view of the forces that influence gender development. This theory has motivated researchers to include a more diverse view of the child's social environment. In addition, it has led to closer scrutiny of the social settings provided for children and the gender-laden messages these settings contain.

3.4 Cognitive Developmental Theory

According to Kohlberg's cognitive developmental theory, children's cognitive organization of the world around them is the basis for their gender development (Kohlberg 1966). Kohlberg proposed that between the ages of 2 and 7 years, children make cognitive judgments about their gender identity (whether they are a boy or a girl) and then come to understand that gender is constant over time and situation. This understanding about gender becomes the foundation for further gender development. A child's recognition of being a male or female determines the value they place on masculine and feminine traits, initiates a child's desire to be similar to same-sex children, facilitates the development of gender stereotypes, and encourages children to hold positive views of their own sex (Kohlberg 1966).

A debate has ensued over whether Kohlberg's theory requires the acquisition of gender constancy as a necessary antecedent to learning stereotypes and attending to same-sex models (Ruble and Martin 1998). Empirical studies suggest that children have only a basic knowledge about gender before the onset of the gender-typing process, and that stereotyping can begin once children are able to label the sexes correctly. Gender constancy may help consolidate gender development, rather than initiating it (Ruble and Martin 1998). Cognitive developmental theory emphasizes the child's active role in gender development. Kohlberg's theory has been enormously influential in shaping the way researchers think about gender development.

3.5 Gender Schema Theory

Gender schema theory is a cognitively based theory that uses an information processing approach to explain how gender development occurs. The basis of this model is the cognitive representation called a schema. A schema is an organizing structure that helps simplify and categorize new information.

There are two types of gender-related schemas (Martin and Halverson 1981). The first is a general 'superordinate' schema that helps children categorize objects, characteristics, and traits into basic male and female categories. The second is a more narrow version of schema, called the 'own-sex' schema, that children

use to identify and learn in-depth information consistent with their own sex. These two schema types allow children to process information about events, objects, attitudes, behaviors, and roles and, in turn, categorize these aspects in terms of masculine or feminine, or as similar or different from the child (Martin and Halverson 1981).

Gender schema theorists (Bem 1981, Martin and Halverson 1981) proposed that gender schemas are developed and applied by children at an early age. In addition, it was proposed that the presence and use of gender schemas can explain the ways in which children and adults attend to, acquire, and remember information; how they behave; and the kinds of attitudes they develop (Martin and Halverson 1981). Gender schemas affect which kinds of information are noticed, encoded, and remembered, with information inconsistent with the existing schema being overlooked and consistent information being processed (Signorella et al. 1997). Research also shows that gender schemas can guide children's preferences, toy choices, and play partner choices.

Because schemas are used to interpret and categorize information, however, they can result in the drawing of incorrect inferences. Distortion of information has also been found to be a consequence of gender schematic processing. Altering information that does not fit into a currently held schema or placing neutral information into a gender schema leads to remembering information in a gender-consistent manner, which may not match with how it exists in reality. These processes illustrate how gender schemas can be easily maintained and how resistant they are to change.

Gender schema theory emphasizes the child as an active processor of information and that the gender schemas that guide thinking also influence children's behavior. The major strength of gender schema theory is in the understanding it provides about the maintenance and power of gender beliefs.

3.6 *Biological Approaches*

Much of the research using a biological approach has focused on understanding differences between the sexes. These approaches often consider, for instance, the influence of hormones and brain structure and function on sex differences.

Sex differentiation begins with the presence of an X or Y sex chromosome in the twenty-third pair of chromosomes. These sex chromosomes carry the genetic information that determines the development of fetal sex organs. The amount of a specific androgen, called testosterone, which is present in both male and female fetuses, determines further development of sex organs. Increased levels of this hormone lead to the formation of the penis and scrotum in males. Female sexual organ development continues in the absence of

increased testosterone levels. Variations in the levels of hormones in these critical periods of development can lead to the formation of sex organs opposite to what would be expected from the sex chromosomes.

Hormones also influence the development of the central nervous system and the brain (Hoyenga and Hoyenga 1993). Hormones are most likely to affect brain development and behavior during two stages of life: during prenatal development and during puberty. Exposure to abnormal levels of hormones (due to genetic disorders or hormonal treatment during pregnancy) allows researchers to study the link between levels of hormones present during developmental processes and behaviors exhibited by individuals postnatally (Hoyenga and Hoyenga 1993). Studies of hormones, their effect on brain development, and the link to behavior suggest that they have an influence on gender development, although it is difficult to disentangle biological effects of hormones from social factors.

The structure and function of the brain have also been studied to compare the similarities and differences in males and females. One area of brain structure that has been studied is brain lateralization, or how much one hemisphere of the brain is responsible for carrying out specific functions. Some studies have shown that women are less lateralized in that they use both hemispheres simultaneously more often than men. It has been hypothesized that differences between men and women in a variety of abilities and behaviors could be due to a gender difference in lateralization. The findings of such studies show mixed support.

Another section of the brain that has been studied as a possible source of gender differences is the corpus callosum (CC). The CC is a group of fibers that connects the right and left hemispheres of the brain and is thought to be related to more intricate processes of thought, including language. Because the links between the structure, function, and size of the CC are not certain, research on differences in brain structure remains difficult to interpret. However, studies have shown that women have a larger CC than do men, and that enlargement of certain parts of the CC is linked to increased language skills.

Controversy surrounds the study of biological information on gender differences, especially concerning the extent of influence of biological and social factors. Most researchers of gender development accept an interactionist position, assuming that both nature and nurture are working together to produce human behavior. Furthermore, neither nature nor nurture should be viewed as steadfast or unmalleable (Hoyenga and Hoyenga 1993).

3.7 *Evolutionary Perspectives*

The evolutionary perspective draws on Darwinian principles of evolution and natural selection to explain

how gender differences evolve. Individuals who are fertile, live long enough to reproduce, and are able to raise their offspring will successfully pass on their characteristics to the next generation. Therefore characteristics of individuals who were successful in passing on their genes are more highly represented in future generations. According to the evolutionary viewpoint, gender differences arise because men and women play different reproductive roles (Trivers 1972).

Gender differences are hypothesized to develop because of differing parental investment levels. Evolutionary theorists contend that, because women are responsible for carrying the child during pregnancy and for producing food for the infant, their levels of parental investment are higher than men's. In contrast, men's reproductive role is limited to producing and investing their sperm. Thus, sexual selection differs for men and women. Because of a woman's increased investment level, she is more selective in her mates and searches for someone with access to resources. In contrast, men look for a woman who will care for the child, so that the child will eventually be fit to reproduce and pass on the family genes. However, because men never know if the child is actually carrying their genes, they are focused on mating more often and with shorter-term investments. Individuals who are most successful at their reproductive role are more likely to have their genes perpetuated, leading to a further divergence in men's and women's behaviors.

Evolutionary theoretical viewpoints have stirred much controversy, and empirical support for the theory's components is mixed. Critiques include the lack of ability to test the theory directly and the heavy dependence on animal studies.

4. Future Directions

The field of gender development moves quickly, making the prediction of future trends difficult. Nonetheless, some topics have garnered attention because they are important issues that have yet to be addressed. First, questions about the relation of mental health and domains of gender are worthy of additional research efforts. Why does the self-esteem of preadolescent girls drop more than that of boys? Second, research that focuses more attention on affective issues is needed to provide answers to questions about the role of self-regulation in gender development. Little is known about the affective and cognitive components that are involved in how people construct a view of their own gender and how others interpret gender identity. Third, many issues about how children are raised and about the contributions that parents, peers, and teachers can and do make to gender development are worthy of future research efforts. Finally, additional studies using multidimensional aspects of gender are important for understanding the complexities of gender development.

For instance, multidimensional studies are necessary to understand how a traditional vs. nontraditional view of one's gender identity relates to a variety of outcomes such as internal psychological states, relationships, and achievement in school and careers. The picture of gender development continues to emerge but is far from complete.

See also: Education and Gender: Historical Perspectives; Family and Gender; Gender Differences in Personality and Social Behavior; Masculinities and Femininities; Psychoanalysis: Overview; Small-group Interaction and Gender; Social Learning, Cognition, and Personality Development; Social Networks and Gender

Bibliography

- Bandura A 1977 *Social Learning Theory*. Prentice Hall, Englewood Cliffs, NJ
- Bandura A 1986 *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice Hall, Englewood Cliffs, NJ
- Bem S L 1981 Gender schema theory: A cognitive account of sex typing. *Psychological Review* **88**: 354–64
- Bussey K, Bandura A 1984 Influence of gender constancy and social power on sex-linked modeling. *Journal of Personality and Social Psychology* **47**: 1292–1302
- Deaux K, LaFrance M 1998 Gender. In: Gilbert D, Fiske S, Lindzey G (eds.) *The Handbook of Social Psychology*, 4th edn. McGraw-Hill, New York
- Eisenberg N, Fabes R A 1998 Prosocial development. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology*. J Wiley, New York, Vol. 3, pp. 701–78
- Halpern D F 1992 *Sex Differences in Cognitive Abilities*, 2nd edn. L Erlbaum Associates, Hillsdale, NJ
- Hoyenga K B, Hoyenga K T 1993 *Gender Related Differences: Origins and Outcomes*. Allyn & Bacon, Boston
- Huston A 1983 Sex-typing. In: Damon W, Hetherington E M (eds.) *Handbook of Child Psychology*, J Wiley, New York, Vol. 4, pp. 388–467
- Kohlberg L 1966 A cognitive-developmental analysis. In: Maccoby E E (ed.) *The Development of Sex Differences*. Stanford University Press, Stanford, CA, pp. 82–173
- Maccoby E E 1990 Gender and relationships: A developmental account. *American Psychologist* **45**: 513–20
- Martin C L, Halverson C F 1981 A schematic processing model of sex typing and stereotyping in children. *Child Development* **52**: 1119–34
- Martin C L, Little J K 1990 The relation of gender understanding to children's sex-typed preferences and gender stereotypes. *Child Development* **61**: 1427–39
- Mischel W 1966 A social learning view of sex differences in behavior. In: Maccoby E E (ed.) *The Development of Sex Differences*. Stanford University Press, Stanford, CA, p. 351
- Moller L C, Serbin L A 1996 Antecedents of toddler gender segregation: Cognitive consonance, gender-typed toy preferences and behavioral compatibility. *Sex Roles* **35**: 445–60
- Ruble D N, Martin C L 1998 Gender development. In: Damon W, Eisenberg N (eds.) *Handbook of Child Psychology*. J Wiley, New York, Vol. 3, pp. 993–1016
- Signorella M L, Bigler R S, Liben L S 1997 A meta-analysis of children's memories for own-sex and other-sex information. *Journal of Applied Developmental Psychology* **18**: 429–45

- Slaby R G, Frey K S 1975 Development of gender constancy and selective attention to same-sex models. *Child Development* 52: 849–56
- Trivers R 1972 Parental investment and sexual selection. In: Campbell B (ed.) *Sexual Selection and the Descent of Man*. Aldine, Chicago, pp. 1871–971
- Zucker K J, Bradley S J 1995 *Gender Identity Disorder and Psychosexual Problems in Children and Adolescents*. Guilford Press, New York

C. L. Martin and L. Dinella

Gender Role Stress and Health

1. The Term 'Gender Role'

The term gender role is used in sociology and psychology to refer to the prescribed behaviors, attitudes and characteristics associated with one's gender status as a female or a male. Attributes associated with gender are the result of learning in accordance with cultural standards or prescriptions.

Gender-role socialization refers to the processes through which individuals acquire attributes appropriate to males and females. Examination of gender roles indicates differences in the content of roles seen as appropriate for men and women. Men are socialized more often than women to be active and to ignore their feelings of weakness. Men are expected to be more independent, autonomous, and strong. It is more acceptable for women to express a wide range of emotions, to pay attention to their moods, and to seek out others for help. While men may be encouraged to be independent, women are taught to be more interdependent. The term interdependence refers to the ability to express interpersonal needs, particularly in emotional relationships, and to relate meaningfully to others in relationships (Greenglass 1982).

2. The Term 'Stress'

Stress is often used to refer to the perception that the demands of the environment exceed the resources of the person, resulting in a threat to well-being. These may emanate from work, family, and daily hassles, as well as significant life events, and include time pressures, conflicts, and frustration in meeting one's goals (see *Stress and Coping Theories*).

2.1 Stress and Health

The way individuals cope with their stress is important for mental health. To the extent that individuals have resources to handle their pressures, they are less likely

to suffer deleterious effects of stress. However, when individuals have difficulties in managing demands and stressors, they experience strain. Strain is a consequence of insufficient personal resources to meet demands. To the extent that individuals experience strain, they are more likely to experience deterioration in their physical and/or mental health. Some of the psychological effects of uncontrollable stress are depression, anxiety, panic attacks, hostility, and anger. High levels of stress may also result in physical symptoms including insomnia, headaches, muscular aches and pains, digestion disturbance, and difficulty in carrying out day-to-day activities (see *Stress and Health Research*).

2.2 Stress and the Feminine Gender Role

Research indicates that women are exposed to more stressful life events than are men and, as a result, have higher rates of psychological distress. Women experience depression, anxiety, and mood disorders more than men. Women may appear to be more distressed as a result of the kind of outcome measures typically used in this research. These studies tend to employ self-report measures of symptomatology. Some argue that women appear more distressed because they are more likely than men to report symptoms of all kinds. Gender-role factors contribute to the degree to which individuals report symptoms and seek help. Traditional masculine and feminine gender roles differ in the amount of vulnerability and permissibility in seeking help from others. Traditionally, men are seen as strong and invulnerable, thus restraining men from showing symptoms of mental illness or seeking help. The traditional feminine gender role allows and even encourages vulnerability to problems, thus resulting in greater reporting of symptomatology in women. It is assumed generally that if an individual is coping reasonably well, she/he is free from anxiety and depression. This conceptualization of effective coping leads to a bias in research findings given that women, more than men, often express distress (Greenglass and Noguchi 1996).

Notwithstanding women's greater 'openness' regarding their psychological symptomatology, research indicates many reasons for women's greater distress. Reproductive life events including menstruation, pregnancy, childbirth, and menopause are stressful life events for women. Within the employment sphere, women tend to be concentrated in low-paying 'female' occupational spheres, including education, healthcare, clerical, domestic, and service industries. Women have fewer promotional opportunities, make less money, and often do not have access to membership of the 'old boys' network,' all of which restricts their advancement in the employment sphere. Women professionals and managers are subject to negative stereotyped attitudes that portray women as doing a poor job

fitting into the masculine work world. Sex discrimination and harassment are ubiquitous in the workplace. Such attitudes not only harm women's salary and promotion prospects, they are also detrimental to their mental health. Women in the labor force also perform most of the domestic work, and on average spend more hours in housework and childcare than men. Employed married women manifest higher rates of distress than employed married men. Employed married women with young children experience higher rates of psychological distress than their female childless counterparts or comparable men. Given the demands on women when combining employment and familial roles, many women experience inter-role conflict and/or role overload (Barnett and Marshall 1991, Baruch et al. 1987).

Women's roles are inherently stressful in that they obligate them to respond to others, including the very young as well as the elderly. Informal care for elderly parents and relatives generally is provided primarily by women. These demands are a significant source of stress, thus playing an important role in women's mental health. Elements of the feminine gender role endanger women's mental health, since provision of care for their families and friends is often emotionally and physically draining, particularly given the relative lack of men's participation in work at home.

2.3 Stress and the Masculine Gender Role

Men have been diagnosed more than women with antisocial personality disorders in which there is a pervasive pattern of disregard for others' rights. Men are also diagnosed more as having schizoid personality disorders, characterized by detachment from social relationships and a restricted range of expression of emotions in interpersonal settings. These diagnoses may be seen as exaggerations of the traditional masculine gender role that stresses independence, separation, strength, and invulnerability (Greenglass 1982).

Research in the area of stress and burnout shows that men experience higher levels of depersonalization than women. Burnout at work commonly is measured using a scale (the Maslach Burnout Inventory) (Maslach and Jackson 1986) which yields three subscales: emotional exhaustion, depersonalization, and lack of personal accomplishment. Depersonalization refers to an attitude characterized by being impersonal, callous, and distant from others (Maslach and Jackson 1986). In this context, depersonalization may be regarded as a reflection of men's repressed emotionality and an ineffective coping form which allows men to continue their work with people yet remain untouched by the others' suffering.

The strict prohibition against emotional expressiveness in men (except perhaps anger) is one of the most powerful cognitive masculine mandates influencing men and their behavior. Male inexpressiveness

is the result of gender role socialization that begins early in childhood. Men are taught that emotional expressiveness is a feminine trait and they learn to distance themselves from femininity. Masculine socialization contributes to and encourages aggressive behavior in men. Men are more involved than women in violent crime, including armed robbery, spouse and child abuse, and rape. While aggression in sports and the military is approved and rewarded socially, the social program associated with masculinity also includes competitiveness and aggression that are socially sanctioned as masculine coping styles.

While women more often express distress in internalized and depressive symptomatology (Chesney and Ozer 1995), men often cope with their feelings differently, since they are more likely to externalize their feelings such as becoming physically aggressive or by distracting themselves from their feelings by engaging in a variety of activities. The violence committed by men against women in its many forms is in part an outcome of a patriarchal society which is dominated by men who see themselves as controlling women in many different contexts. Research suggests that power, sexual control, and aggression are amplified in men in a patriarchal society. The mental health costs of male violence justify re-examination of the prescriptions of the masculine gender role (Greenglass and Noguchi 1996).

Masculine prohibitions and difficulties with emotional expressiveness make it difficult for men to maintain mutually supportive relationships. When under stress, men seem to be less flexible in their coping than women, since they have difficulty expressing positive feelings.

3. Stress, Coping, and Gender Roles

Research shows that the ways people cope with stress differ according to their gender role. Some argue that men are more problem-focused in their coping, and women are more emotion-focused (Folkman et al. 1986). For example, men have been reported to engage more in coping that alters a stressful situation. Some assert that men more often possess psychological attributes, i.e., self-esteem and mastery, which influence their coping. However, gender differences in mastery or self-efficacy are likely to be linked to differences in social experiences rather than to gender *per se*. A consistently unresponsive or negative environment has been found to affect a person's sense of self-efficacy, which in turn leads to anxiety and depression. Studies of gender differences in psychological characteristics have pointed out that models of stress and coping were developed primarily on men without consideration of women's roles and experiences, and may not encompass other aspects of coping including relationship-focused coping techniques.

In research on coping, findings indicate that when education, occupation, and/or position are controlled for, few gender differences are found in coping strategies (Greenglass et al. 1990). The type of coping an individual employs when dealing with stress is often related to the status, resources, and power associated with one's position. Many more women than men hold lower level jobs with less scope for the use of problem-solving and direct action strategies. Women are also more likely to experience high levels of demands, with low control over demands in the workplace (Greenglass 1985).

In coping research, the focus tends to be on individual behavior as the unit of analysis, and often, autonomous, agentic, and independent behavior is seen as synonymous with effective coping. But research shows that coping is highly specific to both the individual and the context. The social milieu in which a particular stressor is experienced should be incorporated into research paradigms that inquire into the relationship between gender and coping. Women may behave in a more passive way in groups with men, where stereotyped roles and social norms may be more strictly enforced. Women may appear different when acting alone or in a group of women.

3.1 Gender, Social Support, and Mental Health

There is a close relationship between positive health outcomes and social support. Close relationships can help a person cope with psychological distress. In such relationships people can disclose and discuss problems, share concerns, and receive advice keyed to their psychological needs. These relationships can also provide useful information, practical advice and morale boosting, all of which can assist individuals in dealing with their stressors (Cohen and Syme 1985). Both men and women benefit from supportive social interactions, especially intimate ones (see *Social Support and Health*).

Women are more emotionally involved in others' lives and serve a nurturing role for a wider network of people. In general, women give and receive more social support than men. Women have more close friends, and emphasize intimacy and disclosure in their friendships. Women's friendships are more involving and emotionally supportive. Social relationships with women are more health promoting than those with men. Wives are better providers of both instrumental and emotional support than husbands. Men tend to benefit from relationships with women; women generally benefit from relationships with their friends of the same sex (Belle 1982).

Research shows that the connection between social support and coping is stronger in women. Women employ more coping forms involving interpersonal relationships. Other research findings suggest that women use support from others through talking with

one another. Women are able to make more effective use of their support networks, since they talk more as a way of coping with stress.

See also: Gender and Cardiovascular Health; Gender and Physical Health; Men's Health; Social Support and Stress; Women's Health

Bibliography

- Barnett R C, Marshall N L 1991 The relationship between women's work and family roles and their subjective well-being and psychological distress. In: Frankenhaeuser M, Lunberg U, Chesney M (eds.) *Women, Work and Health: Stress and Opportunities*. Plenum Press, New York
- Baruch G K, Biener L, Barnett R C 1987 Women and gender in research on work and family stress. *American Journal of Psychology* 42: 130–6
- Belle D 1982 The stress of caring: Women as providers of social support. In: Goldberger L, Bruntz S (eds.) *Handbook of Stress: Theoretical and Clinical Aspects*. Free Press, New York
- Chesney M A, Ozer E M 1995 Women and health: In search of a paradigm. *Women's Health: Research on Gender, Behavior, & Policy* 1: 3–26
- Cohen S, Syme S 1985 Issues in the study and application of social support. In: Cohen S, Syme S (eds.) *Social Support and Health*. Academic Press, Orlando, FL
- Folkman S, Lazarus R, Dunkel-Schetter C, De Longis A, Gruen R 1986 The dynamics of a stressful encounter: Cognitive appraisal, coping, and encounter outcomes. *Journal of Personality and Social Psychology* 50: 992–1003
- Greenglass E R 1982 *A World of Difference: Gender Roles in Perspective*. Wiley, Toronto, ON
- Greenglass E R 1985 Psychological implications of sex bias in the workplace. *Academic Psychology Bulletin* 7: 227–40
- Greenglass E R, Burke R J, Ondrack M 1990 A gender-role perspective of coping and burnout. *Applied Psychology: An International Review* 39: 5–27
- Greenglass E R, Noguchi K 1996 Longevity, gender and health: A psychocultural perspective. *Japanese Health Psychology* 4: 51–72
- Maslach C, Jackson S E 1986 *Maslach Burnout Inventory Manual*, 2nd edn. Consulting Psychologists, Palo Alto, CA

E. R. Greenglass

Genealogy in Anthropology

1. The Concept of Genealogy

Genealogy owes its position in modern cultural and social anthropology to the genealogical method, which will be the main focus of this account, giving it an inevitable bias toward British social anthropology. The English term genealogy has three components: it refers to the descent of a person, family, or group from an ancestor or older forms; to a written or graphic account of those relationships; and to the study of

family pedigrees. English genealogy is closely associated with pedigree. Pedigree refers to the record of an ancestral line or lineage; the origin or history of something; a distinguished ancestry; and the recorded purity of breeding of an individual or strain. The overlap between genealogy and pedigree has been significant for the way genealogy has been incorporated by anthropology, especially in the study of kinship.

Anthropologists distinguished pedigree, as a vernacular account of ancestrally defined relationships, from genealogy as a systematic or scientific account of those relationships. The distinction rests upon the genealogical method, whereby the blood and marriage connections recognized by the people being studied could be systematically recorded. The genealogical framework obtained in this way provided the basis for the comparative study of terminological systems of address and reference, which were also sometimes called systems of relationship; for tracing marriage regulations (prescriptions and proscriptions); and for discerning the rules of descent and inheritance. Genealogies, or parts of them, collected in this way were (and are) often presented in graphic form to illustrate how persons or groups are related to one another. Genealogical information was the bedrock of comparative studies of kinship, descent, and marriage that were characteristic of structural-functionalist analyses of social organization during the first half of the twentieth century. Genealogy was also the basis for terminological and semantic studies of kinship, and used by alliance theorists. The use of genealogy to map the kinship system was even more pervasive, however, becoming a routine part of anthropological practice during the second half of the twentieth century.

2. The Intellectual Context of Genealogy in Anthropology

The genealogical method, developed by W. H. R. Rivers during the Torres Strait Expedition of 1898, was formalized in an article published in 1910 and became an important entry in the 1912 *Notes and Queries in Anthropology* and subsequent editions. The nineteenth century intellectual context of this addition to the anthropological instrumentarium lay in a rich combination of biblical, philological, heraldic, and evolutionary ideas. Darwinian theories concerning the origin of species and the descent of man had recourse both to linguistic and graphic genealogical imagery. The familiarity of biblical and historical family trees made genealogy a perfect vehicle for translating new ideas about the relationships between all forms of life, and the slow transformation (or evolution) of these forms through the immense abysses of geological time that opened up through the discovery of prehistoric sites and remains during the nineteenth century. Graphic phylogenies, such as those drawn by Hae-

ckel (1874), expressed these evolutionary connections in the form of a tree.

Genealogies are central to the creation story in Genesis, which was a source of inspiration for late eighteenth and early nineteenth century philologists and ethnologists who sought to explain the relationship between the languages (and hence peoples) of the world in genealogical terms. Still earlier, culturally diverse objects brought from the new world to Europe from the sixteenth century onward, were accommodated by cabinets of curiosity within the category pagan, which already housed the classical world and barbarian Europe. Discussions about the origins of Incas and Aztecs, for example, revolved around their identity with lost tribes of Israel, and their being remotely descended from Noah (Shelton 1994).

Philological, biblical and Darwinian genealogical ideas influenced L. H. Morgan, author of *Systems of Consanguinity and Affinity of the Human Family* (1871), in his kinship and beaver research among the Objibwa around Marquette, which took place against a background of wilderness receding from American farms (Feeley-Harnik 1999). This was rather different from Darwin's invocation of genealogical imagery in his theory of evolution, which was filtered through the ideas of British animal breeders for whom keeping records was an important means of exercising control over procreation. The fact that the British aristocracy could trace their genealogies, just as they could trace the pedigrees of their favorite animals, such as thoroughbred horses, helped to distinguish them, in terms of descent and breeding, from middle- and working-class Britons.

Rivers borrowed and translated the English aristocratic notion of pedigree into the genealogical method. It became a means of organizing fieldwork and written ethnography, despite being later overshadowed by the long-term fieldwork practices of participant observation. The incorporation of pedigree into anthropological methodology for assimilating knowledge about others is a fascinating maneuver by a middle-class intellectual in the context of turn of the twentieth century Cambridge, which was still very much dominated by a few great families. Borrowing pedigree was simultaneously an assertion about being above this sort of hereditary class system, and yet harnessing the power of its connotations in the form of intellectual capital in the comparative study of mankind.

Genealogical thinking informed arrangements of material culture before the advent of public ethnographic museums, and continued to be influential after these collections had been systematized according to geographical provenance, or as evolutionary typologies. Museum collections played a significant role in the rise of disciplinary knowledge, including anthropology, during the nineteenth century. Rivers' early twentieth century reformulation of the genealogical method scaled down the vast abysses of nineteenth century ethnological time to the more modest tem-

poral requirements of the emerging ethnographic present. In this respect, he made a major contribution to modern ethnographic field techniques by developing an explicitly scientific method for collecting immaterial cultural data, and imbuing it with a distinctive material form.

3. *The Genealogical Method of Social Anthropological Inquiry*

Rivers' genealogical method moved from the concrete level of personal names to the abstract system of relationships. That abstract system was accessed through the terminological categories used by Ego to address and to refer to the persons named as relatives. Rivers explains the steps by which he collected a pedigree (he used the terms genealogy and pedigree interchangeably). First the descendants of the informant's 'real' (as opposed to classificatory) parents were listed, including previous or other marriages, and children listed in order of age, together with their marriages and offspring. The mother and father's respective pedigrees would be collected according to the same procedure. Rivers claimed that the most extensive pedigrees could be collected using only five terms of relationship: 'father', 'mother', 'child', 'husband', and 'wife.'

The fact that Rivers treated some categories—'brother', 'sister', 'cousin', 'uncle', and 'aunt'—as problematic, gave the entire method a lineal bias. Later commentators have argued that lineal relationships were seen as less ambiguous than either collateral or affinal ones (Barnard and Good 1984). Information concerning social condition, residence, totemic, and other group affiliation, along with other relevant facts, were to be recorded for each person on the pedigree as far as possible. An extraordinarily wide range of topics could be studied by collecting information in this way. Kinship was about the social recognition of certain genealogical relationships; and genealogy was assumed to be about biology.

Despite Malinowski's (1930) critique of the over-abstractness of kinship studies and his emphasis on the importance of long-term participant observation, the genealogical method proved to have remarkable staying power as a technique for structuring this kind of fieldwork.

4. *From Corporate Descent Groups to Critique*

One of the main uses of genealogies for Rivers and for subsequent (especially, but by no means exclusively, British) anthropologists lay in studying social organisation. This concern was sometimes connected with colonial administration, seeking to identify and use indigenous social structure for governmental purposes. Interest in descent, in studying ancestral

connections reckoned through the female (matrilineal), or male (patrilineal), or both (bilateral) lines, gradually emerged as a dominant theoretical concern by the 1940s. The juridico-political importance of unilineal descent, especially in the African societies to which anthropologists turned their attention after an earlier interest in the Pacific, led Fortes to distinguish between domestic (internal) and political (external) social domains.

Fortes' (1945) 'Paradigm of the Lineage System' diagram is a particularly striking use of a genealogical-style diagram to represent the all-male lineage structure. Despite Leach's (1961) critique of the paradigm, and his insistence on the ideological rather than empirical nature of corporate lineage groups, he himself made extensive use of genealogical evidence to back up his argument. The continuing recourse to genealogy, despite the eclipse of kinship from center stage after the 1960s, helped to maintain a notion of biology as something separate (like nature) from society, until almost the end of the twentieth century. It is therefore remarkable that Barnes (1967) refers to the collection of genealogies as part of the ethnographer's stock-in-trade, and genealogical diagrams as part of the 'minimum obligation' for making fieldwork intelligible. He did, however, recommend recording the way informant's talked about their kin before embarking on formal questioning.

The ethnologists of France, who set out to study kinship at home in the same way that their colleagues did abroad, developed the technique of talking family, as a complement to genealogy, much more explicitly than their Anglo-Saxonic counterparts. Firth and his colleagues' 1960s study of middle-class kinship in north London collected and made use of genealogies only to discover that these were often part of oral knowledge—as in peasant societies. Schneider, too, constituted genealogies in his studies of American kinship (Schneider 1968), although he engaged in more open-ended questioning.

Bourdieu called for a social history of the genealogical tool, with particular attention to the functions that have produced and reproduced the need for such an instrument (Bourdieu 1977). Schneider (1984) was to criticize 'the doctrine of the genealogical unity of mankind', arguing that all comparative studies of kinship are rooted in our own beliefs about biology, and the enormous value attributed to human sexual reproduction in (what he variously called) Western/European/American cultures. Schneider's critique proved to be the catalyst for a number of new directions taken in the study of kinship. Strathern developed the argument that nature no longer serves to ground the understanding of society, which means that interventions such as the new reproductive technologies are an integral part of both English kinship and wider (political) society. Bouquet (1993) analyzed the cultural specificity of the genealogical method, arguing that it was infused by English notions of

pedigree so specific that it bears little resemblance to the Portuguese notions of relatedness, which were used to problematize English middle-class kinship. Although the genealogical diagram, with its more ramifying cultural-historical roots, simplified and facilitated the divulgation of genealogy in anthropology, it did not cancel the complexities and contradictions of genealogical ideas even within Europe (Bouquet 1996). Franklin (1997) has elucidated the effects of new reproductive technologies on the meaning of genealogy, understood as life's progression and as a way of making knowledge, among prospective parents in northern England. She coined the notion of post-modern genealogy to capture the way technology has come to stand in for and improve upon the natural process of procreation, despite all its attendant uncertainties.

5. The Future of Genealogy

Critical reappraisal of the assumptions implicit to genealogy in anthropology is likely to play a significant role in future studies of kinship. Genealogical thinking is certainly not universal, which renders problematic the anthropological attempt to use genealogy as a standard. Yet anthropologists have already been struck by the growth of popular genealogy in different parts of the world, and through the Internet (Segalen and Michelat 1991), both of which are subjects deserving further empirical study. Indigenous peoples' uses of historically collected genealogies in establishing land claims, constitute another highly significant development: for example, among Torres Strait and Australian Aboriginal communities. The impact of new technologies on genealogy, and new uses to which genealogies collected in the past by anthropologists may be put in the future, are but two examples suggesting that despite recent critical work, genealogy has by no means had its day in anthropology.

See also: Ancestors, Anthropology of; Clan; Kinship in Anthropology

Bibliography

- Barnard A, Good A 1984 *Research Practices in the Study of Kinship*. Academic Press, London
- Barnes J A 1967 Genealogies. In: Epstein A L (ed.) *The Craft of Social Anthropology*. Social Science Paperbacks in association with Tavistock Publications, London, pp. 101–27
- Bouquet M 1993 *Reclaiming British Kinship. Portuguese Refractions of British Kinship Theory*. Manchester University Press, Manchester
- Bouquet M 1996 Family trees and their affinities. The visual imperative of the genealogical diagram. *Journal of the Royal Anthropological Institute* (NS) 2: 43–66
- Bourdieu P 1977 *Outline of a Theory of Practice*. Cambridge University Press, Cambridge, UK

- Feeley-Harnik G 1999 'Communities of Blood': The natural history of kinship in nineteenth-century America. *Comparative Studies in Society and History* 41: 215–62
- Fortes M 1945 *The Dynamics of Clanship among the Tallensi*. Oxford University Press, London
- Franklin S 1997 *Embodied Progress. A Cultural Account of Assisted Conception*. Routledge, London and New York
- Haeckel E 1874 *Anthropogenie oder Entwicklungsgeschichte des Menschen; Stammesgeschichte des Menschen*. Wilhelm Engelmann, Leipzig
- Leach E R 1961 *Pul Eliya, A Village in Ceylon: A Study of Land Tenure and Kinship*. Cambridge University Press, Cambridge, UK
- Malinowski B 1930 Kinship. *Man* 30: 19–29
- Morgan L H 1871 *System of Consanguinity and Affinity of the Human Family*. Smithsonian Institution, Washington, DC
- Rivers W H R 1910 The genealogical method of anthropological inquiry. *The Sociological Review* 3: 1–12
- Schneider D M 1968 *American Kinship, A Cultural Account*. University of Chicago Press, Chicago
- Schneider D M 1984 *A Critique of the Study of Kinship*. University of Michigan Press, Ann Arbor, MI
- Segalen M, Michelat C 1991 L'amour de la généalogie. In: Segalen M (ed.) *Jeux de Famille*. Presses de CNRS, Paris, pp. 193–208
- Shelton A A 1994 Cabinets of transgression: renaissance collections and the incorporation of the New World. In: Elsner J, Cardinal R (eds.) *The Cultures of Collecting*. Reaktion Books, London, pp. 177–202

M. Bouquet

Generalizability Theory

1. Overview

Generalizability theory (G theory) provides a framework for conceptualizing, investigating, and designing reliable observations. It was originally introduced by Cronbach and colleagues (1963, 1972) in response to limitations of the popular true-score-model of classical reliability theory (Spearman 1904, 1910). Classical reliability theory centers around the notion that each observation or test score has a single true score, belongs to one family of parallel observations, and yields a single reliability coefficient (Nunnally and Bernstein 1994). While this model may be reasonable for carefully equated parallel forms of tests, it is overly restrictive and often unrealistic in situations where, for instance, raters differ in the central tendency and variance, observations depend on the context in which they occur, and constructs are patently heterogeneous. Several writers argued in the 1950s and 1960s that the same observation could conceivably belong to more than one set of parallel tests and, thus, have more than one reliability coefficient (Cronbach et al. 1963, Guttman 1953, Tryon 1957). The fact that the internal consistency reliability estimate of a multidimensional measure tends to be low even when test–retest and alternate forms reliability estimates are high illustrates

the contradictions and limitations of the classical reliability model.

It was Cronbach's vision to reinterpret classical reliability theory as a theory regarding the adequacy with which one can generalize from a sample of observations to the universe of observations that was randomly sampled. G theory acknowledges that the reliability of an observation depends on the universe about which the investigator wants to draw inferences. Because a particular measure may, conceivably, be generalized to many different universes, a measure may vary in how reliably it permits inferences about these universes and, therefore, be associated with different reliability coefficients. G theory explicitly requires investigators to specify a universe of conditions over which they wish to generalize. The extent to which different conditions are associated with different observations has implications for designing dependable observations.

G theory not only reinterprets classical reliability theory, but also shows how the traditional distinction between reliability and validity can be overcome to design dependable observations. Traditionally, construct validity is concerned with drawing inferences about a latent construct based on observable measures, whereas reliability is concerned with drawing inferences about a true score from observations across parallel measures. In G theory, a universe, its facets, and the conditions for admissible observations are defined through careful construct explication, the traditional domain of validity theory. G theory defines observations as dependable if they permit accurate inferences about the latent construct (i.e., the universe of admissible observations) they are meant to represent. In so doing, G theory blurs the traditional distinction between reliability and validity. The use of the terms 'dependability' and 'generalizability' instead of 'reliability' reflects the interest in unifying reliability and validity. How to go about designing and investigating dependable measures is the subject of G theory.

2. Key Concepts

2.1 The Decision Maker

Measurements are designed with applications in mind, and their quality has to be examined in the context of their application. This context is symbolized by the 'decision maker' who is interested in measuring a particular construct in a particular population of persons, under a particular universe of conditions, and with particular types of decisions in mind.

2.2 Types of Decisions

The types of decisions for which a measurement instrument is to be used must be known before an

instrument that yields dependable observations can be designed. Two major types of decisions have to be distinguished. The first concerns decisions that rely on the relative ranking of individuals. Accepting the three top-scoring applicants for a position is an example of a decision that relies on the relative interpretation of scores. Measurements that emphasize the interpretation of the relative differences between scores are the domain of classical test theory with its traditional focus on interindividual differences.

The second major type of decision is based on the interpretation of the absolute level of scores. Mastery tests in education provide an example of a decision based on the absolute interpretation of a score. Similarly, pre-established admissions standards for college (e.g., minimum SAT score) or pass/fail standards on job selection tests rely on the interpretation of absolute score levels.

2.3 The Universe of Admissible Observations

A measurement consists of a sample from the universe of admissible observations. What defines a universe for a particular measurement is based on what the decision maker is willing to treat as interchangeable for the purposes of making a decision. Thus, a universe may consist of interchangeable three-letter syllables, symptoms of depression, or interviewers if the decision maker is willing to treat as interchangeable three-letter syllables, depressive symptoms, and interviewers. Or, a universe may consist of interchangeable time points for urine collection, survey questions about voting behavior, or situations for observing altruistic behaviors if the decision maker considers those aspects of measurement to be interchangeable.

2.4 Estimating a Person's Universe Score

Given a particular universe of admissible observations, a person's universe score ($\hat{\mu}_p$) can be defined as the average score based on all admissible observations (X_{pi}) of the universe of interest. The purpose of a measurement is to accurately estimate this universe score ($\hat{\mu}_p$) based on a sample of observations.

The degree to which a measurement is generalizable depends on how accurately the measurement permits us to estimate the universe score. The accuracy of an observation is captured in the variance associated with the different presumably interchangeable observations. The extent to which different items or different observers or different measurement occasions yield different observations determines how dependable a single specific observation is. In other words, if different items, observers, etc., yield similar observations, a single measurement may allow accurate inferences about the universe of observations. If different items, observers, occasions, etc., yield dis-

similar observations, inferences about the underlying universe based on a single observation may be questionable.

2.5 The Universe and its Facets

Universes can be simple or complex, homogenous or heterogeneous, and small or large depending on the construct of interest and the decision maker's interest in investigating different aspects or facets of generalizability. The terms 'facets' and 'conditions' are analogous to 'factors' and 'levels' in the literature on experimental design. A universe of observations is said to consist of one facet if the generalizability of observations regarding one source of variation in the universe, say arithmetic questions of varying difficulty, is at stake. For instance, a particular arithmetic test includes a sample of items, covering different addition, subtraction, multiplication, and division problems of one- and two-digit numbers. The decision maker is interested in general arithmetic achievement, and is indifferent to the particular questions on the test. In the one-facet design, one is interested in estimating the universe score of each person based on the sample of items included in the test. A universe is said to have two (or more) facets if the generalizability of observations regarding two (or more) sources of variation in the universe, say items and occasions, is at stake. For instance, to the extent that arithmetic achievement depends on the difficulty of the items and the test occasions (e.g., beginning or end of school year), the generalizability of a particular measure taken early in the school year on a sample of difficult items may be compromised.

2.6 Generalizability Studies

Generalizability studies (G studies) are conducted to investigate the relationship between an observed score and a universe score. They are designed to provide information regarding the sources of variability (i.e., facets of the universe) that influence the generalizability of observations. Because the dependability of a measure rests on its purpose (and thus the universe of admissible observations it is meant to represent), different G studies may be needed depending on the proposed application of the measure. Alternatively, a single G study has to anticipate the multiple uses of a measurement to provide as much information as possible about potentially important sources of variation.

In the above example, arithmetic items varying in difficulty is an obvious choice of a facet to be examined in a G study. In addition, decision makers may be interested in whether measures taken at different occasions, based on, for example, different response formats or different presentation modes yield different

observations. To examine these issues, a universe has to be defined more broadly to include items of different difficulty, response formats, presentation modes, and taken at different occasions. Having defined such a multifaceted universe, a G study has to be designed to investigate the contributions of these facets and their interactions to the overall variance in observations.

2.7 Decision Studies

G studies are designed to assess the dependability of a particular measurement technique. In contrast, decision studies (D studies) are designed to gather data based on which decisions about individuals are made. D studies rely on the evidence generated by G studies to design dependable measures for a particular decision and a particular set of facets about which a decision maker would like to generalize. The goal is to design a measure that samples sufficient numbers of instances from different facets of a universe of observations to yield a sufficiently dependable estimate of the universe score which the measure is meant to represent.

3. Illustrative Generalizability Design

3.1 Generalizability Design

For illustration, the statistical models and analyses for a one-facet G study will be presented. This most simple G design requires making observations (X_{pi}) on individual subjects ($1, 2, \dots, p, \dots, n_p$) under different conditions of a facet ($1, 2, \dots, i, \dots, n_i$), where subjects and conditions represent random samples from their respective universes. In a one-facet design, the universe of admissible observations consists of observations that differ with respect to one characteristic, such as item difficulty. The generalizability question is: How dependable are observations made under different conditions to draw inferences about the universe consisting of all conditions?

3.2 The Statistical Model

A one-factorial ANOVA model can be used to describe how the observed score of person p under condition i can be partitioned into a grand mean (μ), an effect for each person (π_p), an effect for each item (α_i), and a residual (ε_{pi}).

$$X_{pi} = \mu + \pi_p + \alpha_i + \varepsilon_{pi} \quad (1)$$

In Eqn. (1), the grand mean (μ) is a constant value for all persons, positioning the score for all persons on the particular scale of measurement. The person effect (π_p) describes the difference between a person's universe score and the grand mean ($\mu_p - \mu$), indicating inter-

individual differences in universe scores. The item effect (α_i) indicates the difference between an item's difficulty and the average item difficulty ($\mu_i - \mu$), reflecting the variability associated with the facet of error being investigated. The residual combines the influence of the person-by-item interaction, systematic sources of error other than items, and random events.

According to this model, specific observations X_{pi} vary due to differences between persons (i.e., person effects), differences between items (i.e., item effect), and other unaccounted for sources (i.e., residual). Accordingly (see Eqn. (2)), the variance of a set of observed scores (σ_{Xpi}^2) can be partitioned into components due to persons (σ_p^2), items (σ_i^2), and the residual (σ_e^2). The residual combines two sources of variance that cannot be distinguished from each other in this model. They are the person-by-item interaction (i.e., items vary in difficulty between persons; persons vary in their ranking on different items) and chance fluctuations in scores.

$$\sigma_{Xpi}^2 = \sigma_p^2 + \sigma_i^2 + \sigma_{pi,e}^2 \quad (2)$$

3.3 G study: Estimation of Variance Components

To estimate these variance components, a G study has to be conducted in which data are collected on a representative sample of persons and items. Variance estimates can be derived from the mean squares (MS) reported as part of the standard ANOVA printout of statistical analysis programs (see Eqns. (3)–(5)), in which persons are identified as between-subjects factor and items as within-subjects factor.

$$\hat{\sigma}_{pi,e}^2 = MS_{\text{residual}} \quad (3)$$

$$\hat{\sigma}_p^2 = (MS_p - MS_{\text{residual}})/n_i \quad (4)$$

$$\hat{\sigma}_i^2 = (MS_i - MS_{\text{residual}})/n_p \quad (5)$$

3.4 A Numerical Example

Table 1 presents the ANOVA summary table from a G study in which 1,140 persons rated the frequency with which they experienced 20 depressive symptoms. The table presents the numerical estimates and the percentage of total variance accounted for by persons, items, and the residual.

3.5 Interpretation of Variance Components

Variance components reveal information about how different sources of variability affect the response to a

Table 1

ANOVA mean squares (MS) and estimates of variance components for a one-facet G study

Source of variation	MS	Numerical estimate	Percentage of total variance
Persons (p)	1.5005	$\hat{\sigma}_p^2 = 0.0596$	15
Items (i)	26.4947	$\hat{\sigma}_i^2 = 0.0230$	6
Residual (pi, ϵ)	0.3089	$\hat{\sigma}_{pi,e}^2 = 0.3089$	79
Total	0.3902	$\hat{\sigma}_{Xpi}^2 = 0.3915$	100

single item. For instance, the variance component associated with items in the above example is 0.0230, accounting for 6 percent of the total variance. This variance component indicates how much the mean of any one randomly selected item is expected to vary from the mean of all items in the universe. Clearly, the magnitude of a variance component depends on how the items were scored (e.g., 1, 2, 3, 4 or 10, 20, 30, 40). In this example, items were scored 1, 2, 3, 4, where 1 indicates the absence of depressive symptoms and 4 indicates the symptoms were present all of the time. The average item mean (i.e., estimate of the mean of all items in the universe, $\hat{\mu}_i$) was 1.38 indicating that most respondents reported the absence of depressive symptoms. Taking the square-root of the variance component $\sqrt{0.023}$ yields 0.1517, indicating that we would expect approximately 95 percent of item means to fall between ± 0.2972 of the overall mean (i.e., $CI_{95\%} = [1.08; 1.68]$). Given that the possible range of item scores is 1 to 4, the expected range indicates that the item means in the universe of admissible items will mostly vary between 1 and 2. The variance component for persons is 0.0596, or approximately 15 percent of the total variance. This variance component estimates the universe score variance, that is, the between-person variability in the average item scores across all the admissible items of the universe. Based on the variance component, the standard deviation is 0.2441, indicating that 95 percent of universe scores are expected to fall between ± 0.4885 of the grand mean (1.0, 1.86).

Similar to the reliability coefficient in classical measurement theory, G theory yields a coefficient of generalizability (ρ^2) based on a particular G design. It is defined as the ratio of universe- and observed-score variances. In this example, the G coefficient for a single item is $\hat{\rho}^2 = \hat{\sigma}_p^2 / \hat{\sigma}_{Xpi}^2 = 0.1522$. Similarly, intra-class correlation coefficients can be determined. For one item it is $\alpha_{(1)} = (MS_p - MS_{\text{res}}) / (MS_p + (n_i - 1)MS_{\text{res}}) = 0.1617$, and for the set of 20 items it is $\alpha_{(20)} = (MS_p - MS_{\text{res}}) / MS_p = 0.7941$. The latter is also known as the Cronbach α coefficient or KR20 formula.

By far the largest variance component is due to the residual (i.e., 0.3089), accounting for approximately 79 percent of the total variance. This suggests that there are important sources of variance not accounted

for by differences between persons or differences in item difficulties (i.e., item means). Because the residual consists of the item-by-person interaction, this variance component reflects the fact that subjects indicate somewhat different levels of depressive symptoms depending on which item they are responding to. Based on the variance component, the standard deviation is 1.0893, indicating that the response to a specific item is expected to fall within ± 1.09 of a person's universe score in 95 percent of observations.

In summary, the results show that responses to a single item are mostly influenced by measurement error. Information about inter-individual differences make up only 15 percent of the total variance, and information about differences in test items make up only 6 percent of the total variance in item responses. Clearly, a single test item is of little use in estimating a person's universe score. Consequently, decision studies must sample many items to achieve more dependable estimates. Exactly how many items is the subject of D studies.

3.6 D study: Designing Generalizable Observations for Decisions

To conduct a D study, the relevant universe of observations and its facets must have been investigated in the G study. The type of decisions to be made must be known and the accuracy necessary to make the decisions specified. With this information at hand, the number of items from each facet of the universe can be determined such that dependable inferences can be drawn about the entire universe of observations at stake.

For instance, the decision maker in the previous example may want to design a measure that yields a 95 percent margin of error of ± 0.25 for the estimate of a person's universe score. That is, the decision maker wants to be 95 percent confident that a person's universe score is within ± 0.25 of the observed score. Based on this 95 percent margin of error (i.e., $0.25 = 1.96 * \sigma_{\text{error}}$), the error variance of the universe score can be determined: $\sigma_{\text{error}}^2 = (0.25/1.96)^2 = 0.0163$.

For decisions based on the relative ranking of scores, the required number of items is:

$$n'_i = \frac{\sigma_{\text{pl,e}}^2}{\sigma_{\text{error|relative}}^2} \quad (6)$$

For decision based on the absolute ranking of scores, the required number of items is:

$$n'_i = \frac{\sigma_i^2 + \sigma_{\text{pl,e}}^2}{\sigma_{\text{error|absolute}}^2} \quad (7)$$

Thus, for relative decisions, we need to sample $n'_i = 0.3089/0.0163 = 18.95$ or 19 items. And, for absolute

decisions, we need to sample $n'_i = (0.0230 + 0.3089)/0.0163 = 20.36$ or 21 items. In this example, very little variance is contributed by differences between items in their difficulties (i.e., σ_i^2), indicating that, regardless of which items are sampled, similar absolute scores will be obtained. Consequently, only two additional items have to be sampled to yield equally dependable estimates of universe scores for relative and absolute decisions.

4. Extensions

The principles presented here for a simple generalizability scenario can be extended to more complex designs consisting of two or more factors, which may be random or fixed, and which may be crossed or nested. The interested reader is referred to the works by Cronbach et al. (1972) and Shavelson and Webb (1991). G theory can also be applied to the analysis of score profiles, composites, and difference scores. Furthermore, extensions have been developed to estimate variance components via maximum likelihood, Bayesian, and covariance structure methods, and for studying the dependability of any facet of observation. The interested reader is referred to the work by Marcoulides (1996, 1999), Shavelson et al. (1989), Wittmann (1988), and Webb et al. (1983).

See also: Classical (Psychometric) Test Theory; Content Analysis; Content Validity; Generalization: Conceptions in the Social Sciences; Indicator: Methodology; Measurement, Representational Theory of; Reliability: Measurement; Test Bias

Bibliography

- Cronbach L J, Gleser G C, Nanda H, Rajaratnam N 1972 *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley, New York
- Cronbach L J, Nageswari R, Gleser G C 1963 Theory of generalizability: A liberation of reliability theory. *British Journal of Statistical Psychology* **16**: 137–63
- Guttman L 1953 A special review of Harold Gulliksen's theory of mental tests. *Psychometrika* **18**: 123–30
- Marcoulides G A 1996 Estimating variance components in generalizability theory: The covariance structure analysis approach. *Structural Equation Modeling* **3**: 290–9
- Marcoulides G A 1999 Generalizability theory: Picking up where the Rasch IRT model leaves off. In: Embretson S E, Hershberger S L (eds.) *The New Rules of Measurement: What Every Psychologist and Educator Should Know*. Lawrence Erlbaum, Mahwah, NJ, pp. 129–52
- Nunnally J C, Bernstein I H 1994 *Psychometric Theory*, 3rd edn. McGraw Hill, New York
- Shavelson R J, Webb N M 1981 Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology* **34**: 133–66

- Shavelson R J, Webb N M 1991 *Generalizability Theory: A Primer*. Sage, Newbury Park, CA
- Shavelson R J, Webb N M, Rowley G L 1989 Generalizability theory. *American Psychologist* **44**: 922–32
- Spearman C 1904 The proof and measurement of association between two things. *American Journal of Psychology* **15**: 72–101
- Spearman C 1910 Correlation calculated from faulty data. *British Journal of Psychology* **3**: 271–95
- Tryon R C 1957 Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin* **54**: 229–49
- Webb N M, Shavelson R J, Maddahian E 1983 Multivariate generalizability theory. *New directions in Testing and Measurement* **18**: 67–82
- Wittmann W W 1988 Multivariate reliability theory: Principles of symmetry and successful validation strategies. In: Nesselroade J R, Cattell R B (eds.) *Handbook of Multivariate Experimental Psychology*, 2nd edn. Plenum Press, New York, pp. 505–60

G. E. Matt

Generalization: Conceptions in the Social Sciences

All social scientists use observed particulars to draw conclusions about abstract entities. In measurement theory, the argument that ‘IQ is what an IQ test measures’ is discredited, and informed debates take place to interpret what the observed scores mean in general language. Some qualitative researchers are also reluctant to generalize beyond observables, but their language is replete with summarizing statements and they invite readers to construct theoretical meanings beyond the descriptions offered. In laboratory research, scientists often study local student samples, not to construct theories about them or about students in general, but to make inferences about humans. Some laboratory researchers use animals to substitute for humans, and their case for cross-species generalization has to depend on arguments—e.g., that cats and humans share the anatomical and physiological attributes central for understanding whatever aspect of vision is under study. This entry describes and evaluates many of the procedures for generalization that social scientists actually use.

Cronbach (1982) has perhaps the most comprehensive framework for understanding generalization. We use his theory, though slightly modifying it. He claims that social scientists typically draw conclusions about four not completely independent entities—human populations, physical settings, causes, and observables (including effects and moderators of cause–effect relationships). Time, understood as historical period, might also be added. Cronbach further contends that scientists work with samples, instances,

cases, exemplars, or particulars (terms he uses almost interchangeably), and that these represent study-specific populations, categories, entities, universes, classes, constructs, or categories (terms he also uses interchangeably). A major task of all scientists is to identify the class a given set of lower-order instances represents.

Cronbach also understands generalization in terms of inferences about classes that the sampling specifics manifestly do not represent. Such knowledge transfer is particularly salient, for instance, when animals are studied as proxies for humans or when laboratories substitute for other social settings. But transfer is also involved when research on American senior citizens, say, is used to extrapolate to Americans in mid-life or to seniors in Western Europe; or when findings about factories are extrapolated to schools on grounds that both are complex organizations. Extrapolation also pertains to causes and effects, as when studies of healthcare subsidies are applied to housing subsidies, or when a finding about a subsidy’s effect on hospital utilization is applied to effects on mental health. Cronbach believes that methods are needed both for justifying inferences from particulars to generals (as from samples to populations or from measures to constructs) and for justifying extrapolations beyond the categories studied.

His schema also includes subcategories—e.g., the racial groups within some human population or the components constituting a construct-like stress. Researchers routinely partition data both to test a relationship’s stability across the subcategories examined and to identify the conditions under which it varies. The more heterogeneous the subpopulations are over which a relationship holds, the greater its generality is usually assumed to be. In addition, the more specifically identified are the boundary conditionals limiting a relationship, the better its generality is understood. Campbell’s external validity (Campbell 1957) also emphasizes generalization across entities, but without restricting it to subtypes. For him, a relationship found among Austrians should be examined not just by partitioning Austrians into regional or generational subgroups, but also by examining the same relationship across other nations. Findings that dependably replicate across very different circumstances are the foundation of all general science.

Using the Cronbach framework, we structure this exposition into four parts: (a) methods based on formal sampling theory, as they are and are not used for generalizing to clearly designated universes of people, settings, times, causes, and effects; (b) purposive sampling methods, especially as they are used in validating constructs; (c) methods for assessing the range of circumstances over which a causal connection holds; and (d) methods for extrapolating knowledge to entities with different characteristics from those on which the knowledge was originally established.

This phrasing acknowledges that many social scientists use universe (and population) differently from construct, though the later Cronbach tends to use them almost interchangeably. In statistics, populations are ostensive; their elements are real and can be pointed to. But constructs are hypothetical and more obviously theory dependent. Moreover, the formal methods statisticians prefer when sampling elements from a population cannot be used with constructs, because the necessary enumeration and sampling of elements cannot be readily achieved with measures of abstract constructs. This is why Cook and Campbell (1979) used external validity to refer to people and settings and construct validity to refer to instances of more hypothetical causes and effects. However, the distinction is partly arbitrary. Constructs have constitutive elements theoretically specified as their components, and instances of any one construct vary in which components they incorporate. Moreover, human populations are not totally ostensive *per se*. Despite official definitions, there is still room to disagree about what being an Australian means: what about someone with an Australian passport who has always lived abroad, or the illegal immigrant who has always lived in Australia without a passport? There is room for legitimate debate about the definition of any human class and about whether certain individuals fit within that class. Also, whether human populations or abstract constructs are involved, the functional task is always the same: justifying inferences from particular cases to general categories. Hence, we see no fundamental reason for treating populations and constructs differently, though we do respect social science terminology and often use the one to refer to people and settings and the other to cause and effect variables as well as causal moderators.

1. Formal Sampling Theory

The best way to represent a population is to enumerate its members before selecting a random sample from that population. When properly implemented, this guarantees that the sample will formally represent the population within known limits of sampling error. The sample and population distributions will match on all measured and unmeasured attributes. This makes random sampling the centerpiece of survey research, often complemented by other sampling details. To reduce costs, cases can be clustered within a geographic area before being randomly selected. To ensure that subpopulations are reliably represented, stratification can also occur before random selection. Formal sampling theory provides an impeccable justification for generalization to well-designated populations. And methods for implementing are now so improved that its use in describing human populations and some settings is almost routine.

The problem with random selection is its own limited generality. It rarely applies to causes and other abstract constructs (or to time). One cannot readily imagine enumerating all instances of a specific causal agent before selecting a random sample of these instances for study. In practice, causal agents are purposively chosen because of a presumed correspondence between substantive theory and operational instance(s). Likewise, it is rare for all possible measures of a construct to be enumerated before some are randomly chosen. The selection of measures to represent a construct is again purposive. Also, it is hardly practical to sample historical periods at random for generalizing, if not to eternity, then at least to some meaningful interval. Detailed measures are not always available from the past; and anyway, many theories predict temporal contingency rather than constancy. This last justifies the purposive selection of time periods for the unique light each throws on some causal process.

Even when selection is at random, unambiguous inference still depends on large samples that are not always readily forthcoming. How often can we randomly sample 50 or 100 time points for assessment? Or 50 or 100 institutions, unless the aim is to describe a particular class of institutions? What about manipulating causal agents within institutions? What proportion of those asked to participate will agree to do so? What types are these volunteer organizations likely to be—the most typical or the most proud?

Random sampling rarely applies to assessing the conditions under which cause and effect are related. We know of only six experiments where individuals or households were randomly selected before being randomly assigned to a treatment (see Shadish et al. 2001). However, survey researchers often use random sampling to test causal hypotheses, using item responses to identify variation in a possible causal agent and to provide outcome data. But this is widely acknowledged to generate inferior causal conclusions compared to experimenter-controlled intervention. So how can causal relationships be better generalized via the random selection that surveys afford if these relationships are more uncertain because a survey was used? Moreover, surveys do not routinely use random selection for other generalization-relevant features of the research. For instance, data are typically collected at one or a few time points only, and then in one type of setting (usually the home),

Random selection was not developed for extrapolation, generalizing beyond the sampling particulars. So, it is not surprising that it is not relevant to this function. Indeed, none of the extrapolation methods we later detail depends on random sampling. Other premises are involved, as they are for generalizing causal relationships and for selecting measures to represent constructs. Thanks to its logical and empirical warrant, random selection is *the* method of choice for generalizing to populations of persons

and sometimes settings. But it is rarely relevant to other tasks with which a comprehensive theory of generalization must struggle.

2. Purposive Selection and Construct Validity

Purposive samples are selected without randomization and require only that the selected cases have certain prespecified characteristics. If these characteristics coincide with those of a population, then a match is achieved between the details in the population description and the details achieved in the sampling particulars. Unfortunately, such proximal matching cannot guarantee that the cases and categories are matched on all unmeasured characteristics, let alone perfectly matched on the measured ones. Purposive selection is less well theoretically justified than random selection. By itself, it justifies claims about category membership but not category representativeness.

Yet purposive selection is used throughout the social sciences. Even passionate advocates of random sampling measure behavior by first describing the construct to be assessed and then ensuring that measures of it include all attributes in the theoretical description. Empirical patternmatching between constructs and measures is the basis for generalization in this case, and substantive theory provides the template for choosing how to measure (or manipulate). The correspondence between construct and operational representation is usually multidimensional, with the most important attributes being the prototypical ones that most clearly define the construct's uniqueness (Brunswick 1956).

Such pattern matching (also called proximal similarity) is not restricted to measurement contexts. When random selection is infeasible it also applies to selecting persons, settings, and times. Thus, experimenters might partition their data into gender subgroups. The resulting subsamples clearly correspond to the categories 'male' and 'female,' as inferred from body shape or valid self-reports. Selecting this way ensures a proximal similarity between the class 'female' and the sampled instances of that class. But do they represent their gender if all attend the University of Berlin? In such a case, the gender specifics and the construct match on some prototypical attributes; but the sample is much more internally homogeneous than the population with respect to age, intelligence, geography, and culture, among other factors. Belonging in a class does not entail formally representing that class. Proximal similarity is necessary but not sufficient for inferring the representativeness of sampled instances. Sophisticated construct validation requires more.

One auxiliary principle concerns 'the heterogeneity of irrelevancies' (Campbell 1988), ensuring that the sample is not homogeneous on attributes that are irrelevant to the inference being made—such as age,

intelligence, and nation in the case of female students in Berlin. To make these conceptual irrelevancies heterogeneous requires adding samples of women of different ages, social classes, and countries. In measurement contexts, heterogeneity requires multiple measures of each construct or varying multiple instances of a treatment, provided that each instance has unique irrelevancies and that no irrelevancy is shared by all instances. It also requires testing the degree of triangulation achieved across the sources of presumed irrelevancy.

Another way to deal with homogeneous irrelevancies requires focused comparison. If female students in Berlin are compared with their male counterparts, this holds constant irrelevancies like country, language, and city of testing, while also varying gender. But such internal comparisons are not always feasible; and groups may differ in other unmeasured irrelevancies that affect responding—e.g., if female students in Berlin have higher intelligence than males on the average, or if data collection there occurs in chemistry classes that attract different kinds of females than males. Whatever the construct being targeted—gender, intelligence, or being Australian—purposive sampling requires critically examining whether the instances selected are more homogeneous than their presumed population.

Construct validation also requires 'discriminant validity' (Campbell and Fiske 1959), demonstrating that a relationship obtained with constructs labeled one way is not due to another construct that is correlated but nonetheless conceptually unique constructs. Thus, the validity of an anxiety measure requires not only showing that its measured attributes behave as hypothesized but also that they, as a whole, behave differently from conceptually related concepts like depression. Likewise, testing a hypothesis about how new computer firms are unique requires sampling more than just this type of firm. Also needed are firms making new products in other sectors. Relationships that hold across sectors can be generalized to firms in general, including computer firms. However, only relationships limited to computer firms help define what is unique about computer firms.

All sophisticated construct validity is about testing unique predictions entailed by the substantive theory of a population. We always validate interpretations of categories rather than categories themselves, even with human populations (Cronbach 1989). Lawyers might pronounce that an Australian is anyone eligible for an Australian passport. But this definition is not adamant. It includes individuals with Australian parents who have never lived in the country; it excludes illegal immigrants who have lived there for decades; and what about persons eligible for a passport who deliberately choose not to claim one despite living in the country for decades? No specification of a construct or population is beyond reasonable criticism and so each preferred interpretation must be defended.

Usually, the interpretation can be formulated as hypotheses about some combination of components, antecedents, consequences, or correlates. Construct validation requires specifying and testing the nearest to unique of these hypotheses. Only then can we be truly confident about generalizing from purposive samples and measures to their target constructs rather than to some other entity correlated with them. However, constructing the arguments needed is more laborious and uncertainty-riddled than random sampling. Still, most social scientists sample purposively most of the time, though they vary greatly in how much auxiliary information they collect after a modicum of proximal similarity has been achieved.

3. Generalizing Causal Connections

Much effort goes into testing causal hypotheses—statements that variation in one construct leads to variation in another and that no alternative cause can account for their covariation. Generalizing causal statements depends on methods other than formal sampling theory, given that experiments with both random selection and random assignment are rare and surveys provide causal tests that are flawed when compared to experiments. So, how is generalization achieved when the object to be generalized is a causal claim?

One way is when a population has recently been described using measures that are also available on a purposive sample of persons or sites. Then, any causal relationships obtained at the sample level can be weighted to the population level, using the empirical relationship between sample and population for this purpose. But such weighting only approximates what the population effect size would have been if the sample had been randomly selected. Correct inference depends on the sample level effect sizes not being affected by unmeasured sample characteristics that are not fully captured by the shared, measured variables. Still, we can anticipate that weighting procedures will be used more often in the future as methods for social experiments and estimating missing data advance. After all, survey researchers now describe human populations from biased sample data using basically the same weighting technique.

Sometimes, substantive theory is specific enough to detail the conditions under which a causal relationship holds. Then, a program of research can test explicit hypotheses about specific sources of population, setting, or time variation in the strength or direction of the causal relationship. This involves deliberate attempts to falsify the hypothesis that a causal connection is general (Gadenne 1976). Rejecting the hypothesis entails identifying specific causal moderators; failure to reject implies that the hypothesis of a general cause–effect relationship remains viable. A less formal variant of this describes common practice in the laboratory sciences. A causal claim is made and

then independently replicated, often under conditions that approximate the original ones. Successful replication leads to concluding that the causal relationship is provisionally true—that is, true until future failures to replicate indicate the need to invoke causal moderator variables. In both approaches, special emphasis is on identifying moderators that reduce the causal relationship to zero or that change its sign, as opposed to those that merely vary its size without affecting its sign.

Purposive sampling is also used in other less formal generalization probes. In clinical trials, hospitals recruit patients with a given disease and then adhere to the study's intervention and data collection protocols. Hospitals are basically chosen for their willingness to comply and for their heterogeneity with respect to, say, region of the country, predominant ethnic group served, or type of hospital sponsorship. Statistical power is often only adequate for cross-site analysis, suggesting that the goal is not to identify site-level main and interaction effects, but to see if the causal estimate exceeds zero despite heterogeneity in the hospitals studied. If it does, the causal effect is held to be general. If it doesn't, several alternative claims are viable and need to be tested, especially the following: there is no effect; the causal sign varies by hospital type and obscures a main effect; or the methods used cannot detect an effect of the size expected. Only increasing the sample of sites allows direct tests of which hospital types are involved in site-level interactions. Even so, relying on a purposive sample of sites that are heterogeneous on many attributes except for volunteering is only logically defensible if researchers know from other sources that no confounds are associated with all the hospitals being volunteers for the clinical trial.

Exploratory analyses often involve measuring many constructs from among a large array of possible causal modifiers in order to probe which ones affect the strength of a causal relationship. Without *a priori* hypotheses about modifiers, the approach can be time-consuming and run afoul of error rate problems due to the many tests conducted and the small sample sizes for some values on the moderator. Nonetheless, subcategory breakdowns are often used within experiments to probe how robust a relationship is across persons, settings, times, and measures. This can prevent the overgeneralization that occurs when researchers conclude that a relationship found with the entire sample holds in each subgroup within the sample. The entire sample allows one to conclude that the effect is found despite the internal heterogeneity achieved. It does not imply there is independent replication within any of the individual subgroups. The story is essentially the same when partitioning data from a single survey. With adequate sample sizes per group, one can probe the robustness of a relationship across the subgroups sampled. But only those potential conditional variables directly sampled can be assessed.

The best tests of causal conditionals come from synthesizing multiple studies on a topic rather than from subgroup breakdowns within a single study (Cooper and Hedges 1994). Experiments and surveys relevant to the same causal hypothesis accumulate and can be used in meta-analysis, the best-known form of synthesis. Meta-analysis combines data across multiple studies asking the same question, taking advantage of the between-study variation in measures, manipulations, populations of people, settings, and times. The dependent variable is each study's standardized effect size(s) and these are manipulated (a) to generate an average effect size across all the studies and all the irrelevant sources of heterogeneity they contain, and (b) to identify the specific contingency factors responsible for variation in the effect sizes.

What logic buttresses meta-analysis? It is not random sampling since individual studies are selected for their relevance to the hypothesis and the hope is to achieve a census of all published and unpublished studies. However, the more relevant target population is all studies that could be conducted on the topic, not all those done on it. In meta-analytic practice, generalization depends on the sample of achieved studies including a wide range of human populations, setting types, historical periods, and operationalizations of the cause and effect. Each study is first coded to label its sampling particulars in terms of theoretical constructs and to examine the sources of irrelevancy with which the particulars are and are not collectively imbued. Thus in a hospital experiment, hospitals might be described as being university-based, public, or private, and as being large or small. Patients might be described in terms of diagnosis, age, and gender mix, etc. The treatment might be described in terms of its components, duration, and manner of delivery; and the outcome might be described in similar terms. Then researchers typically test whether the same causal relationship holds despite these sources of sampled irrelevancy.

Sample sizes permitting, researchers also test whether the relationship holds across individually identified sources of potential irrelevancy so as to generalize to hospitals in general, people diagnosed with cancer, or whatever the target inferences might be for a particular study. Important here are the number and especially the heterogeneity of the types examined. Of course, some sources of variation might have *a priori* substantive meaning, as when university hospitals are a special target. Then one has to do a discriminant validity test and show that the effect size for such hospitals is different from that for other hospitals. Otherwise, there is nothing unique about university hospitals. So, meta-analysis depends on the very tests for proximal similarity, heterogeneous irrelevancy, and discriminant validity that are so central in construct validation. In addition, to achieve such tests multiple replication is needed across studies that may each be individually very narrow in the range of

people, settings, times, and cause and effect operations it includes. Central is the heterogeneity across studies and the pattern of results achieved despite this sampled heterogeneity.

4. *Extrapolating to Nonsampled Entities*

Some social scientists probably think it silly to use samples for generalizing to populations that are not even proximally similar to the samples used. So, many animal researchers abjure all talk of generalizing to humans. But their language is often anthropomorphic, and their research is usually funded because others think it is relevant to humans. So, elaborate reasoning has evolved over time to justify generalizing from animals to humans.

In this regard, consider the use of cats in visual research (Blake 1988). The rationale for their use is in two parts. The first is entirely pragmatic and stresses the conveniences of working with cats (or *Drosophila* in genetics research), usually because of cost and ethics. The second is theoretical and is based both on cats sharing some anatomical and physiological features with humans and on behavioral research showing that cats and humans behave similarly when confronted with similar visual stimuli. However, the empirical work necessary for showing such correspondences has also revealed that cats are much less like humans in some physical ways—especially features related to night vision—and that some behaviors are not replicated across species when the same stimulus is presented. So, a contingency theory has evolved of the conditions under which cats are better or worse substitutes for humans.

But when asked, honest vision researchers will avow that convenience is the paramount rationale for using cats. Would they rather work with chimpanzees? Undoubtedly, since the chimps' visual system is much closer to humans' than the cats'. But the expense of maintaining chimps precludes using them. Would the researchers rather use humans than chimps? Undoubtedly, if ethical research procedures are possible—which they often are not. However, now that new imaging technologies are available that allow observing the human eye reacting to planned stimuli in great micro-detail and real time, many vision researchers are turning to work with humans rather than cats because the pragmatic obstacles to using humans now hardly apply. Thus within the community of vision researchers the preference for generalization to humans via within-species proximal similarity is clear. So, too, is the distaste for arguments that depend on the (only partial) proximal similarity between cat and human on some vision-related attributes given their obvious dissimilarity in others. A similar story could be told for conducting research in the laboratory rather than in the field. Convenience is paramount, and the reasons given for preferring the lab often

contain elements of rationalization. But even so, extrapolation across species and settings is surely superior when there is an empirically defensible contingency theory of the conditions under which the extrapolation is better and worse warranted.

How is extrapolation otherwise achieved? Causal explanation is often held to play a major role, explaining how or why two or more things are related (Mackie 1974). Once an explanatory process has been identified, the argument is that it can be reproduced in other settings, with other kinds of people, at future times, and with unique ways of instantiating the process. Indeed, this is the major public rationale for basic science. Once one knows why fire happens, one can replicate it with new kinds of materials (say, some newly invented polymers), in new settings (an uninhabited desert island), and with people who have never seen fire before (so long as they can understand the explanatory recipe). The problem with this view is that full explanation is hard to conceptualize and achieve. It is difficult enough to use data for discriminating among explanatory claims even at one level of explanation, let alone for simultaneously investigating factors at the biological, individual, and social levels. Even so, causal explanation is the holy grail of science, not just for its aesthetic parsimony, but also for the utilitarian causal generalization it promotes in the form of extrapolation.

Extrapolation is also achieved through careful induction. If patient education reduces the length of hospital stays and does so with patients suffering from 25 different diagnoses, then it is likely to do so with the 26th. Of course, the logical warrant for such induction is limited, and seems more reasonable the more numerous and heterogeneous are the patient types already examined. If all 25 were different forms of cancer diagnosis, or involved patients aged between 25 and 40, or were restricted to men, this limited heterogeneity would reduce our confidence in any inductive-empirical extrapolation. Meta-analysis is particularly important in this regard, especially when it permits replication across multiple different sources of irrelevancy. A relationship that has proven to be robust in whatever form it was operationalized is more likely to re-occur when other ways of operationalizing its components are attempted—more likely, but not certain. We see here the value traditionally assigned in science to discovering both general dependencies and the contingencies that describe when a given relationship does and does not hold.

The final context for extrapolation relates to formal modeling. When subgroups can be arrayed along a single continuum, the functional form of a causal relationship can then be described. Is it stable? Is there a linear increase or decrease? If a pattern clearly emerges, one might want to argue that the strength of the relationship can be inferred for unobserved subpopulations with values that fall within the continuum range. Simple interpolation is involved here, while the

strength of the relationship for values beyond the continuum is inferred by extrapolating from the observed function. But comfort with such extrapolation depends on how distant the nonobserved subpopulation values are from the sampled ones.

Model-based extrapolation is always tricky, but especially so when the models are more complex. Consider much of macroeconomics. Models of a given economic system are developed from theory and past findings, and changes in parameter values are introduced to see how the system reacts to this perturbation. This is basically an extrapolation exercise to discover what would happen if some details of the economic system were not as they actually are. Can we extrapolate from what we think we know, in order to learn what would happen under novel circumstances? The validity of any extrapolation depends on the model being correctly specified with respect to the variables included, the relationships postulated between them, and the strengths of these relationships. Be wrong about any of these assumptions and any conclusions about the effects of the novel perturbation will be biased to some unknown extent. The assumption that the model is complete and accurate is difficult to test, though crucial to the credibility of the results (Glymour et al. 1987). Still, the creation of valid multivariate models of causal interdependency is another major goal of science, and it is not difficult to see why. It promises the extrapolation of results beyond whatever circumstances were studied in the past.

5. Conclusion

In the behavioral and social sciences at the beginning of the twenty-first century, theory and practice are most developed when individuals or households are sampled to describe a human population. Here, best practice requires randomly sampling units from the population. Theory and practice are also well developed for generalizing from a measure to an abstract construct or from an experimental treatment to a more general causal agent. Here, best practice requires an explicit theory of construct validity that necessarily invokes proximal similarity, but preferably also the heterogeneity of irrelevancies, discriminant validity, and causal explanation. However, the purposive selection and theory-linked methods needed for construct validation are not as formally well supported as random selection is. Theory and practice are even less developed for generalizing to settings and historical time periods and for identifying the conditions under which a causal relationship holds. However, various deliberate sampling and replication strategies are relevant here, including meta-analysis. At their best, such strategies borrow much of the logic underpinning sophisticated construct validation. Least well developed are theory and practice for extrapolating results to nonstudied circumstances. But even here,

understanding causal explanatory processes can permit re-creation of the same causal forces in novel settings. Also, there are now decades of research and reflection about using animals to extrapolate to humans and about using the laboratory to extrapolate to other social settings. In addition, substantive modeling is sometimes used to examine what would happen to a system if one or more of its parameters were changed. Of all the methods examined, only random sampling provides an impeccable formal rationale for generalization. But its reach is, alas, limited. It is a great irony that the best formal method for generalization in the behavioral and social sciences can be used for only a small part of the generalization tasks that practicing social scientists routinely face. Given this, practicing social scientists and theorists of method reflecting on these practices have cobbled together more complex methods of generalization that depend more heavily on substantive theory, purposive sampling, heterogeneity of irrelevancies and demonstrated replication and the identification of causal contingencies.

See also: Causation (Theories and Models): Conceptions in the Social Sciences; Causation: Physical, Mental, and Social; External Validity; Generalizability Theory; Internal Validity; Sample Surveys, History of; Sample Surveys: The Field

Bibliography

- Blake R 1988 Cat spatial vision. *Trends in Neuroscience*. **11**: 78–82
- Brunswik E 1956 *Perception and the Representative Design of Psychological Experiments*, 2nd edn. University of California Press, Berkeley, CA
- Campbell D T 1957 Factors relevant to the validity of experiments in social settings. *Psychological Bulletin* **54**: 297–312
- Campbell D T 1988 *Methodology and Epistemology for Social Science: Selected Papers* [ed. Overman E S]. University of Chicago Press, Chicago
- Campbell D T, Fiske D W 1959 Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* **56**: 81–105
- Cook T D 1993 A quasi-sampling theory of the generalization of causal relationships. In: Sechrest L, Scott A G (eds.) *New Directions for Program Evaluation: Understanding Causes and Generalizing about Them* (Vol. 57). Jossey-Bass, San Francisco
- Cook T D, Campbell D T 1979 *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Rand-McNally, Chicago
- Cooper H, Hedges L V (eds.) 1994 *The Handbook of Research Synthesis*. Russell Sage Foundation, New York
- Cronbach L J 1982 *Designing Evaluations of Educational and Social Programs*. Jossey-Bass, San Francisco
- Cronbach L J 1989 Construct validation after thirty years. In: Linn R L (ed.) *Intelligence: Measurement, Theory, and Public Policy*. University of Illinois Press, Urbana, IL, pp. 147–71
- Gadonne V *Die Gültigkeit psychologischer Untersuchungen*. Kohlhammer, Stuttgart
- Glymour C, Scheines R, Spirtes P, Kelly K 1987 *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling*. Academic Press, San Diego, CA

Mackie J L 1974 *The Cement of the Universe: A Study of Causation*. Oxford University Press, Oxford, UK

Shadish W R, Cook T D, Campbell D T 2001 *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton-Mifflin, Boston

T. D. Cook

Generation in Anthropology

In anthropology, the term generation has two primary meanings. In one of its senses, generation refers to a group of people who have lived through a time period together and have developed some kind of shared consciousness. Generation has been used even more commonly by anthropologists in its genealogical, kinship-related sense, to refer to the relationship between parents and their children, these children and their children, and so on. Using one or both of these senses of the term, anthropological interest in generation has focused on four primary topics, pertaining to the ways people across cultures organize and envision social relationships, identities, and the processes of human social-cultural life. This article will look at each approach in turn.

1. The Cultural Construction of Intergenerational Ties

Work on the cultural shaping of intergenerational ties focuses on generation in its genealogical sense as a descent relationship. Interest in intergenerational family relationships has existed since the beginnings of anthropology as a discipline, largely due to the analysis paid by anthropologists to kinship systems. Since about the 1970s, the majority of work on intergenerational ties has come out of the subfield of anthropology of aging, where one of the central research issues has surrounded the degree to which the elderly can expect to receive various forms of support from their younger relatives. Research on intergenerational ties, however, goes well beyond the (interesting and important) problem of how the elderly are cared for, to speak to broader conceptions about the nature of the family, the relationship between individuals, families and the state, and the social-moral order.

Lamb (2000) examines such problems through a focus on West Bengal, India. She finds that Bengalis perceive intergenerational family relations to entail long-term bonds of reciprocal indebtedness extending throughout life and even after death. Juniors provide care for their elderly parents, reconstruct relations with parents as ancestors after death, and ritually nourish these ancestors, as a means of repaying the tremendous 'debts' owed for producing and caring for them in infancy and childhood. What parents once

gave their children to produce and raise them—such as a body in birth, food, material goods, money, a home, forms of love, the cleaning of urine and excrement—children are expected to provide in return for their parents years later in old age and after death. (Children, for instance, construct new spirit-bodies for their parents as ancestors, parallel to the bodies parents create for their children in birth). Other anthropologists have found that relationships of lifetime intergenerational reciprocity, conceptualized and practiced in varying ways, are not uncommon cross-culturally (see, e.g., Cattell 1990).

The shaping of intergenerational ties can bear significance beyond the family as well, to implicate the social-moral-political conditions of a broader society and nation. Indian discourses, for instance, present close, multigenerational families as signifying a 'good' and 'Indian' society; whereas disintegrating generational ties—the break-up of joint families, uncared-for old people, independent daughters-in-law—are held to be paradigmatic signs of a widely degenerate 'modern,' postcolonial, and urbanizing era (Cohen 1998, Lamb 2000).

Some anthropologists argue that the shaping of intergenerational ties can be understood through scrutinizing surrounding cultural values. Simic (1983) suggests links between an American model of intergenerational relations and a broader cultural orientation toward self-fulfillment and individualism. Children in middle-class, Anglo-American households maintain their own rooms and possessions, and by about age 18 are expected to move out and live separately from their parents. Though parents often extend financial aid to their children after they have left home, Simic observes, 'it is almost always with the overt rationale of helping them get established as independent units, rather than creating a binding reciprocity of mutual dependence between the two generations.' In later life, this same ideology prevails: among the aged economic independence is a fundamental value; dependency is 'demeaning.'

In general, research has shown that in developing countries, intergenerational family relations tend to be closer and more mutually interdependent than in capitalist, industrial settings, although many important exceptions can be found (see Albert and Cattell 1994, Sokolovsky 1990). A wealth of careful, detailed, cross-cultural data on generational ties has been produced by anthropologists, revealing a profound variety in the ways intergenerational relationships are conceived, maintained, practiced, and given value.

2. Generation as an Element of Social Organization

Anthropologists have also examined the ways social and cultural groups employ some concept of generation to organize social relationships beyond the

family. Many cultures group people into generation-defined categories, taking some notion of genealogical succession within the family and extending it to think about broader social relations (Kertzer 1979). Examining generation in this sense, as a principle of social organization, has allowed especially social anthropologists and those concerned with social structure to address questions about the very ordering and continuity of societies (e.g., Baxter and Almagor 1978, Eisenstadt 1956, Fortes 1984).

In some societies, social generational ordering is highly formalized and elaborate. Rituals, variations in dress, forms of speech, permitted behaviors, specialized tasks, and places of residence may sharply distinguish persons of one generational group from another. The most highly institutionalized organizations of generations into social groupings, often referred to as generation-set systems, have been found among a number of East African groups (Baxter and Almagor 1978). In such groups, the generation model is extended so widely as to provide the basis for regulating social relationships in an entire society. Every male is a member of a named generation set, and the generation set a man is in is determined by the generation set his father is in.

The crucial rule in such generation-set systems is that successive genealogical generations of fathers and sons must be socially and politically separated. Some scholars suggest that this is a way of minimizing generational conflict and competition, for instance over wives, land, and political leadership, for members of different generation-sets generally cannot compete over the same resources and privileges. Generation-set systems have almost always been organized around males, and women may be tied to them peripherally as daughters, wives and mothers. Versions of such systems have also been found in Melanesia, among Australian aborigines, and among some Native American groups.

Even in societies where people are not organized into formal generation sets, some notion of generation as a principle of social organization is almost universally employed (see Kertzer 1979). For instance, the social roles or identities a person assumes may depend on whether he or she represents the junior or senior generation in a family. In many societies, a person's transition to social adulthood is linked to the birth of his or her children; and in India, the transition to a 'senior' stage in life and society comes about not so much through reaching a particular chronological age but through the marriage of one's children (Lamb 2000).

It is necessary to distinguish between generation and chronological age as principles of social organization and of cultural value. Most societies use some version of both in ordering social relations, but these categories are often confused or overlapping in social scientific analyses and among the people being studied (Kertzer 1979). Age, however, is distinct from gen-

eration: it has to do with the chronological passage of time, and rests on the formally isolated individual as the unit of reference whereas generation is family-generated. There are many situations in which relationships by generation conflict with relationships by age. For instance, among the Tallensi, no man, even if very advanced in age, can achieve jural autonomy until his father dies (Fortes 1984). In the US, social criticism may ensue if a man marries a woman 'young enough to be his daughter' (the woman as wife becomes the man's generation, but has the chronological age of his children's generation).

Several anthropologists have been intrigued by the fact that what is essentially an intrafamily principle of genealogical generations has been used so widely by human groups to think about social relations outside the family. People of diverse cultures, it seems, find families 'good to think with.'

3. Generational Identity as an Axis of Social Difference

A third way that anthropologists have approached the study of generation relates to what is probably the most common usage of the term in English: generation as a cohort of people who have lived through a time period together and share a sense of identification with others of their era. Sociologist Mannheim's classic essay, *The Problem of Generations* (1952), is often cited as the piece that popularized this sense of generational studies within the social sciences. To Mannheim, a generation is not simply a cohort—a group of people having a statistical factor, such as age or decade of birth, in common; a generation is made up of people who, born during the same time period, feel the impact of common historical events and cultural forces, and develop a shared consciousness or identity around these experiences. Anthropologists have expanded upon this sense of generation as a form of identity, using it to heed the multiple, competing perspectives, voices, and experiences that make up any culture. Along with gender, class, race, and ethnicity, generation can be a salient axis of social difference.

Newman's work (e.g., 1996) highlights such a sense of generation as an important theoretical concept for anthropology. In one study, she collected life-history data from divorced mothers in the US who had been pushed into downward mobility. The study did not begin as a generational story, yet she found that the life history texts bifurcated strongly by generation. Women who were born during the Depression were invested in different versions of American culture than those who were raised amidst the plenty of the 1950s and 1960s. 'It became clear,' Newman reflects, 'that the source of their views was a fairly tightly woven and self-conscious generational culture, or symbolic dialect,' which formed a grid through which their life circumstances were interpreted (1996).

Comaroff and Comaroff (1999) take a powerful look at generation as one of the dominant idioms through which youths in postcolonial South Africa frame their identities and place in the world. Widespread anxiety over the hardening materialities of life in an era of global capital translates frequently into bitter generational opposition. Many young black men blame their current incapacity to ensure a viable future for themselves on a perceived sinister, all-consuming aged elite, referred to as 'witches' or 'old ladies' (even when they are men). The Comaroffs cite one example of a group of youths pelting an old woman with sticks and stones. 'Why are you killing me, my grandchildren?' she wailed. Her assailants responded, 'Die, die you witch. We can't get work because of you!' The major lines of opposition in contemporary South Africa, the Comaroffs find, are not race or class, but generation, mediated by gender.

Such work reveals the potential salience of generation as a category of social experience and analysis. A society at any given moment may be perceived, by the people within it, and/or by the anthropologists studying it, as being crosscut by generations representing powerfully divergent values, beliefs, perspectives, and agendas.

4. Generation and Social Change

Finally, anthropologists have used generation as a means to explore social change, history, and time. Just as generation is one of the key images people around the world use to represent social change (for instance, by telling of how children are growing up in a world vastly different from that of their parents or grandparents), social theorists often think about processes of social or cultural change in generational terms. Mannheim speculated in the 1920s that cultural transformation comes about through the continuous emergence of new generations: as each generation comes into 'fresh contact' with its social and cultural heritage, it remodels what is found (1952). Anthropologist Fortes mused several decades later that such a generational model is an 'apt imagery by means of which to depict continuities and discontinuities in a community's social and cultural life over a stretch of time' (1984). It is within the framework of the succession of generations, Fortes explains, that all people acquire their cultural heritage. 'Mediated from earliest infancy by the natal family, the society readily takes on a generational form to the individual.' The concept of culture has always been defined by anthropologists as something that is 'transgenerational' at the same time that it is open to creative revision and change.

In the contemporary era of globalization and rapid social transformation, perceived 'gaps' between the lifeways of senior and junior generations may become, for many, ever more profound. Generation will be a

provocative place for anthropologists to look to apprehend the ways global processes of change take shape in particular places and lives.

5. *The Importance of Generation in Future Work*

Generation in anthropology has been a key topic in the study of aging, kinship and social organization, but it has otherwise remained rather undeveloped. That is, conceived as a form of subcultural identity, axis of social difference, or dimension of social change, generation has not taken center stage. The category of generation can be fruitfully exploded, however, to explore problems of current critical interest to the discipline—such as the ways processes of globalization play out in local lives, and diverse voices and identities make up particular cultural worlds.

See also: Age: Anthropological Aspects; Age, Sociology of; Age Stratification; Age Structure; Ancestors, Anthropology of; Cohort Analysis; Culture: Contemporary Views; Family, Anthropology of; Family as Institution; Family Theory; Economics of Intergenerational Relations; Generations in History; Generations, Relations Between; Generations, Sociology of; Globalization, Anthropology of; Youth Culture, Anthropology of; Youth Culture, Sociology of; Youth Movements

Bibliography

- Albert S M, Cattell M G 1994 *Old Age in Global Perspective: Cross-cultural and Cross-national Views*. Macmillan, New York
- Baxter P T W, Almagor U (eds.) 1978 *Age, Generation and Time: Some Features of East African Age Organisations*. Hurst, London
- Cattell M G 1990 Models of old age among the Samia of Kenya: Family support of the elderly. *Journal of Cross-cultural Gerontology* 5: 375–94
- Cohen L 1998 *No Aging in India: Alzheimer's, the Bad Family, and Other Modern Things*. University of California Press, Berkeley, CA
- Comaroff J, Comaroff J L 1999 Occult economies and the violence of abstraction: Notes from the South African postcolony. *American Ethnologist* 26: 279–303
- Eisenstadt S N 1956 *From Generation to Generation: Age Groups and Social Structure*. Free Press, Glencoe, IL
- Fortes M 1984 Age, generation and social structure. In: Kertzer D I, Keith J (eds.) *Age and Anthropological Theory*. Cornell University Press, Ithaca, NY
- Kertzer D I 1979 Generation and age in cross-cultural perspective. In: Riley M W, Abeles R P, Teitelbaum M S (eds.) *Proc. AAAS Selected Symposium 1979. Aging from Birth to Death: Volume II: Sociotemporal Perspectives*. Westview Press, Boulder, CO
- Lamb S 2000 *White Saris and Sweet Mangoes: Aging, Gender and Body in North India*. University of California Press, Berkeley, CA

Mannheim K 1952 (1927) The Problem of generations. In: Mannheim K (ed.) *Essays on the Sociology of Knowledge*. Oxford University Press, London

Newman K 1996 Ethnography, biography, and cultural history: Generational paradigms in human development. In: Jessor R, Colby A, Shweder R A (eds.) *Ethnography and Human Development: Context and Meaning in Social Inquiry*. University of Chicago Press, Chicago

Simic A 1983 Aging in the United States and Yugoslavia: Contrasting models of intergenerational relationships. In: Sokolovsky J (ed.) *Growing Old in Different Societies: Cross-cultural Perspectives*. Wadsworth, Belmont, CA

Sokolovsky J (ed.) 1990 *The Cultural Context of Aging: Worldwide Perspectives*. Bergin & Garvey, New York

S. Lamb

Generations in History

Generations—cohorts of humans born about the same time and growing up through the same slice of society and history—have played a more important role in the human past than is generally acknowledged. During the nineteenth and twentieth centuries, younger generations in particular played colorful, sometimes crucial roles in history. As awareness of generational differences and conflicts has grown, furthermore, an evolving theory of social generations has developed to explain the phenomenon. The writings of such thinkers as Karl Mannheim, José Ortega y Gasset, and a host of more recent scholars have analyzed the impact of changing sociocultural influences and traumatic historical events in forging social generations which may then themselves become powerful forces for change in history.

1. Generations: The Term and the Concept

The term generation has been commonly used in a variety of senses, some of them marginal, others of central relevance to the generational historian's task. This opening section will suggest six quite different uses of the term generation and of the more precisely focused concept of the social generation in history.

Genealogical accounts of the history of individual families commonly refer to each new wave of siblings as a new genealogical generation. Useful as it may be, however, this tight focus on an individual family evidently lacks the scope of developed generational studies, which typically seek to survey an age-defined subdivision of society as a whole.

Family history more broadly conceived may probe relations between biological generations—parents and children—across a whole society. Even this broader approach, however, still suggests that generational

relations consist exclusively of relations between biological parents and children—another unnecessary limitation on the range of generational history.

Prosopography, offering a group portrait of coevals who occupy the same social niche as politicians, writers, or business leaders, for instance, does offer a form of generational history. Emphasizing similarities of backgrounds, beliefs, careers, and achievements as well as generational contemporaneity, such analyses clearly move beyond the family into the larger sphere of society as a whole.

Another form of generational analysis divides up historical time itself into generational periods of about 20 years, the time believed to elapse between the birth of one biological generation and the next. While this approach may uncover significant differences of generational perspective, it may also impose a series of generations of unconvincingly equal length upon the alternating longueurs and crisis periods of human history.

Perhaps the most useful form of generational analysis has been that which sees social or historical generations as birth cohorts, people born about the same time. This sort of generational analysis moves beyond the family and the single professional group and recognizes that different social generations may be of varying lengths and of greater or lesser significance, depending on social or historical circumstances. Above all, cohort analysis provides a solid foundation for generational studies in a given society at a specific moment in its history.

2. Age Groups in History

Age groups—groups of coevals—have been recognized in most societies and have been elaborately institutionalized in some. Like other subsets of society such as race, class, and gender, age groups have sometimes been both active forces in social conflict and creative agents of social change.

2.1 Age Groups in Human Societies

Anthropologists have detected age groups (age-grades or age-sets) among many pre-urban peoples. Progress from one age grade to the next may be celebrated by elaborate rites of passage, and membership in a particular age group often brings both duties and privileges. Young men will be expected to fight wars, young men and women may look forward to marriage, and older people of both sexes may serve as decision-making elders or perform important ceremonial functions.

But age groups are detectable in more complex urban societies also. Apprentices in guilds, students in schools, and even conscript armies may segregate many of the young from most of their elders. At the

other end of life, retirement homes and communities separate older age groups from the rest of society.

In most pre-urban societies, younger and older people apparently tend to accept the same traditions and taboos. In more complex societies, however, conflict between age groups seems to have become considerably more common.

2.2 Conflict Between Age Groups

There is evidence of generational conflict in earlier urban societies than ours. Aristotle saw differences between youth and age at the heart of the social wars that frequently rent the fabric of Greek city-states. Medieval observers were dismayed by the riotous behavior of guild apprentices or the town and gown violence of university students. Many of the leaders of the European Reformation were young preachers, and historians have noted that the founding fathers of the USA were actually young men.

The clearest and most numerous examples of generational conflict, however, have been detected over the last two centuries. Nineteenth-century European commentators noticed the youth of the barricade-building ideological revolutionaries who challenged authority in the 1820s, 1830s, and 1840s. Twentieth-century observers around the world recognized the youthfulness of the ideologically inspired guerrillas, urban-terrorist groups, and revolutionary cadres who played such a central role in the history of Russia, China, and the developing societies of Asia, Africa, and Latin America, as well as in the most developed nations, in the 1900s.

3. The Generational Idea in History

Theoretical understanding of the generational idea has evolved in tandem with this apparently increasing historical significance of generational conflict. Earlier insights into the significance of generations in human society were brief and suggestive. The nineteenth and particularly the twentieth century, however, have seen much more elaborate and sophisticated development of the concept.

3.1 Early Awareness of Generations in Society

Old Testament genealogies and Homeric poetry reveal clear concern with the succession of human generations. The traditional father of history, Herodotus, posited 33-year generations in his *History of the Persian War*, and the Arab world historian Ibn Khaldun proposed a basic theory of generations in his *Muqaddimah: An Introduction to History*.

Only in the nineteenth century, however, did social thinkers of the caliber of Auguste Comte, John Stuart

Mill, and Wilhelm Dilthey begin to develop a coherent and convincing theory of social generations. Julián Marías (1949), a leading historian of the generational idea, emphasizes their major contributions. Comte, he points out, clearly recognized the succession of generations as an important causal factor in social change. Mill was aware that the generational concept must be expanded beyond a familial or genealogical framework. Dilthey attempted to analyze the basic structure and significance of social generations.

3.2 Classic Formulations of the Theory of Social Generations

The first half of the twentieth century, particularly the period between the two world wars, saw classic formulations of the generational idea by the Spanish philosopher José Ortega y Gasset (1923, 1933) and the German sociologist Karl Mannheim (1928). Ortega's account of the phenomenon focuses on the internal development of the typical social generation through four 15-year stages: childhood, youth, initiation to positions of power, and domination of society. This stages-of-life perspective has influenced many generationists since.

Mannheim's analysis emphasizes the roots of generational differences in changing social and cultural circumstance. Arising in different social locations (e.g., middle or working class, rural or urban environment) and shaped by new trends in society at large, new social generations develop different points of view and patterns of behavior from those prevalent among their elders. Mannheim also analyzes the structure of the social generation, suggesting that a given birth cohort may be subdivided into a number of distinct generational units, subgroups who work up the common generational experience in very different, often ideologically definable ways.

3.3 Development and Validation of the Generational Idea

In the wake of the global pandemic of youth revolt of the 1960s and earlier 1970s, finally, sociologists, political scientists, and other social scientists researched the realities and refined the theory of generational development and interaction as never before. Such theoretical and empirical students of the generational phenomenon as Vern L. Bengtson, Richard G. Braungart, T. Allen Lambert, Robert S. Laufer, and Marvin Rintala elaborated a convincing model of the social generation in history. These scholars probed the sociopsychological roots of generational consciousness in immense detail. They documented the persistence of generational identity long after the strong orientation toward one's youthful peer group has faded. They collected masses of contemporary evidence of the existence of generational conflict and of its influence on social change.

4. The Theory of Social Generations

As the twenty-first century begins, then, one form or another of generational analysis is used routinely, not only by historians, political scientists, sociologists, anthropologists, and critics of literature and the arts, but by journalists, politicians, marketing experts, and many others. These diverse commentators explore age groups just as they do groups defined by gender, class, or ethnicity. The brief exposition of a developed, essentially Mannheimian view of the sociohistorical generation that follows will try to build on this broad awareness.

4.1 Social Circumstances Shape Social Generations

A social generation, then, consists of a birth cohort, a group of individuals born in similar circumstances during the same span of years, and thus shaped by common social and cultural influences. The critical period in the shaping of a generation comes in its earlier years, when its members are most malleable and open to the influence of their sociocultural environment.

During the childhood of a birth cohort, common family structures, child-rearing methods, and educational practices begin this socialization process. Thus a patriarchal family structure, a child-rearing philosophy of spare the rod and spoil the child, and an education built around rote memorization may tend to produce one sort of younger generation. A more egalitarian family structure, permissive child-rearing methods, and schools committed to developing the individual propensities of each child may produce a very different social generation.

During the adolescence and youth of a birth cohort, the larger society in which it develops may also exert powerful developmental influences. Hierarchical or egalitarian relations in society as a whole, traditionalist or cosmopolitan cultures, authoritarian or democratic politics, all have an impact on young people just emerging into the world beyond the family. Within the society as a whole, furthermore, the particular generational location within which youth matures will also have its influence. Young people raised in working-class ghettos will grow up with a very different view of the world from that of their coevals raised in middle-class suburbs, even if both these locations are within the same larger society. At every level, the books young people read, the films they see, the songs they hear and sing, shape the beliefs and values of a social generation in ways we are only beginning to understand.

4.2 Historical Traumas Define Rebellious Generations

The socialization process thus shapes the members of each new generation from their earliest years well into

their twenties. To these influences of home, school, and society at large, however, historical study of the generational phenomenon would add a special emphasis on the traumatic historical event in the shaping of a generation.

Popular accounts of specific social generations have tended to stress the impact of history on the young. Wars may have a much more profound impact on the lives and attitudes of the young men who fight them than on their elders. A major economic depression can be far more traumatic for young people discouraged by poverty from starting careers and families. Such historic events, furthermore, frequently challenge the view of the world the young have absorbed from other sources, producing confused, disillusioned, angry, or rebellious younger generations.

As Lewis Feuer pointed out, traumatic events may thus undermine the authority of both traditional institutions and of the older generation that imposes them (Feuer 1969). In this way the mores and morals of the young may change drastically as they shrug off the prescriptions of a de-authorized older generation. Where anti-establishment ideas and attitudes are already present in the culture, such historic traumas may transform hitherto isolated advocates of change into leaders of large-scale generational revolts. Any historic dissonance between widely held social ideas and changing social realities may produce generational disillusionment. A traumatic historical event may multiply this effect, shaking the foundations of society and calling its deepest principles into question in the minds of the young.

4.3 *Generations Shape the Course of History*

It is sometimes difficult to identify the historic victories or social transformations brought about by generational revolts. Youthful rebellions like the European political revolutions of 1848 or the very different American and Chinese cultural revolutions of the 1960s may fail to achieve many of their declared goals. Yet the new commitments and the youthful activism of social generations who march to a different drummer do often make a difference.

When young black Americans, for instance, took to the streets in the early 1960s, laws and attitudes changed. As a result, doors to social advancement opened, and the lives of at least middle-class black Americans were significantly altered. Similarly, the dominantly young people who rallied to the banners of leaders like Vaclav Havel in Eastern Europe in 1989 toppled political and economic institutions, radically transforming life in such nations as Poland, the Czech Republic, East Germany, and Hungary.

The openness of younger birth cohorts to new cultural influences may also have transforming social consequences. Around 2000, the spread of capitalism, computer technology, and even such specific ideas as freedom of thought and expression, or gender equality

began to change the lifestyles of the young, and through them their societies, in many parts of the world.

5. *Problems and Potential of Generational Analysis*

Major theoretical problems remain for the generational historian. There does, however, appear to be real potential for future development and application of the generational approach to history.

5.1 *Problems of Generational Theory*

Like any broad attempt to explain society and history, the theory of social generations has generated its share of challenging questions.

How long is a generation, for instance? Thirty-three years as Herodotus declared, 15 years with Ortega, or even three years, as some modern demographic studies might suggest? Historically, the length of a significant social generation seems to depend on the pace of change. The more rapidly things change, the more quickly enough changes accumulate in the shaping social matrix to produce a significantly different generational point of view.

But how important are social generations? People unsympathetic to the approach often ask this question, frequently having a rival theoretical framework to suggest, such as social class. However, generational theory makes room for class differences in the concept of generational location, which Mannheim saw as an expanded form of class identity. There would in fact seem to be room for analyses of the same population in terms of its component races, classes, genders, or social generations—all important bases for analysis, none more fundamental than the others.

Finally, are social generations historical realities at all? Some observers have noted that even when the streets are full of young demonstrators, most young people seem to share their parents' views. Certainly the socialization process does succeed more often than not, even in eras rent by generational discord. In history, however, numerical majorities are not necessary to produce significant consequences. When socialization fails with even a critical minority of the young, the resulting conflict of generations can produce historic changes in society.

5.2 *Historical Research on Generational Themes*

The uses to which generational analysis has been put during the twentieth century are manifold, and new topics well suited to interpretation in terms of generations emerge with every passing decade.

The generational approach has been widely utilized by historians and critics of art and literature. Students

of such increasingly important features of modern history as immigration have also found generational analysis invaluable. And there is a vast literature on younger generations mobilized in student or youth movements, countercultures, or revolutionary cadres.

During the late twentieth century, the generational approach was also applied to changing attitudes and patterns of activity in many non-Western lands. India's crisis of values has been explored in a number of such studies. Several Chinese student movements, as well as the massive Maoist Cultural Revolution, have been analyzed as revolts of younger generations. And more research is needed on the generational factor in the postwar colonial liberation movements and in the jolting changes that have accompanied urbanization, modernization, and globalization in the developing world since.

In developed societies with rapidly aging populations, finally, the other end of the spectrum of age groups would seem to demand historical study. With the progress of gerontology and our growing awareness of the persistence of generational memory as background, historians might profitably probe generational relations in past cultures in which older age groups have enjoyed both prestige and real power.

5.3 Related Concepts

Both the historic reality and the historiographical concept of the generation are ancient ones, yet we are often surprised when generational differences surface once again in history. Given the likelihood that such challenges will continue to occur, the theory of social generations would seem to merit both further development and more vigorous application to the study of society and history.

See also: Age: Anthropological Aspects; Age, Sociology of; Age Stratification; Age Structure; Diffusion: Anthropological Aspects; Education: Anthropological Aspects; Elites, Anthropology of; Generation in Anthropology; Generations: Political; Generations, Relations Between; Generations, Sociology of; Life Course in History; Life Course: Sociological Aspects; Prosopography (Collective Biography); Right-wing Movements in the United States: Women and Gender; Social Movements and Gender; Youth Culture, Anthropology of; Youth Culture, Sociology of; Youth Movements; Youth Movements, History of

Bibliography

- Bengtson V L, Laufer R S (eds.) 1974 Youth, generation, and social change. *Journal of Social Issues* 30: issues 2 and 3
 Braungart R G, Braungart M M 1986 Life course and generational politics. *Annual Review of Sociology* 12: 205–31
 Eisenstadt S N 1956 *From Generation to Generation: Age Groups and Social Structure*. Basic Books, Glencoe, IL

- Esler A 1972 *Bombs Beards and Barricades: 150 Years of Youth in Revolt*. Stein and Day, New York
 Feuer L S 1969 *The Conflict of Generations: The Character and Significance of Student Movements*. Basic Books, New York
 Lambert T A 1972 Generations and change: Toward a theory of generations as a force in historical process. *Youth and Society* 4: 1, 21–45
 Le Vine V T 1968 Generational conflict in Africa: A paradigm. *Civilisations* 18: 399–418
 Mannheim K 1928 Das Problem der Generationen. *Kölner Vierteljahrshefte für Soziologie* 7: 157–85, 309–29 [English translation 1952 The problem of generations. In: *Essays on the Sociology of Knowledge*, Oxford University Press, New York, pp. 276–320]
 Marías J 1949 *El Método Histórico de las Generaciones*. Revista de Occidente, Madrid, Spain, [English translation 1970 *Generations: A Historical Method*. University of Alabama Press, University, AL]
 Ortega y Gasset J 1923 *El tema de Nuestro Tiempo*. Revista de Occidente, Madrid, Spain [English translation 1933 *The Modern Theme*. Norton, New York]
 Ortega y Gasset J 1933 *En torno a Galileo*. Revista de Occidente, Madrid, Spain [English translation 1958 *Man and Crisis*. Norton, New York]
 Peyre H 1948 *Les Générations Littéraires*. Boivin, Paris
 Pinder W 1926 *Das Problem der Generation in der Kunstgeschichte Europas*. Frankfurter Verlags-Anstalt, Berlin
 Rintala M 1963 A generation in politics: A definition. *Review of Politics* 25: 509–22
 Sinha D 1972 *The Mughal Syndrome: A Psychological Study of Intergenerational Differences*. McGraw-Hill, Bombay, India
 Spitzer A B 1973 The historical problem of generations. *American Historical Review* 78: 1353–85
 Wohl R 1979 *The Generation of 1914*. Harvard University Press, Cambridge, MA

A. Esler

Generations: Political

'Political generations' can be defined as politically distinctive age groups delimited by their years of birth or some major event, such as their year of entry into the electorate. Rintala (1979, p. 8) defines a political generation as 'a group of human beings who have undergone the same basic historical experiences during their formative years.' According to Braungart and Braungart (1989, p. 283) 'a political generation is an age group in history that mobilizes to work for political change,' although an age group could work to maintain the status quo and still be a political generation.

1. Generation and Cohort

Birth, aging, and death gradually transform the composition of any society. New generations supplant the old, but transforming the population may or may not lead to social and political change. The young may resemble the old in their political attitudes and

behaviors and, even when young people do differ from their elders, they may come to resemble them when they themselves age.

But although each generation perpetuates society, it may also change it. If the young could be isolated from their elders, it would be easier to create a new social order, and thus Plato argued that philosophers could establish the ideal polity more quickly if they expelled all citizens above the age of 10 from the republic. At times a generation may be too corrupt to establish a new political order. The Book of Numbers relates that the generation freed from Egypt, which again and again showed its faithlessness, was condemned to have its carcasses consumed in the wilderness until a new generation, born in freedom, was reared to conquer Canaan.

Although the concept of generations has biblical and Greek roots, the social scientific study of generations began with Mannheim's essay, 'The Problem of Generations' [1928] (1952).

According to Mannheim (1952), the concept of generations is both biological and sociological. Generations are produced through 'the biological rhythm in human existence,' but 'individuals who belong to the same generation, who share the same year of birth, are endowed, to that extent, with a common location in the historical dimension of the social process' (p. 290).

Mannheim (1952) maintains that historical experiences are most likely to influence persons in their late adolescence and early adulthood, postulating that 'personal experimentation with life' usually begins at about age 17. Whether young people are actually influenced by historical events depends on social and historical conditions. For example, young adults living in different cultures may experience quite different events, and, even within a single society, some persons may be relatively isolated from the mainstream of social events and thus be unaffected by major historical movements. Nor do all historical periods produce conditions that influence young adults in distinctive ways. Mannheim (1952) argues that distinctive generations are more likely to form in periods of rapid social change. But his basic point is that there is no guarantee that socially or politically distinctive generations will emerge: 'Whether a new *generation style* emerges every year, every thirty, every hundred years, or whether it emerges rhythmically at all, depends entirely on the trigger action of the social and cultural process' (p. 310).

Following Mannheim's (1952) cautions, it is unwise to label a group of people a political generation unless one can establish that they are actually distinctive in their attitudes and behaviors. Therefore, social scientists need a term describing persons born during the same years, but making no other assumptions about their distinctiveness. The term 'cohort' is used for this purpose. This term derives from Roman military terminology, since the cohort was a unit within a

Roman legion, but the term now has a much broader usage. Scholars usually define a cohort as a group of people born during the same years, but there are many other factors that can define a cohort: the entry class of a school, persons married during the same period, or persons who vote for the first time in the same election. But the most basic experiences are usually birth and death, and the former is usually used to define a cohort.

There is no definitive answer to how many years a cohort should include. In the 1960s, and even into the early 1970s, survey researchers often used arbitrary age categories to code years of birth, often grouping respondents into 10-year age categories in order to save one column of usage on the IBM punch cards that were used to store data. Scholars who want to work with these early data sets must employ these arbitrary age categories. With computerized storing techniques, survey age is usually coded by year of birth, and one has great flexibility in analyzing survey results. For example, a major analysis by Converse (1976) divides respondents into three-year age categories, and throughout their study of the US electorate, Miller and Shanks (1996) generally divide their results into cohorts that are four years in width.

2. Life Cycle, Generational, and Period Effects

Whenever young adults differ from their elders there are usually two plausible interpretations for these differences. A life cycle interpretation maintains that these differences result mainly from the inexperience of the young, and postulates that the young will become more like their elders when they age. A generational interpretation maintains that these differences result from the differences between the formative experiences of young adults and those of their elders.

One of the most extensive controversies in political science research has been the well-documented finding that young US citizens tend to have weaker party loyalties than their elders. This finding was reported in *The American Voter*. Lacking longitudinal data, Campbell and his colleagues (1960) relied upon an indirect test comparing the partisan loyalties of persons of the same age who have held their partisan loyalties for different lengths of time. Even though their test was indirect, Campbell and his colleagues (1960) reached strong conclusions that favored a life cycle formulation. The weak party loyalties of young Americans, they argued, resulted from their youth and political inexperience. These conclusions were strengthened through a cross-national analysis, re-analyzing Almond and Verba's study of the USA, Britain, West Germany, Italy, and Mexico. As Converse (1969) showed, *The American Voter* formulation could explain a process through which partisan strength among the electorate is held in equilibrium.

Even though older voters with strong party ties are dying and being replaced continuously by younger voters with weak partisan loyalties, the dynamics that increase partisan strength with age could prevent the erosion of partisan strength among the electorate. If, as Campbell and his colleagues (1960) argued, party loyalties among the electorate contribute to political stability, the life cycle explanation, if valid, had major implications for the political system.

Although *The American Voter* explanation was unchallenged for a decade, in the early 1970s scholars began to question it, based upon the finding that partisan loyalties did not become stronger as cohorts were tracked across the life cycle. Converse and his supporters acknowledged that partisan loyalties after 1964 did not increase among cohorts as they aged, but maintained that this resulted from the political dislocations of the 1960s and 1970s. In other words, the failure of partisanship to increase with age was held to result from the events of the post-1964 period. By the late 1980s at least 20 articles had focused on this controversy (for a summary, see Abramson 1989) and several more appeared in the 1990s. But the space devoted to this controversy has declined as scholars have increasingly recognized the intractable difficulties of distinguishing among life cycle, generational, and period effects. Instead, scholarship has turned to studying the impact of generational replacement upon political attitudes and behavior.

3. The Impact of Generational Replacement

There is no debate that the population is gradually transformed as a result of generational replacement. (Some scholars use the term 'cohort succession.') Replacement can play a major role in transforming the political attitudes and behaviors among mass publics. For example, during the 1930s and 1940s, replacement helped make the Democrats the majority party in the USA, and during the postwar years replacement was a major force eroding US party loyalties. Replacement benefited the British Labour party during the 1960s, but it contributed to the erosion of British partisan loyalties between 1964 and 1987. Between 1969 and 1988, replacement contributed to the decline of the Labor Party Alignment in Israel. During the postwar years, replacement may have contributed to a more partisan electorate in West Germany. Replacement may also help transform political values. For example, replacement helped to increase political tolerance in the USA between the mid-1950s and the mid-1970s. (For the citations to the sources documenting these trends, see Abramson and Inglehart 1995.) Women who came of age in the USA after the enactment of the Nineteenth Amendment played a major role in increasing women's turnout (Firebaugh and Chen 1995).

The most extensive analysis of the impact of generational replacement upon value change among mass publics has been conducted by Inglehart and his

colleagues (Abramson and Inglehart 1995) who focus on changing levels of postmaterialism. Respondents who give high priority to postmaterialist values tend to emphasize freedom and quality of life concerns, whereas materialists emphasize economic and physical security. In all of the European societies they study, young adults are more likely to have postmaterialist values than their elders. During the two decades that Abramson and Inglehart (1995) studied, about two-fifths of the population of these societies had been replaced. As Europeans with higher levels of materialism were dying out of these societies, and as they were being replaced continuously by Europeans with postmaterialist values, population turnover had affected the overall distribution of values in these societies. Analysis of a combined sample of the European societies that can be studied between 1970 and 1992 suggests that about nine-tenths of the overall shift toward postmaterialism resulted from replacement.

4. Methodological Issues and Future Research

Future research on the impact of generational replacement will face methodological problems. As cohorts are tracked over time, the older cohorts die. Samples of the very old become small and nonrepresentative and, eventually, dwindle away. The standardization procedures used by Abramson and Inglehart (1995) need to assign some value to cohorts that can no longer be tracked, but imputing any such value is problematic. The best one can do is to present estimates using a variety of assumptions, some of which have been suggested by Firebaugh (1989).

See also: Age, Sociology of; Age Stratification; Cohort Analysis; Generations in History; Generations, Relations Between; Mannheim, Karl (1893–1947); Youth Culture, Sociology of; Youth Movements

Bibliography

- Abramson P R 1989 Generations and political change in the United States. In: Braungart R G, Braungart M M (eds.) *Research in Political Sociology*. JAI Press, Greenwich, CT, pp. 235–80
- Abramson P R 1992 Of time and partisan instability in Britain. *British Journal of Political Science*. **22**(July): 381–95
- Abramson P R, Inglehart R 1995 *Value Change in Global Perspective*. University of Michigan Press, Ann Arbor, MI
- Braungart R G, Braungart M M 1989 Political generations. In: Braungart R G, Braungart M M (eds.) *Research in Political Sociology*. JAI Press, Greenwich, CT, pp. 281–319
- Campbell A, Converse P E, Miller W E, Stokes D E 1960 *The American Voter*. Wiley, New York
- Converse P E 1969 Of time and partisan stability. *Comparative Political Studies* **2**(July): 139–71
- Converse P E 1976 *The Dynamics of Party Support: Cohort-Analyzing Party Identification*. Sage, Beverly Hills, CA
- Firebaugh G 1989 Methods for estimating cohort replacement effects. In: Clogg C C (ed.) *Sociological Methodology*, 1989.

- American Sociological Association, Washington, DC, pp. 243–62
- Firebaugh G, Chen K 1995 Vote turnout of nineteenth amendment women: The enduring effects of disfranchisement. *American Journal of Sociology* **100**(January): 972–96
- Mannheim K 1952 The problem of generations. In: Kecskemeti P (ed.) *Essays on the Sociology of Knowledge*. Oxford University Press, New York, pp. 276–322
- Miller W E, Shanks J M 1996 *The New American Voter*. Harvard University Press, Cambridge, MA
- Rintala M 1979 *The Constitution of Silence: Essays on Generational Themes*. Greenwood, Westport, CT

P. R. Abramson

Generations, Relations Between

The concept of intergenerational relations has different meanings that are not always sufficiently distinguished. One of its aspects is the relations between age groups demarcated by demographic criteria. To the extent that these relationships are regulated by the institutions of the welfare state, one speaks of a 'generational contract.' The concept of intergenerational relations also encompasses the personal relations between the members of different generations. As a major focus of theory building and empirical research, this latter aspect—intergenerational relations within the family—is the subject of this article, which also touches on intergenerational relations with friends and acquaintances. Because the modern (or now perhaps postmodern) family is usually compared with the traditional family, a brief historical review is included as well.

1. Historical Patterns of Intergenerational Relations

Historical research shows that the rise of the bourgeois family in the late eighteenth century implied a change in family relations. This change is described in terms of growth in emotional closeness, intimacy, or affective individualization (Stone 1990). Although the finding of increasing affective attachment between parents and young children has become accepted as a matter of course, historical research has been less successful at changing popular beliefs about intergenerational relations between adult children and their parents in traditional society.

Many people, including a number of sociologists and especially politicians, cling to the 'misbelief' (Laslett 1991, p. 111) that adult children in premodern times lived with their parents under the same roof. Actually, types of families in preindustrial European society were quite diverse. The stem family, in which the members of the younger generation continued to

live with their parents, was quite common in south and southeast Europe. Another type of family was the *zadruga*, an arrangement whereby the families of brothers lived together. By contrast, the nuclear family was (and still is) the main type in north and northwest Europe (Burguière and Lebrun 1997). In that part of Europe, the adult children did not provide housing, financial support, and care for their old parents, though the opposite claim has often been made. The relatively few persons who did survive into old age had to work as long as they could. If they were poor, frail, or sick, they received a very low level of support from the local community. People who owned property or land or who had a business decided themselves when one of their children should succeed them. Monthly contributions were regulated by contract, with the heir having to provide the parents with either money or natural products.

The nature of intergenerational relations changed at the end of the eighteenth and beginning of the nineteenth century. The new emotional quality of the early parent-child relationship became encoded in new norms that defined the conduct of a good mother. Similarly, normative ideas began to spread about obligations that adult children had toward their parents. Until the seventeenth century, there seems to have been no legal provisions that committed children to support their old parents (Grimm 1899). However, Prussian common law (*Allgemeines Preussisches Landrecht*) of 1794, the most advanced codification of its time, ruled that children had to care for their parents in case of need (Paragraph 63). It was only after life spans (longevity) increased and after the industrial economy put old people out of work that the number of three-generation households increased (Ehmer 1982).

2. Intergenerational Relations at the Turn of the Second Millennium: Parents and Their Adult Children

Contrary to the assumptions in sociological theories of modernization, premodern norms of solidarity do not seem to have been eroded by the dynamics of modernization. The norm that adult children owe their parents support and affection has been spread by modern middle-class society. Simultaneously, however, the idea has emerged that children have a right to make decisions about their own lives. Parent-child relationships have thus come to be characterized by two sometimes incompatible dimensions: attachment and the pursuit of independence.

Additional changes in intergenerational relationships have been due to demographic factors in the twentieth century. Secular decline in the birth rate, which has been particularly apparent since the 1960s, has reduced the number of siblings, uncles, aunts, cousins, nephews, nieces (a family's 'horizontal' di-

mension) but has increased the number of generations within the family (the 'vertical' dimension of a family). In the 1990s, for example, 81 percent of all children ranging from 10 through 14 years of age in Germany had a living grandfather or grandmother (Federal Ministry 1998).

Research on intergenerational relations in late phases of the family cycle began in the 1960s, when it became apparent that the rising number of old and frequently frail people was going to confront society with a problem. In response to the structural 'isolation of the conjugal unit' (Parsons 1943, p. 27), researchers have sought evidence that relationships between old parents and their adult children are characterized by solidarity (Litwak 1965). Rosenmayr and Köckeis (1965) summarized their research results with the words 'intimacy at a distance' (p. 204), a formulation that became known worldwide and that still seems to hold for the relations between old parents and their adult children, at least in Western industrialized countries. Numerous studies have meanwhile shown that parents and children provide each other instrumental, emotional, and financial support (Attias-Donfut 1995, Schütze and Lang 1996, Silverstein and Bengtson 1997). Because intergenerational financial transfer flows from the older to the younger generation, it seems plausible that some of the public transfers to the elderly are channeled back to the young through the family (Attias-Donfut 1995, Kohli 1999).

Although researchers have particularly addressed questions about the stress and feelings of gratification experienced by middle-aged children who become caregivers for their parents, care is primarily given by spouses. Often, adult children become the caregivers only if one of the parental spouses is not available. Daughters are more likely than sons to assume this responsibility, and the daughters experience much greater stress, social isolation, and disruption in their daily lives than sons do (Horowitz 1985). The number of studies on conflict and consensus in parent-child relationships, particularly on caregiving relationships, is probably due to political concerns about whether the family will continue to perform this function or whether it will fall to state institutions. The focus on sociopolitical issues has led to a relative neglect of theoretical questions, such as the transformation of the parent-child relationship during the late phases of the family cycle.

3. Theoretical Perspectives on Intergenerational Relations

Erikson (1959) occasionally referred to the developmental task of overcoming dependency stemming from the period of childhood. In a similar vein Blenkner (1965) developed the construct of filial maturity, arguing that a transformation occurs in the

late phases of the family cycle irrespective of the quality of the early parent-child relationship. When the children cannot help but realize that their old parents are becoming frail, they experience a filial crisis because they recognize that the parents are having to rely completely on themselves. In the subsequent phase of filial maturity, the children acquire a new perspective, conceiving of their parents as individuals with their own rights, needs, and limitations. With this new perspective, there emerges a sense of responsibility that did not exist before. The concept of filial maturity may well be a way in which to integrate autonomy and attachment, moral obligation and affection, distance and closeness.

Taking Blenkner's concept of filial maturity as a model, Nydegger (1991) developed the concept of parental maturity. Parental maturity means that parents cease interference in their children's lives. Mature parents successfully cope with the fact that some hopes they had had for their children have not been met. They recognize whatever differences there may be between their values and those of their children. Mature parents also have learned to accept help and do not claim a degree of independence that is no longer realistic. Unfortunately, neither the concept of filial maturity nor that of parental maturity has been empirically researched (for an exception, see Marcoen 1995).

Another theoretical framework for the analysis of late parent-child relationships can be derived from the view of generations offered by Mannheim ([1928] 1952). Mannheim held that the members of different generations (cohorts) are located at different places in the historical process and that they therefore differ in the way they experience and cope with historical events. Mannheim applied this concept to classifications of age groups in the population, though not to family lineage. It seems legitimate, however, to ask how location in the historical process affects relations and mutual perceptions in the family. In a cross-cultural comparison, for instance, Attias-Donfut (1998) found much less convergence between value orientations and much greater emotional distance between old and younger generations (age groups) in Germany than in France. The author suggested that this generation gap in Germany may be a consequence of the Nazi period. The findings are consistent with those from the literature on family therapy, in which it is shown that Germany's Nazi past weighs heavily on relationships within the family as well (Stierlin 1981, Rosenthal 1997).

4. Intergenerational Relations beyond the Family

In both sociological theory and popular understanding, it is assumed that people choose friends on the basis of similarity, for friends mostly have similar social status, converge in their lifestyles, and belong to

the same sex and age group (Merton and Lazarsfeld 1954, Rosow 1967; for more recent literature see Filipp and Mayer 1999). Homogeneity of age, however, is viewed positively in relations between children or adolescents but negatively (as an indicator of discrimination) in relations that old people have. In reality, it is questionable whether the low number of friendships between older and younger people is evidence of discrimination. After all, friendship is usually between members of the same age group.

All in all, it is primarily through the family that the members of two, three, or even four generations maintain personal relations. Within these relations cultural traditions and family-specific experiences and narratives are passed on. At the same time, however, new forms of dealing with each other are created as well. There is no evidence to support suspicions of cultural pessimists that this balance between continuity and change is going to collapse.

See also: Age, Sociology of; Age Stratification; Age Structure; Cohort Analysis; Families as Educational Settings; Family as Institution; Family Processes; Generation in Anthropology; Generations in History; Generations, Sociology of; Kinship in Anthropology; Life Course: Sociological Aspects; Mannheim, Karl (1893–1947); Population Aging: Economic and Social Consequences; Tribe

Bibliography

- Attias-Donfut C (ed.) 1995 *Les solidarités entre générations. Vieillesse, famille, état*. Nathan, Paris
- Attias-Donfut C 1998 Generationenverhältnis und sozialer Wandel. In: Köcher R, Schild J (eds.) *Wertewandel in Deutschland und Frankreich*. Leske & Budrich, Opladen, Germany, pp. 173–205
- Blenkner M 1965 Social work and family relationships in later life with some thoughts on filial maturity. In: Shanas E, Streib G F (eds.) *Social Structure and the Family: Generational Relations*. Prentice-Hall, Englewood Cliffs, NJ, pp. 46–59
- Burguière A, Lebrun F 1997 Die Vielfalt der Familienmodelle in Europa. In: Burguière A, Klapisch-Zuber C, Segalen M, Zonabend F (eds.) *Geschichte der Familie*. Campus Verlag, Frankfurt am Main, Germany, pp. 13–118
- Ehmer J 1982 Zur Stellung alter Menschen in Haushalt und Familie. Thesen auf der Grundlage von quantitativen Quellen aus europäischen Städten seit dem 17. Jahrhundert. In: Konrad H (ed.) *Der alte Mensch in der Geschichte*. Verlag für Gesellschaftskritik, Vienna, pp. 62–103
- Erikson E H 1959 *Identity and the Life Cycle*. International Universities Press, New York
- Federal Ministry for Family, Senior Citizens, Women, and Youth 1998 *Zehnter Kinder- und Jugendbericht. Bericht über die Lebenssituation von Kindern und die Leistungen der Kinderhilfen in Deutschland*. Federal Ministry for Family, Senior Citizens, Women, and Youth, Bonn, Germany
- Filipp S-H, Mayer A-K 1999 *Bilder des Alters. Altersstereotype und die Beziehungen zwischen den Generationen*. Kohlhammer, Stuttgart, Germany
- Grimm J 1899 *Deutsche Rechtsalterthümer*, 4th edn. Dietrich'sche Verlagsbuchhandlung, Leipzig, Germany
- Horowitz A 1985 Sons and daughters as caregivers to older parents: Differences in role performance and consequences. *The Gerontologist* 25: 612–17
- Kohli M 1999 Private and public transfers between generations: Linking the family and the state. *European Societies* 1: 81–104
- Laslett P 1991 *A Fresh Map of Life: The Emergence of the Third Age*. Harvard University Press, Cambridge, MA
- Litwak E 1965 Extended kin relations in an industrial democratic society. In: Shanas E, Streib G F (eds.) *Social Structure and the Family: Generational Relations*. Prentice-Hall, Englewood Cliffs, NJ, pp. 290–323
- Mannheim K [1928] 1952 The problem of generations. In: Kecskemeti P (ed.) *Essays on the Sociology of Knowledge*. Routledge and Kegan Paul, London, pp. 276–320
- Marcoe A 1995 Filial maturity of middle-aged adult children in the context of parent care: Model and measures. *Journal of Adult Development* 2: 125–35
- Merton R, Lazarsfeld P 1954 Friendship as a social process. In: Berger M, Abel T, Page C (eds.) *Freedom and Control in Modern Society*. Van Nostrand, Princeton, NJ, pp. 18–66
- Nydegger C N 1991 The Development of paternal and filial maturity. In: Pillemer K, McCartney K (eds.) *Parent-Child Relations throughout Life*. Erlbaum, Hillsdale, NJ, pp. 93–111
- Parsons T 1943 The kinship system of the contemporary United States. *American Anthropologist* 45: 22–38
- Rosenmayr L, Köckeis E 1965 *Umwelt und Familie alter Menschen*. Luchterhand, Neuwied, Germany
- Rosenthal G (ed.) 1997 *Der Holocaust im Leben von drei Generationen. Familien von Überlebenden der Shoah und von Nazi-Tätern*. Psychosozial-Verlag, Giessen, Germany
- Rosow J 1967 *Social Integration of the Aged*. Free Press, New York
- Schütze Y, Lang F R 1996 Integration in family, kinship and friendship networks. In: Mollenkopf H (ed.) *Elderly People in Industrialised Societies: Social Integration in Old Age By or Despite Technology?* Edition Sigma, Berlin, pp. 25–40
- Silverstein M, Bengtson V L 1997 Intergenerational solidarity and the structure of adult child–parent relationships. *American Journal of Sociology* 103: 429–60
- Stierlin H 1981 The parent's Nazi past and the dialogue between the generations. *Family Process* 26: 379–90
- Stone L 1990 *The Family, Sex and Marriage in England 1500–1800*. Penguin Books, London

Y. Schütze

Generations, Sociology of

Generations constitute—just as the closely related categories of age, youth, adulthood and old age—seemingly universal categories in all human societies signifying the different stages of human life; the same is true of tensions between generations, as well as of concepts of the age group or set—in particular the youth group—as a special social category with its own distinct self-consciousness. But it is only in very specific civilizational and historical circumstances that these features come together. Only in special circumstances does historical change become articulated in terms of generational consciousness in general, and that of youth in particular.

It was Karl Mannheim—as well as J. Marias and J. Ortega y Gasset—who emphasized the apparent universality of generational consciousness or at least posed it as a major problem of research; but this tendency, at least as reported by them, was indeed very closely connected to rather specific civilizational traditions—the European ones—and to specific historical circumstances in the development of this civilization. Accordingly a central problem in the analysis of generations is under what conditions does historical change become articulated in terms of generational consciousness in general and that of youth in particular?

1. The Cultural Definitions of Age and of Youth

Generational consciousness is naturally connected with that of age and age differences, and therefore it might first of all be worthwhile to specify the general conditions under which specific age groups and generational tensions and consciousness develop in human societies.

Age differences in general and youth in particular constitute a universal phenomenon. They are first of all a biological phenomenon. Although the basic biological processes of maturation (within the limits set by such factors as relative longevity) are probably more or less similar in all human societies, the definition of age differences in general and of youth in particular is always cultural. Their cultural definition varies from society to society, at least in its details.

The cultural definitions of age and age differences contain several different yet complementary elements. First, these definitions often refer to the social division of labor in a society, to one of the criteria according to which people are assigned to various social positions and roles; for instance, in many societies certain roles—especially those of married men, full citizens, independent earners—are barred to young people, while others—such as certain military roles—are specifically allocated to them. Second, the cultural definition of age is one important constituent of a person's self-identity, his self-perception in terms of his own psychological needs and aspirations, his or her place in society, and the ultimate meaning of his life.

Within any such definition, the qualities of each age are evaluated according to their relation to some basic, primordial qualities, such as vigor, physical and sexual prowess, the ability to cope with material, social and supernatural environment, wisdom, experience, or divine inspiration. Different ages are seen in different societies as the embodiments of such qualities. These various qualities seem to unfold from one age to another, each age emphasizing some out of the whole panorama of such possible qualities.

The various qualities attributed to different ages are usually interconnected in many ways, constituting a series of sorts. The characteristics of any one 'age,'

therefore, cannot be fully understood except in relation to those of other ages. Only when taken together do these different 'ages' constitute the entire map of human possibilities and limitations as envisaged in different societies; and, as every individual usually must pass through them all, their complementariness and continuity (even if defined in discontinuous and contrasting terms) become strongly emphasized and articulated. The same holds true for the age definitions of the two sexes, although perhaps with a somewhat different meaning. Each age-span is often defined differently for either sex, and these definitions are usually related and complementary, as the 'sexual image' and identity always constitute basic elements of man's image in every society.

This close connection between different ages necessarily stresses the problem of transition from one point or stage in a person's life to another as a basic constituent of any cultural definition of an 'age.' The personal transition or temporal progress, or change, tends to become closely linked with the designation or perception of cosmic and societal time—with the rhythms of nature or history, the cycles of the seasons, the unfolding of some cosmic plan (whether cyclical, seasonal, or apocalyptic), or the destiny and development of society. The nature of this linkage often constitutes the focus round which an individual's personal identity becomes defined in cultural terms and through which personal experience, with its anguish, may be given some meaning in terms of cultural symbols and values.

2. Youth as a Liminal, Transitory Phase; Kernels of Generational Consciousness

The problems of age definition and the linkage of personal time and transition with cosmic time become very often accentuated in that age-span usually designated as youth—which constitute in most societies the most liminal of age definitions—the transition from childhood to full adult status, to full membership in the society. In this period the individual is no longer a child (especially from the physical and sexual point of view) and is ready to undertake many attributes of an adult and to fulfil adult roles. But he is not yet fully acknowledged as an adult, a full member of the society. Rather, he or she is being, as it were, 'prepared,' or is preparing himself for such adulthood.

This image of youth—the cultural definition of youth—contains all the crucial elements of any definition of age, usually in a specially articulated way, and often in confrontation with other age categories. Moreover, in this phase the problem of the linkage of the personal temporal transition with cosmic or societal time becomes extremely acute.

The fullest, the most articulate and definitive expression of these archetypal elements of youth and of their liminal potentials is best exemplified in the ritual

dramatization of the transition from adolescence to adulthood, such as the various *rites de passage* and ceremonies of initiation in primitive tribes and in ancient civilizations. In an even more diluted form, these elements may be found in the various spontaneous initiation ceremonies of the fraternities and youth groups in modern societies (Bettelheim 1954, Erikson 1963, pp. 54–93). Hence the youth period and its symbolism, and the liminal situations in which it becomes expressed, may become the focus both of continuity of the social order as well as of protest against it, and of its change.

The very emphasis on the transitory nature of this stage and on its essentially preparatory character, however, often create a somewhat paradoxical situation—namely an image of youth as the purest manifestation and repository of ultimate cultural and societal values of all the major human virtues and primordial qualities. Such an image may be reinforced by the identification of physical vigor in youth with a more general flowering of the cosmos or the society. It is this situation that contains the kernels of the development of strong generational consciousness especially around the image of youth, with strong potentialities of protest.

The social organization of any age and especially of the period of youth is constructed in all societies according to two major criteria. One is the extent to which age in general and youth in particular form a criterion for the allocation of roles in a society, whether in politics or in economic or cultural activity—aside from the family, of course, in which they always serve as such a criterion. The second is the extent to which within society there develop specific age groups, specific corporate organizations, composed of members of the same 'age,' such as youth movements or old men's clubs.

3. *The Social and Cultural Contexts of Different Patterns of Age and Generational Symbolism and Organization*

The importance of age as a criterion for allocating roles in a society is closely related to several major aspects of social organization and cultural orientation. The first aspect is the relative complexity of the division of labor. In general, the simpler the organization of the society, the more influential age will be as a criterion for allocating roles. Therefore, in parallel tribal or preliterate traditional societies, age and seniority constitute basic criteria for allocating social, economic, and political roles.

The second such aspect is that the major value orientations and symbols of a society, especially the extent to which they emphasize certain general qualities, or types of activity (such as physical vigor, the maintenance of cultural tradition, the achievement and maintenance of supernatural prowess) become

expressed and symbolized in specific age-spans. Thus for instance, in those societies in which military values and orientations prevail, young adulthood is emphasized as a very important—sometimes even as the most important—age, while those in which sedentary activities prevail emphasize older age. Similarly, within some traditional societies, where older people may serve as important 'storers' of collective memory, old age may be emphasized if it is seen as the most appropriate one for expressing major cultural values and symbols—for instance, the very upholding, symbolization, and perpetuation of a given cultural tradition.

The social and cultural conditions that determine the extent to which specific age groups and youth groups develop differ from the conditions that determine the extent to which age serves as a criterion for the allocation of roles. At the same time, the two kinds of conditions may be, in some situations or societies, closely related. Age groups in general and youth groups in particular tend to arise in those societies in which the family or kinship unit cannot ensure (it may even impede) the attainment of full social and political status on the part of its members. These conditions appear especially, although not uniquely, in societies in which family or kinship groups do not constitute the basic units of the social division of labor. Membership in the total society (citizenship) is not defined in terms of belonging to any such family, kinship group, or estate, nor is it mediated by such a group.

The emergence of distinct age and generational organization and consciousness, as Meyer Fortes has shown in one of his last papers, lies in the political realm, in the search and definition of citizenship beyond the family sphere and not through the family, while generational differences are mostly elaborated in the structure of the family and kinship unit.

4. *Age Organization and Generational Consciousness in Tribal, Archaic, and Historical Societies*

The type of social division of labor and of political and cultural symbolism that generates such conditions is found in varying degrees in different societies—primitive, historical, or modern. Such a division of labor has existed, for example, in several primitive tribes: in Africa, among the chiefless (segmentary) tribes of Nandi, Masai, or Kipigis; in the village communities of Yako and Ibo; in the more centralized kingdoms of the Zulu and Swazi; and among some of the Indian tribes of the Plains, as well as among some South American and Indian tribes. Such division of labor likewise existed to some extent in several historical societies (especially in city states such as Athens or Rome), although most great historical civilizations were characterized mainly by a more hierarchical and ascriptive system of the division of

labor, in which there were greater continuity and harmony between the family and kinship groups, and the broader institutional contexts.

The fullest development of this type of social division of labor is to be found in modern, especially industrial, societies in which inclusive membership is usually based on universal criteria. In these societies the family does not constitute a basic unit of division of labor, especially not in production and distribution, and even in the sphere of consumption its functions become more limited; occupations are not transmitted through heredity. Similarly, the family or kinship group does not constitute a basic unit of political or ritual activities. Moreover, the general scope of the activities of the family has been continuously diminishing, while various specialized agencies tend to take over its functions in the fields of education and recreation.

At the same time in most modern—as well as in some historical—societies, there developed a certain dissociation between organizations of groups based on age and the symbolic distinction between different age categories. It is such dissociation that constitutes the background for the potentialities for the crystallization of specific generational consciousness formed around the image of youth and in which there develops a strong articulation of the dialectics between, on the one hand, the ‘rebellious’ protest-oriented aspects of youth organization and generational consciousness, and, on the other hand, continuity-oriented aspects (Bettelheim 1954, Erikson 1963, pp. 54–93).

5. Generational and Youth Symbolism and Consciousness in Modern Societies

5.1 Generational Consciousness in Modern Society

The generational consciousness, about which Mannheim, Marias, and Ortega y Gasset and many other modern writers wrote, has developed against the background analyzed above. Such consciousness is indeed connected, in a rather paradoxical way, with the fact that in modern societies the development of specific youth organizations is linked with the weakening of the importance of age in general and youth in particular as definite criteria for the allocation of roles in society become weakened. Yet this very weakening of the importance of age, as a criterion of allocation of roles, has been connected with intensive development of youth groups and movements—and lately also, perhaps paradoxically, among ‘retired’ people. Indeed, it is very interesting to note that in some of the contemporary industrial societies there tends to develop a distinct organization based on age as well as generational consciousness that are again coming together at the other end of the life-span—in

old age—among the older people, that is, those who have retired from work.

It is because of these developments—some incipient phases of which can be also discerned in different archaic and historical societies, but which have not been studied adequately (Nash 1978)—that in modern societies there also arises the possibility of the development of very strong generational consciousness in general, and of the articulation of such consciousness in terms of youth in particular. It denotes the general development of the self-consciousness of generations, which develops in the public realm—whether the realm of politics or of culture—and it contains very strong potentialities of social, political, or cultural protest.

6. Historical Change, Revolutionary Orientations, Generational Consciousness in Modern Youth Movements

The development in modern societies of such a dissociation between generations, and the concomitant possibility of the development of a special type of a distinct generational consciousness, are due not only to the complex division of labor that characterizes these societies. They are also closely connected to the relatively quick tempo of historical change, and above all to the consciousness of such change. Such consciousness is often connected with changes rooted in the revolutionary orientations which constituted important components of the construction of modern civilization, cultural values, and orientations. These revolutionary orientations were most fully articulated by the major social movements that have developed as part and parcel of modern life, of the constructions of modern social and political order and of struggle around such construction. The possibility of the development of distinct generational consciousness has been enhanced in modern societies by the occurrence of rapid historical changes or ruptures, the consciousness of which is shared by members of the same age group or age cohort.

But while historical change is particularly rapid and the consciousness of such change very widespread in modern societies, it does not always give rise to fully articulated generational consciousness—although to some degree such consciousness is probably rather common in modern societies and is not necessarily confined to the symbolism and ideology of youth. It is only in rather special historical circumstances and situations—like wars, changes of regime, and the like, especially of events which entail strong ruptures—that such generational consciousness has become articulated in modern societies in terms of youth symbolism.

Among such situations in modern societies two are probably of special importance. One was developed at the end of the nineteenth and the beginning of the twentieth century, and found its fullest expression in

the free German youth movements, though it was also present in the less spectacular youth movements of Central Europe and to some extent in the youth culture of various more flexible youth groups that developed in most modern societies. In all these movements very strong attempts have been made to overcome the dislocation between personal transition and societal and cultural time. It was in them that dreams of a new life, a new society, freedom and spontaneity, a new humanity, and aspirations to social and cultural change found utterance. It was in these youth movements that the forging of youth's new social identity became closely connected with the development of new symbols of collective identity or new social-cultural symbols and meanings, which thus become foci of protest aimed at the reconstruction of society and culture.

While some of these orientations could be found in many of the youth groups that developed in most modern societies—be they in England or the United States—they developed most fully in situations where these groups became closely interconnected with major social movements oriented at the reconstruction of the centers and symbols of their respective societies. This was the case above all in the major nationalistic and revolutionary movements in Central and Eastern Europe, in Italy, Germany, Hungary, and Russia, or in the Zionist movement, where these orientations and student movements played a crucial role.

One of the major characteristics of many, probably most, of these movements was that in their ideology the symbolism of youth was very closely connected with that of the construction of new modern, especially national, collectivities that were being constructed at the period, and among the aims of these youth movements the reconstruction of such collectivities—national, social, or cultural—was central. These movements aimed at changing many aspects of the social and cultural life of their respective societies, to reconstruct the centers and premises of their respective societies in terms of some of the major revolutionary orientations of modernity. They depicted the present in a rather shabby light; they censured it for its materialism, exploitation, restrictive views, and lack of opportunity for self-fulfilment and creativity. At the same time they held out hope for the future—seemingly the not very far-off future—when both self-fulfilment and collective fulfilment could be achieved and the materialistic civilization of the adult world shaken off. They tried to appeal to youth to forge its own self-identity in terms of these new collective symbols, and this is why they have been so attractive to youth, for whom they provided a set of symbols, hopes, and aims to which to direct its activities.

The influence of such movements and orientations was very great, not only political and educational but also in many cultural arenas. In a way, education itself has tended to become a social movement. Many schools and universities, and many teachers, have

been among the most important bearers of collective values. The expansion of education itself is often seen as a means by which a new epoch might be ushered in.

7. *The Routinization of Revolutionary Movements; the Demystification of the World; and Youth and Student Rebellion of the 1960s*

The second historical situation in modern societies in which generational consciousness, couched in the symbolism of youth, became very prominent in the new social movements burgeoned in the late 1960s and early 1970s, and became most fully manifest in the various student rebellions and movements of the time.

The background to these new developments was a combination of the growing institutionalization of many of the collective values and symbols of the social, national, and youth movements discussed above, and the continuous expansion of education and changes in occupational structure. These developments had several important repercussions on the structuration of the life spans of youth.

The area of social life that the specific youth or student culture encompasses has tended to expand continuously. First, it has extended over a long period of life, reaching, through the impact of the extension of higher education, to what before was seen as early adulthood. Second, it tends more and more to include areas of work, leisure time activity, and many interpersonal relations. Third, the potential and actual autonomy of these groups, and the possibility their members have for direct access to work, marriage, and family life, to political rights, and to consumption, have greatly increased, while their dependence on adults has greatly decreased.

Paradoxically enough, the growing direct access of young people to various areas of life has given rise to a growing insecurity of status and self-identity and to a growing ambiguity of adult roles. This insecurity and ambiguity has been enhanced, first, by the prolongation of the time-span between biological and social maturity and by the extension of the number of years spent in 'preparatory' (educational) institutions, and, second, by the growing dissociation between the values of these institutions and the future—especially in terms of occupational and parental roles—of those participating in them. A third factor is that for a long period many 'young' people may as yet have no clear occupational roles or responsibilities and may be dependent on their parents or public institutions for their economic needs, while at the same time they constitute an important economic force as consumers and certainly exercise political rights. In turn this situation may become intensified or aggravated by the growing demographic preponderance of the 'young' in the whole population and by the increasing possibilities of ecological mobility.

All these developments gave rise to new types of generational and youth consciousness—a new combination between the symbolism of youth and collective symbolism. The older classical movements of protest and youth movements tended to assume that the framework and centers of the nation-state constituted the major cultural and social reference points of personal identity and of the charismatic orientations to some sociocultural orders, and that the major task before them was to facilitate the access of broader strata to these centers and full participation in them and possible changes in them. In contrast, the new movements of protest have been characterized by their scepticism towards these new modern centers and by lack of commitments to them—all of which denoted important aspects of what has been called by Weber the demystification of the world.

The foci of protest promulgated by these movements shifted from demands for greater participation in national political centers or from attempts to influence their socioeconomic politics to new directions. The most important of these directions were: first, attempts to 'disrobe' these centers of their charismatic legitimacy and perhaps of any legitimacy at all; second, continuous searches for new loci of meaningful participation beyond these centers and the concomitant attempts to create new centers which would be independent of the existing center; third, attempts to formulate the patterns of participation in their centers not so much in sociopolitical or economic terms but more through symbols of primordial or of direct social participation.

These developments have often been connected with the breakdown of consciousness of historical continuity and of the conception of the future in terms of such continuity. The image of the societal future has become flattened and has been deprived of its allure. No sharp ideological discontinuity has been conceived between present and future.

Through these developments, youth has been, as it were, robbed of the full experience of the dramatic transition from adolescence to adulthood and of the dramatization of the difference between present and future. Their own changing personal future has become dissociated from any changes in their societies or in cultural activities and values. As a result of all these processes, the possibility of linking personal transition to societal and cosmic time—which was so strongly emphasized in the youth movements and observable also to some extent even in the earlier, looser youth culture—has become greatly weakened.

One of the most important results of these developments has been the development of a new configuration of youth symbolism and generational consciousness—the major characteristic of which was the development of strong antinomian intellectual tendencies as a central component of such symbolism and consciousness.

It is not just the contents of these antinomian

tendencies that are new, but the convergence of these contents with the spread and location of these movements. Of special importance in this context is the fact that this type of protest was not confined only to small, closed intellectual groups but involved widespread circles of novices and aspirants to intellectual status. Given the expansion of the modern educational system and the parallel effects of the spread of mass-communication media, a very large part of the educated public have become aspirant to such status, and impinge on the centers of intellectual creativity and cultural transmission and become integral, even if transient, parts thereof.

This rather unique combination of intellectual antinomianism with youth rebellion gave rise to a new type of youth ideology that was connected to the reversal of the hitherto existing relation between the definition of different age-spans and the possibilities of social and cultural creativity. Unlike in the past—even the not-so-distant past—youth became more and more seen not only as a preparation for independent and creative participation in social and cultural life, but as the very embodiment of permissive, often unstructured creativity—to be faced later on with the constants of a relatively highly organized, constrictive, meritocratic, and bureaucratic environment. Thus youth now became, even more than before, the transmitters and symbols of a continuous antinomian confrontation between the pristine values it embodies and a demystified social reality, and the impact of this new type of generational consciousness has been of crucial importance in the crystallization of what has been designated as the postmodern trends in contemporary societies. While the intensity and the concrete manifestation of this consciousness will naturally change from one situation to another, the consciousness itself with the various antinomian tendencies analyzed above seems to have become more and more prevalent.

See also: Age: Anthropological Aspects; Age, Sociology of; Age Stratification; Age Structure; Cohort Analysis; Generation in Anthropology; Generations in History; Generations: Political; Generations, Relations Between; Life Course in History; Life Course: Sociological Aspects; Mannheim, Karl (1893–1947); Structure: Social; Youth Culture, Anthropology of; Youth Culture, Sociology of; Youth Movements

Bibliography

- Aron R 1969 *The Elusive Revolution: Anatomy of a Student Revolt*. Pall Mall Press, London
- Bauer R A, Inkeles A, Kulckhohn C 1956 Generational differences. In: R A Bauer et al. (eds.) *How the Soviet System Works: Cultural, Psychological and Social Themes*. Harvard University Press, Cambridge, MA, pp. 190–8
- Bettelheim B 1954 *Symbolic Wounds, Puberty Rites and the Envious Circle*. Free Press, Glencoe, IL

- Bettelheim B 1963 The problem of generations. In: Erikson E H (ed.) *Youth: Change and Challenge*. Basic Books, New York, pp. 54–93
- Bourges H 1968 *The French Student Revolt: The Leaders Speak*. Hill and Wang, New York
- Eisenstadt S N 1956 *From Generation to Generation: Age Groups and Social Structure*. Free Press, Glencoe, IL
- Eisenstadt S N 1963 Archetypal patterns of youth. In: Erikson E H (ed.) *Youth: Change and Challenge*. Basic Books, New York, pp. 24–43
- Erikson E H (ed.) 1963 *Youth: Change and Challenge*. Basic Books, New York
- Fortes M 1984 Age, generation and social structure. In: Kertzer D I, Keith J (eds.) *Age and Anthropological Theory*. Cornell University Press, Ithaca, NY, pp. 99–123
- Glazer N 1970 *Remembering the Answers: Essays on the American Student Revolt*. Basic Books, New York
- Gusfield J 1957 The problem of generations in an organizational structure. *Social Forces* 35: 323–30
- Heberle R 1951 *Social Movements: An Introduction to Political Sociology*. Appleton Century-Crofts, New York [See especially 'The Problem of Political Generations,' pp. 118–27]
- Lipset S M, Altbach P G (eds.) 1969 *Students in Revolt*. Houghton Mifflin, Boston
- Lipset S M, Wellin S S (eds.) 1965 *The Berkeley Student Revolt. Facts and Interpretations*. Doubleday, Garden City, NY
- Mannheim K 1952 The problem of generations. In: Kecskemeti P (ed.) *Essays on the Sociology of Knowledge*. Oxford University Press, New York, pp. 276–320
- Marias J 1949 *El Metodo Historico de las Generaciones*. Revista de Occidente, Madrid, Spain
- Nash L L 1978 Concepts of experience, Greek origins of generational thought. *Daedalus* Fall: 1–23
- Ortega y Gasset 1933 *The Modern Theme*. Norton, New York
- Pinder W 1926 *Kunstgeschichte nach Generationen*. Pfeiffer, Leipzig, Germany
- Preston T R 1927 *My Generation of Politics, and Politicians*. Rose, Toronto
- Rintala M 1963 A generation in politics: A definition. *Review of Politics* 25: 509–22
- Zeitlin M 1966 Political generations in the Cuban working class. *American Journal of Sociology* 71: 493–508

S. N. Eisenstadt

Generative Grammar

Generative grammar is the label of the most influential research program in linguistics and related fields in the second half of the twentieth century. Initiated by a short book, Chomsky's *Syntactic Structures* (1957), it became one of the driving forces among the disciplines jointly called the cognitive sciences. The term *generative grammar* refers to an explicit, formal characterization of the (largely implicit) knowledge determining the formal aspect of all kinds of language behavior. From the start the program had a strong mentalist orientation instance, documented, for, in a fundamental critique of Skinner's *Verbal Behavior* (1957) by Chomsky (1959), arguing that behaviorist

stimulus–response theories could in no way account for the complexities of ordinary language use. The 'Generative Enterprise,' as the program was called in 1982, went through a number of stages, each of which was accompanied by discussions of specific problems and consequences within the narrower domain of linguistics as well as the wider range of related fields such as ontogenetic development, psychology of language use, or biological evolution. Four stages of the Generative Enterprise can be marked off for expository purposes.

1. Transformational Structure and Levels of Description

The foundation of all further developments is to be found in the Chomsky's monograph *The Logical Structure of Linguistic Theory* (1975, henceforth LSLT), which was published only 20 years after the manuscript was completed. Starting from the notion of structural analysis as developed by Bloomfield and made precise by Harris (1951), LSLT introduces a radically new perspective on the nature of linguistic structure. On the one hand, LSLT takes up Harris's notion 'level of linguistic structure,' in terms of which the form of utterances, i.e., their phonetic and syntactic shape, is characterized. On the other hand, however, LSLT develops a notion of linguistic structure that differs fundamentally from the empiricist orientation of American structuralism. More specifically, Chomsky proposes a notion of grammar that provides a full and explicit account of the tacit knowledge a speaker/hearer has of the form of utterances of his or her language. Under this perspective, a grammar *G* is a theoretical construct, dealing with internal states of knowledge that relate only indirectly to overt behavior of the speaker. In other words, *G* is viewed as an empirical hypothesis about the mental structure underlying ordinary speech behavior. To make this notion precise, LSLT highlights the importance of intuitive judgments a speaker is able to make about properties of utterances, such as the grammatical well-formedness of the famous nonsense-sentence (1) as opposed to the ungrammaticalness of (2), or the rather different properties of (3) and (4), in spite of their superficial similarity:

- (1) colorless green ideas sleep furiously
- (2) furiously sleep ideas green colorless
- (3) this picture was painted by a real artist
- (4) this picture was painted by a new technique

Hence *G* of English has to account for the properties on which intuitions like these rely. The study of the general features of *G* can furthermore be assumed to reveal principles of the internal organization of mental states in general and of knowledge of language in particular. This is in fact the basic motivation underlying the research program of Generative Grammar.

More technically, LSLT gives the following account of the concepts needed for this program. First of all, a linguistic level **L** is defined as an algebraic structure of the following sort:

- (5) $\mathbf{L} = [L, R_1, \dots, R_m, \mu, \Phi, \varphi_1, \dots, \varphi_n]$, where
 L is the set of primes of **L**;
 R_1 to R_m are sets of relations in **L**,
 such as linear ordering, inclusion,
 identity;
 μ is a set of so-called **L**-markers,
 generated by the primes of **L**;
 Φ is a mapping that maps μ onto
 the set of grammatical utterances;
 φ_1 to φ_n are operations relating **L** to other
 levels.

Roughly speaking, an **L**-marker is the representation of the structure of an utterance on the level **L**. The particular levels assumed in LSLT include the phonetic level **Pt**, the phonemic level **Pm**, the morpheme level **M**, the word level **W**, the level of word classes **C**, the phrase structure level **P**, and the transformational level **T**. For **Pm**, the set μ consists of strings of phonemes, for **W**, the elements of μ are sequences of words, and for **P**, the set μ consists of phrase markers, roughly tree structures indicating the constituency of an utterance. The relation between the different levels defined by the operations φ_i depends on particular conditions of the various levels and is, as Chomsky argues, anything but trivial. The full structural representation of an utterance comprises the **L**-markers assigned to it at the various levels in question. The mapping Φ relating the levels of representation to the actual (or possible) grammatical utterances is correspondingly indirect and complex. The assumption in LSLT is that it may be sufficient to define Φ for phonetic representations, i.e., strings in **Pt**, provided that the relation to **Pt** is defined for all other levels. This accounts for the fact that the structure of an utterance cannot be arrived at by simple operations of segmentation and classification, as assumed in classical structuralism, but depends on highly indirect and complex relations.

Second, a grammar **G** is a system of operations or rules that generate the sentences of a language with their structural representations. Technically, **G** determines the set of **L**-markers for each level **L** and the operations by which they are related to the other levels. LSLT deals with the general form of phonological rules, defining the relation between the phonemic and phonetic level, and more extensively with phrase structure rules, generating **P**-markers, and with grammatical transformations. The role of transformations is a matter of central interest in LSLT (such that for quite a while the term 'transformational grammar' was taken as largely equivalent to 'generative grammar'), they are assumed to capture structural properties **P**-markers could not account for. For instance, the difference between (3) and (4), which are not distinct in terms of constituency, is explained by

different transformational derivations, informally indicated in (3') and (4'):

- (3') a real artist painted this picture = T_{Passive}
 $\Rightarrow$ this picture was painted by a real artist
 (4') someone painted this picture by a new
 technique = $T_{\text{Passive}} \Rightarrow$ this picture was
 painted by a new technique by someone
 = $T_{\text{Agent-deletion}} \Rightarrow$ this picture was
 painted by a new technique

An important point of the theory developed in LSLT is the assumption that grammars must systematically be evaluated. To this effect, a simplicity-metric is introduced by which a grammar is preferred if and only if it expresses more natural generalizations than its competitors. For example, a grammar that accounts systematically for the different properties of sentences like (3) and (4) is to be preferred over one that can only assign to them the same structure. This is the reason why a grammar with a transformational component comes out as simpler than one without it. One of the factors entering the evaluation is the principle of rule ordering, which recognizes the possibility that rules might have a different structural effect, if they operate in different ordering. As a simple case in point, in (4') the Agent-deletion can operate only after the Passive Transformation has applied. (The importance of rule ordering had already been observed in the Sanskrit grammar of Panini.)

The productivity of the proposed framework is demonstrated by an analysis of large sections of English with respect to phrase structure and transformations in LSLT and in a stimulating monograph on nominalization by Lees (1960). A seminal study of phonology is given in Halle (1959) for Russian and in Chomsky's unpublished master's thesis on the *Morphophonemics of Modern Hebrew*.

A feature that LSLT inherits from the notion of transformational analysis proposed by Harris (1957) is the notion that transformations not only relate individual sentences to their more elementary basis, as shown in (3') and (4'), but are also used to decompose complex sentences into their constituent parts made up. Thus (6c), for instance, is composed by embedding of (6a) into (6b), replacing the bracketed pronoun:

- (6) (a) this picture was painted by a new technique
 (b) the director suspected [it]
 (c) = $T_{\text{Embedding}} \Rightarrow$ the director suspected
 [this picture to be painted by a new
 technique]

This assumption requires a distinction between simple transformations like T_{Passive} and so-called generalized transformations like $T_{\text{Embedding}}$. Another distinction that must be made is that between optional and obligatory transformations. Passivization, for instance, is optional, while the fronting of relative and wh-pronouns is obligatory, as the ungrammaticality of (7b) shows:

- (7) (a) I know the man who you met last night
 (b) *I know the man you met who last night

The clarification of these and a wide range of other complexities concerning the level **T** is one of the driving forces in further developments of Generative Grammar.

Another, more general feature that the Generative Enterprise inherits from American Structuralism is the assumption that all grammatical structure, from the most concrete level **Pt** of phonetics to the most abstract level **T** of transformational structure, is a matter of linguistic form which must be embedded in a wider context of a theory of semiotics that accounts for sense, reference, and use of linguistic expressions. Although there have been several proposals to include a level of semantics into the framework of generative grammar, the main emphasis has been on the knowledge of linguistic form throughout.

2. The Standard Theory of Generative Grammar

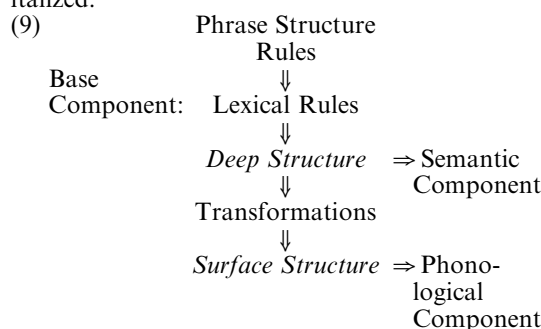
Immediately after the publication of *Syntactic Structures* (actually lecture notes excerpted from LSLT), the theory it proposed attracted interest from various sides, including mathematical linguistics and psychology. An overview of the formal properties of grammatical rules, especially of phrase structure grammars, in Chomsky (1963) established what became known as the Chomsky hierarchy of formal grammars as a branch of subrecursive function theory. Miller and Chomsky (1963) gave a first outline of the role of Generative Grammar for the study of language comprehension and production. Furthermore, Katz and Fodor (1963) made a first proposal to extend the theory of grammar by a semantic component, using so-called projection rules to integrate compositionally the meaning of lexical items according to their syntactic relations.

The main development, however, was due to continuous efforts to constrain systematically and strengthen the theory, increasing its explanatory power by reducing the range of possible grammars compatible with a given set of data. Such a reduction reduces at once the task assigned to the evaluation metric. Efforts to this effect led to what was later called the 'Standard Theory' of Generative Grammar. Based on preparatory ideas discussed, *inter alia*, in Katz and Postal (1964), it is formulated in Chomsky's *Aspects of the Theory of Syntax* (1965). The core point is a reassessment of the nature of transformational structure. Instead of a separate level **T** of transformations, grammatical transformations are now construed as operations that define the relation between two levels of syntactic representation, both of which are characterized by P-markers. This modification is related to one of the most influential distinctions proposed in the Standard Theory, namely that between *deep* and *surface structure*. According to this view, each sentence is assigned a deep and a surface P-marker, with trans-

formations mapping the former on the latter. Thus except for the different choice of personal or wh-pronouns, (8)(a)–(d) are just different surface realizations of the same deep structure, where the bracketed Agent phrases might furthermore optionally be deleted by $T_{\text{Agent-deletion}}$, known from (4')

- (8) (a) John may expect her to meet him
 (b) she may be expected (by John) to meet him
 (c) who does John expect her to meet
 (d) who may be expected (by John) to be met (by her)

On this account, the semantic interpretation of linguistic expressions could apparently be determined on the basis of their deep structure, which would be an interesting step towards an integration of grammar into an account of meaning and reference. Another important point of the Standard Theory is a more systematic distinction between (a) the lexical system, i.e., the complex system of rules that introduce lexical items with their (idiosyncratic) phonetic, syntactic, and semantic properties, and (b) the grammatical rules that integrate the lexical information into the structure of complex linguistic expressions. Finally, the distinction between levels of representation on the one hand and components of grammar generating and inter-relating these representations is made more perspicuous than in LSLT. The syntactic levels that appeared to be crucial are deep- and surface-structure, while word- and morpheme-structure do not have an independent status. The components of grammatical rules are (a) the base component, including the lexical system and the phrase structure rules, (b) the transformational component, (c) the phonological component, and, possibly (d) the semantic component. Rules and representations are inter-related according to the schema in (9), with levels of representation italicized:



The base component and the transformations together constitute the syntactic component. The semantic component raises problems to be taken up below, while the organization of the phonological component, extensively discussed by Chomsky and Halle (1968), is assumed to have a fairly clear organization, parallel to that of the syntactic component in relevant respects. The underlying phonemic structure is defined by the phonological features of lexical items and their arrangement in surface structure:

- (10)
- | |
|--|
| <i>Phonemic Structure</i> |
| ↓ |
| Phonological Component: Phonological Rules |
| ↓ |
| <i>Phonetic Structure</i> |

While there are still optional and obligatory transformations, generalized transformations are dispensed with, the deep structure of complex sentences being generated already in the base component. A particularly elegant feature that emerges from this reorganization is the principle of cyclic operation of the rules, which is added to the earlier assumption that rules are ordered. Cyclic operation apparently captures a deep property of natural languages: the same sequence of operations applies to increasingly larger parts of an utterance, defining its syntactic properties, its phonetic patterns, and the compositional integration of its semantic structure.

An important general effect of the Standard Theory is the perspicuous account it allows for the traditional notion of linguistic universals, or Universal Grammar (UG), considered as the innate, biologically fixed structure underlying the species-specific language capacity. The range of possible rules and the architecture of their interaction, as schematized in (9) and (10), could now be construed as formal universals, the potential primitive elements entering the rules and representations as substantive universals. Under this perspective, the set of possible grammars *G* accessible as instantiations of UG determines the diversity of possible natural languages, including the range of variation delimiting processes of linguistic change. By the same token, UG provides the innate, formal basis and predisposition that makes language acquisition possible. These accompanying considerations of the Standard Theory led to a new, principled approach to problems of language acquisition, as well as linguistic change.

3. Towards the Principles and Parameters Theory

Shortly after the Standard Theory was formulated, empirical analyses as well as efforts to improve the structure of the theory created modifications of the model. Controversial proposals concerned the role of semantics as indicated in (10). Lakoff (1971) and others proposed a model, according to which semantic representations are the underlying structure of linguistic utterances, from which surface structures are step by step derived by means of transformations and lexical rules. The insights and difficulties related to these proposals created a counterproposal, called the Revised Standard Theory. Based on observations discussed in Jackendoff (1972) it became obvious that meaning is not unaffected by surface structure phenomena. For example (11) and (12) are based on the same deep structure, with passivization applying to (12) but not to (11); they do not in general have the same meaning:

(11) Three students in the class wrote only one paper

(12) Only one paper was written by three students in the class

(Due to what is called the scope of quantification, in (11) the restriction of having written one paper is ascribed to three students, while in (12) the restriction of being written by three students is ascribed to a single paper.) As phenomena of this sort are by no means marginal, the Revised Standard Theory recognized a more complex relation between syntax and semantics than that indicated in (9). Only certain aspects of meaning are determined by deep structure, while scope and certain other relations depend on surface structure, such that transformations—contrary to the Standard Theory—might now affect the semantic interpretation.

A technical way to handle these problems is the so-called Trace Theory proposed in Fiengo (1977). According to this proposal, a constituent that is moved by a transformation leaves behind a trace that is phonetically empty, but may play a role in semantics, as indicated in the following example, where e_i is the trace left in the initial position of the co-indexed pronoun:

(13) who_i do you want to talk to e_i

In many cases, the trace turns out to be similar to what logicians consider a variable bound by the operator contained in the moved constituent. In this sense, invisible elements of the surface structure account for its contribution to semantic interpretation. Other empty elements of the surface structure are 'invisible' pronouns, as in (14)(a) and (b), where the actor of the infinitive clause is controlled by (i.e., co-referential with) different constituents of the main clause:

(14) (a) John_i promised Mary [e_i to take the train]

(b) John recommended Mary [e_i to take the train]

The structural aspect that emerges from transformations, as for instance in (13), cannot be the Deep Structure of (9), but it also differs from Surface Structure, as it contains elements like traces and empty pronouns. This more abstract surface, called S-structure, is assumed to provide the information for phonetic as well as semantic interpretation. On this account, (9) might be replaced by the general schema (15), where LF, namely the Logical Form, represents all the information *G* contributes to the meaning of an expression, just as PF, that is the Phonetic Form, specifies the conditions *G* imposes on its pronunciation:

- (15)
- | |
|----------------|
| Deep Structure |
| |
| S-structure |
| / \ |
| PF LF |

The theory of traces and other 'empty' elements is not a purely technical modification, however. It is

rather part of a general orientation striving for systematic restrictions on the still much too complex possibilities to construct a grammar according to the Standard Theory. One of the reasons motivating this orientation has been called the logical problem of language acquisition. This problem consists of the requirement to specify the conditions under which an appropriate grammar can be identified on the basis of the restricted and partially defective evidence the learning child is normally exposed to. This problem cannot be approached seriously if UG is as unconstrained as it would be with all the complexities initially allowed for as transformations. On the other hand, it turned out that many of the formal complexities exploited, for instance, in operations like the standard passive transformation, can be reduced to more elementary operations if the application and interaction of these operations is subject to more systematic constraints.

Induced by the Trace theory, a number of systematic conditions have been explored which determine the conditions of empty elements of different sorts. These conditions allow transformations to be reduced ultimately to just one simple operation, called Move α , where α is any constituent. The systematic and by no means trivial conditions determining the possible choice of α , the domain and bounds of movement, the consequences of movement (such as leaving a trace), are all expressed by general principles needed on independent grounds. One of the conditions concerns relations lexical elements impose on their necessary or optional complements. This is called their Argument- or Theta-Structure. Another rather general condition entering these principles is a systematic reorganization of phrase structure rules in terms of the so-called X-bar Theory. This theory takes up an observation made already in Harris (1951) about the restricted options according to which constituents X and Y combine in order to form a complex constituent $\bar{X}$. In somewhat simplified terms, properties of $\bar{X}$ are determined by its head X, which furthermore selects Y as its complement or is modified by Y as a free adjunct.

These and a number of other developments led to a new version of the overall framework called the Principles-and-Parameters Theory, formulated in Chomsky's *Lectures on Government and Binding* (1981). The basic idea of this theory is to replace the notion of different types of highly complex rules, from which individual grammars are made up, by a system of universal principles that constrain the effect of rather basic operations like Move α and the options of X-bar theory. The principles in question determine the relation between the positions of a moved constituent, the domain of movement, the possible candidates for movement, and the conditions that require the operation. They constitute separate subsystems, called Government, Binding, Bounding, Control, Case, and Thematic Structure. The properties distinguishing different languages are no longer seen as consequences

of intricate systems of different and rather complex rules. It rather seems possible to account for the relevant phenomena by the interaction of these general principles. The obvious differences between different languages is assumed to be reducible to different values of a restricted set of parameters contained in these general principles. Thus the head X of a constituent $\bar{X}$ may be in initial or final position, an embedded clause may need an introductory element or not, etc. Most (or all) of these parameters seem to be related to classes of particular lexical items so that all language particular properties are basically related to lexical information. This leads to a rather different notion of how UG determines an individual grammar G: besides lexical items making up the language particular dictionary, grammars of different languages might differ only by their choice of parameter values, while the principles these parameters rely on are given by UG and hence identical for all languages. Under this perspective, language acquisition consists essentially of identifying the lexical system of a language, thereby fixing the particular values of the parameters contained in UG. Presupposing the schema of levels given in (15), the organization of a grammar G within the Principles-and-Parameters theory can be indicated roughly as follows:

- (16) (a) Lexical System, fixing the idiosyncratic phonetic, semantic, and syntactic properties of lexical items
- (b) X-bar syntax, Move α
- (c) Principles of Government, Binding, Bounding, Thematic Structure, Control, and Case-Assignment
- (d) PF-component
- (e) LF-component

Among the numerous stimulating consequences of this model is a new perspective on historical and typological variation: like language acquisition, linguistic change and typological differences are not only constrained by UG, but also closely related to choice of parameter values on the basis of universal principles.

4. *The Minimalist Program of Generative Grammar*

Research within this framework explored the principles listed in (15c) in more detail, keeping to the overall orientation of the Generative Enterprise, namely improving the explanatory power by constraining the theory as far as possible. This led in particular to the attempt to derive the principles in question from general conditions of cognitive organization plus the boundaries that are conceptually necessary to identify the domain of language. Minimal assumptions in this sense require linguistic expressions to compositionally relate patterns of the perceptual and articulatory system A-P to representations in the range of con-

ceptually and intentionally organized experience C-I. This requires the internal or I-language determined by G to provide at least the interfaces PF and LF, relating language to A-P and C-I, respectively, as indicated in (17):

$$(17) \text{ signal} \Leftrightarrow \text{A-P} \Leftrightarrow \text{PF} \leftrightarrow \text{LF} \Leftrightarrow \text{C-I} \Leftrightarrow \text{environment} \\ \text{G}$$

Thus G must determine pairs $\langle \pi, \lambda \rangle$, where π and λ are representations in PF and LF, respectively. One must assume furthermore, that each pair $\langle \pi, \lambda \rangle$ must be based on a selection N of elements of the lexical inventory of the language. The research strategy of the Minimalist Program proposed in Chomsky (1995) pursues the aim to derive the properties of UG as far as possible from these minimal assumptions together with the hypothesis that UG is subject to conditions of structural economy. This requires first the lexical items entering the selection N to contain all and only the idiosyncratic specifications by means of which N participates in determining the pair $\langle \pi, \lambda \rangle$. This includes, besides phonetic and semantic features interpreted in P-A and C-I, respectively, certain grammatical or formal features controlling the computation of $\langle \pi, \lambda \rangle$ from N . The minimal assumption to be made about this computation is an operation *Merge* which combines two expressions X and Y into a complex expression Z , made up from the features of X and Y , adding the formal features of the head to characterize the formal properties of Z . As a matter of fact, the operation *Merge* is the minimalist version of X-bar syntax, reviving, moreover, a reduced version of the generalized transformations combining two expressions of arbitrary complexity. With this background, it is an interesting, empirical fact about natural language that the derivation from N to $\langle \pi, \lambda \rangle$ cannot in general be restricted to the simplest possibility requiring only the operation of *Merge*, but needs the effect of what was called *Move α* in order to account for facts like *Who_i did he_j talk to e_j*. As a matter of fact, the operation can be simplified to *Move*, as it is some formal feature of the moving constituent α , which triggers the operation if and only if the feature cannot be eliminated otherwise. More technically, then, the computation of linguistic expressions proceeds as indicated by (18):

$$(18) N \Leftrightarrow \text{Merge, Move} \Leftrightarrow \text{PF, LF} \rangle \text{ Spell Out PF}$$

A wide range of facts in different languages have been shown to follow from appropriate lexical information together with minimal assumptions about *Merge*, *Move*, plus two sorts of general conditions regulating the economy in representation and derivation of expressions of I-language. What this leads to is indicated in (19):

- (19) (a) Conditions on lexical items i.e., sets of phonetic, semantic, and formal features
- (b) Operations of the computational system of language: *Merge*, *Move*
- (c) Economy Principles:
- (i) Representational: Full Interpretation FI

(ii) Derivational: Shortest derivation from N to $\langle \text{PF}, \text{LF} \rangle$ -pairs

The requirement of Full Interpretation presupposes that *Merge* and *Move* check and eliminate step by step the formal features, leaving only information that can appropriately be interpreted in terms of articulatory and conceptual conditions at the two interface levels. Derivational economy is a fairly abstract notion, the appropriate formulation of which is still under exploration. In any event, violation of these principles is now taken to be the source of ungrammaticalness: only optimal derivations allowing for complete interpretation yield well-formed expressions. The control of these conditions is a crucial effect of formal features, coming primarily with special lexical items called functional categories. These include grammatical words like Determiners, Auxiliaries, but also inflectional elements like Tense, Number, or Case.

From a more general perspective, the conditions in (9) are rather general, compared with the still rather special principles in the Principles-and-Parameters model (16), very much like these principles were a strong generalization compared with the rule systems of the Standard Theory in (9). The shift from (9) to (16) replaced rules of individual grammars by universal but still language-specific principles; the organization of UG indicated in (19) replaces these language-specific principles with even more general conditions of cognitive organization. Language specificity is now essentially a matter of formal features and their effect on *Merge* and *Move*, that is the operations that compositionally relate P-A to C-I, modules of cognitive organization which are largely independent of the language capacity.

From this point of view, the species-specific language capacity might consist essentially of the disposal of discrete infinity and in fact be related, as Chomsky (1982) speculates, to the computational capacity underlying arithmetic. The language capacity clearly recruits and enhances general cognitive and communicative capacities, but its evolution, which Pinker (1994) persuasively argues to be due to adaptive selection, might well be an independent step accompanying evolutionary changes in the architecture and size of the human brain.

5. General Perspectives

The intellectual force of generative grammar comes from the combination of two equally important factors. First, it provides a suggestive methodological and theoretical perspective for the study of language as part of the biological endowment of the human organism, supporting a central system of the overall cognitive capacity. Second, it shows how this perspective can be made precise, turned into an effective research program, and pursued with respect to a large domain of empirical facts not recognized before. It is

indeed the wide range of phenomena—from syllable structure and inflection to quantifier scope and syntactic embedding—as well as the diversity of languages—from English, German, and Chinese to Hebrew, Malayalam, and Walbiri—that gives the Generative Enterprise its unusual place in the field. The explicit formulation of the theory, its technical means, and its cognitive orientation are the reason for its strong impact on a wide range of subdisciplines from research in language comprehension and production, acquisition and aphasia to typology, historical linguistics, computer science, poetics, theory of music, and philosophy of language.

As can be seen from the above, various offshoots developed different, though related, ideas. These are, *inter alia*, Categorical Grammar, Dependency Grammar, Case Grammar, Lexical Functional Grammar (LFG), Generalized Phrase Structure Grammar (GPSG), Head-driven Phrase Structure Grammar (HPSG), Relational Grammar, to mention the more influential proposals. Another type of offspring is the Optimality Theory (OT), which develops certain ideas of the Minimalist Program in a different way, assuming that conditions on linguistic structure may be of different strength and can be violated.

A peculiar relationship must finally be noted between Generative Grammar and the various theories of semantics and pragmatics. Sticking to the initial notion that grammar can and must account for the form, but not properly for the complexities of interpretation and use of language, Chomsky did not extend his theory of language to the domain of semantics and pragmatics. It might be noted in this respect that he always considered Logical Form as a syntactic level, pertaining to the form rather than the interpretation of language. Therefore, Formal Semantics as developed by Montague (1974), Discourse Representation Theory as proposed by Kamp and Reyle (1993), Speech Act Theory as conceived by Searle (1969) and a fair number of related approaches to aspects of interpretation and use are compatible with and in part stimulated by Generative Grammar, but they have not been included into its proper domain. Even though this situation was occasionally met with disappointment, it could not in general reduce the attraction exerted by the Generative Enterprise.

See also: Constituent Structure: Syntactic; Lexical Functional Grammar; Linguistics: Theory of Principles and Parameters; Montague Grammar; Movement Theory and Constraints in Syntax; Syntax; Syntax–Phonology Interface; Syntax–Semantics Interface

Bibliography

Chomsky N 1957 *Syntactic Structures*. Mouton, S'Gravenhage, The Netherlands

- Chomsky N 1959 Review of Skinner's *verbal behavior*. *Language* 35: 26–58
- Chomsky N 1963 Formal properties of grammars. In: Luce R D, Bush R R, Galanter E (eds.) *Handbook of Mathematical Psychology*. Wiley & Son, New York, Vol. II, pp. 323–418
- Chomsky N 1965 *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA
- Chomsky N 1975 *The Logical Structure of Linguistic Theory*. Plenum, New York
- Chomsky N 1981 *Lectures on Government and Binding*. Foris Publications, Dordrecht, The Netherlands
- Chomsky N 1982 *The Generative Enterprise: A Discussion with Riny Huybregts and Henk van Riemsdijk*. Foris Publications, Dordrecht, The Netherlands
- Chomsky N 1986 *Knowledge of Language: Its Nature, Origin, and Use*. Praeger, New York
- Chomsky N 1995 *The Minimalist Program*. MIT Press, Cambridge, MA
- Chomsky N, Halle M 1968 *The Sound Pattern of English*. Harper and Row, New York
- Fiengo R 1977 On trace theory. *Linguistic Inquiry* 8: 35–62
- Halle M 1959 *The Sound Pattern of Russian*. Mouton, S'Gravenhage, The Netherlands
- Harris Z S 1951 *Methods in Structural Linguistics*. University of Chicago Press, Chicago
- Harris Z S 1957 Co-occurrence and transformations in linguistic structure. *Language* 33: 293–340
- Jackendoff R S 1972 *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, MA
- Kamp H, Reyle U 1993 *From Discourse to Logic*. Kluwer Dordrecht: The Netherlands
- Katz J J, Fodor J A 1963 The structure of a semantic theory. *Language* 39: 170–210
- Katz J J, Postal P 1964 *An Integrated Theory of Linguistic Descriptions*. MIT Press, Cambridge, MA
- Lakoff G 1971 On generative semantics. In: Steinberg D, Jakobovits L (eds.) *Semantics*. Cambridge University Press, Cambridge, UK, pp. 232–96
- Lees R B 1960 *The Grammar of English Nominalizations*. Indiana University Press, Bloomington, IN
- Miller G A, Chomsky N 1963 Finitary models of language users. In: Luce R D, Bush R R, Galanter E (eds.) *Handbook of Mathematical Psychology*. Wiley & Sons, New York, Vol. II, pp. 419–91
- Montague R 1974 *Formal Philosophy*. Yale University Press, New Haven, CT
- Pinker S 1994 *The Language Instinct*. Harper Collins, New York
- Searle J R 1969 *Speech Acts*. Cambridge University Press, Cambridge, UK
- Skinner B F 1957 *Verbal Behavior*. Appleton-Century-Crofts, New York

M. Bierwisch

Generative Syntax

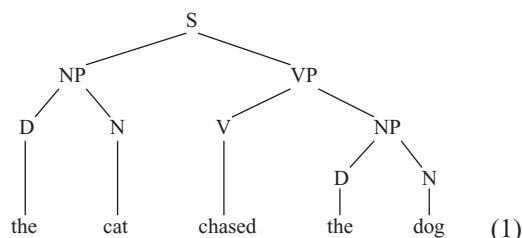
The term generative syntax refers to research on the syntax of natural languages within the general framework of generative grammar. As in generative grammar, the term 'generative' means explicit. Thus, generative syntax differs from much other research on syntax in its emphasis on precise and explicit specifica-

tion both of general theories and of specific analyses. Most work in generative syntax like most work in generative grammar has also been seen as an investigation of I-language (internalized language), the cognitive system underlying the ordinary use of language. Like generative grammar, generative syntax has its origins in Chomsky's *Syntactic Structures* (1957). It is often seen as the central component of generative grammar and it has always been the most prominent.

1. Origins

Syntax was very much a secondary concern for pre-Chomskian linguists and their ideas about syntax were largely unformalized. Chomsky's first concern was to make these ideas explicit and to investigate their adequacy. He argued that they were unsatisfactory in various ways and proposed a richer conception of grammar; known as transformational grammar (TG).

Most pre-Chomskian work on syntax implicitly assumed that words can be classified and that they combine together into phrases which can also be classified. In other words, it assumed that sentences have a constituent structure (or more than one if they are structurally ambiguous). Constituent structures are commonly represented by tree diagrams like (10):



This tells us what categories the words in the sentence belong to and what phrases the sentence contains. Chomsky pointed out that constituent structures could be characterized by phrase structure (PS) rules. A PS rule has the form in (2) and the interpretation in (3).
 $C_0 \rightarrow C_1 C_2 \dots C_n$ (2)
 A C_0 can consist of a C_1 , followed by a C_2 , followed by ... followed by a C_n (3)
 Chomsky argued that grammars consisting just of PS rules could not provide an illuminating account of various syntactic phenomena, and argued for a richer conception of syntactic structure in which some constituents are introduced into the structure somewhere other than their superficial position and then moved to their superficial position. For example, he argued that the subject of a passive sentence such as (4) originates in object position.

Kim was arrested (4)

This provides an explanation for the fact that 'arrested' is acceptable here without an object although it

normally requires one, as the ungrammaticality of (5) shows.

* The policeman has arrested. (5)

It also provides an explanation for the fact that 'Kim' is understood in the same way in (4) as in (6), i.e., as the person undergoing the action.

They have arrested Kim. (6)

Phenomena like these led Chomsky to propose that sentences have an underlying constituent structure characterized by PS rules and that the superficial constituent structure is derived from this by various transformational rules. Thus, for most versions of TG, the full syntactic structure of a sentence is a sequence of constituent structures. The underlying structure has been called deep structure or D-structure and the most superficial structure, surface or S-structure.

Since 1957 a variety of different theories have been developed within generative syntax and they have been applied to many different languages.

2. Data

Generative analyses make complex predictions about what is and what is not grammatical, i.e., in accord with the speaker's I-language. The most efficient way of investigating these predictions is through the judgments or intuitions of native speakers. Hence, generative syntax has always been based mainly on intuitions. Of course, intuitions have to be treated with caution. The fact that a speaker finds some sentence unacceptable does not necessarily mean that it is ungrammatical for her. A speaker might find (7) unacceptable because it is false, and might find (8) unacceptable because it is misleading. (It is natural initially to assume incorrectly that 'her' forms a constituent with 'contributions to the fund.')

Three fours are thirteen. (7)

Without her contributions to the fund would be quite inadequate. (8)

However, if a sentence is unacceptable and there is no other explanation for the unacceptability, it is reasonable to assume that it is ungrammatical.

None of this implies that other kinds of data, e.g., corpora of naturally occurring speech or writing, cannot be useful. In fact generative syntacticians have often used corpus data. Moreover, there is no alternative to corpus data if the speakers are dead. Hence, generative work on dead languages is inevitably based on a corpus supplemented by intelligent guesses about meanings. Corpus data has also played a central role in generative work on the early stages of first language acquisition.

The emphasis on intuitions has encouraged researchers to work on their own language. In the early years of generative syntax, most researchers were native speakers of English, and as a result English had more attention than any other language. However, as Newmeyer (1986, p. 40) emphasizes, generative syntax

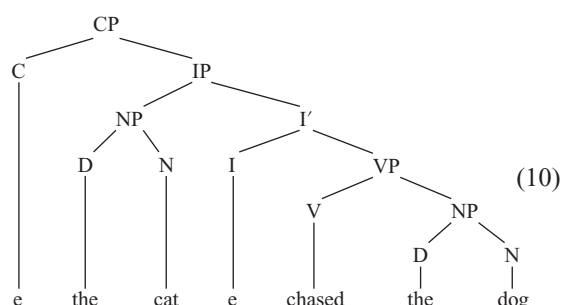
was never limited to English. Moreover, languages other than English have been increasingly prominent since about 1980 as a result of the growing numbers of generative syntacticians who are native speakers of other languages. Work in generative syntax often focuses on a single language, but work that compares and contrasts a pair of related languages is also quite common, and some work deals with a broad range of languages.

3. Structures

As has been noted, most versions of TG assume that the full syntactic structure of a sentence is a sequence of constituent structures. Different versions of TG differ in how complex these constituent structures are. Typically, since the early 1980s, they have been much more complex than (1) as a result especially of the presence of various phonologically empty elements. These include traces, empty categories 'left behind' by movement processes. Thus, (4) will have a trace in object position left behind by movement of the NP 'Kim.' Empty elements also include empty functional categories. Functional categories include C (complementizer) and I (inflection), exemplified by 'that' and 'may' in (9), respectively.

I think that the cat may chase the dog. (9)

It is standardly assumed that a simple sentence contains empty members of these categories. Thus, instead of (1), one might have something like (10)



A variety of functional categories, which are commonly empty, have been assumed in recent TG work.

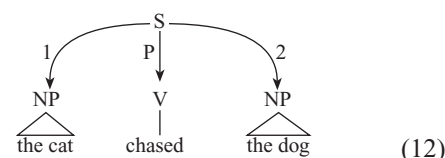
The assumption that the syntactic structure of a sentence is a sequence of constituent structures has been largely abandoned within the recent minimalist version of TG (Chomsky 1995, Radford 1997). Within this framework, movement processes apply within a constituent before it is combined with another constituent. In (11), for example, 'Kim' is moved to subject position in the subordinate clause before this clause is combined with the verb 'think.'

I think Kim was arrested. (11)

Hence, there is no level of structure after all processes of combination have applied and before any move-

ment processes apply. In other words, there is no level of D-structure.

Much work in generative syntax has argued for different sorts of structures. Standard constituent structures do not incorporate labels like subject and object. It is assumed within TG that these notions can be defined in structural terms. Assuming structures like (10), one might define subject as an NP that is an immediate constituent of IP and object as an NP that is an immediate constituent of VP. Proponents of Relational Grammar (RG) argued in the early 1970s that such functional notions should be primitives and that constituent structure trees should be replaced by relational networks like (12), where 1 and 2 stand for subject and object, respectively, and P stands for predicate:



In early work in RG, the structure of a sentence is a sequence of such relational networks and the order of constituents is only significant in the most superficial network. In more recent work, the structure of a sentence is a single complex network and order is determined by its most superficial properties. (For discussion, see Blake 1990.)

The idea that certain functional notions should be primitives is also a feature of Head driven Phrase Structure Grammar (HPSG) (see Pollard and Sag 1994). Here the structure of a sentence is a single quite simple constituent structure enriched with information about whether a constituent is a head (the constituent which determines the identity of the larger expression, a noun in a noun phrase, a verb in a verb phrase, etc.), a subject, a complement, etc. (Object is not a primitive in HPSG.) A somewhat different position is found in Lexical-Functional Grammar (LFG). Here the full structure of a sentence consists of a fairly conventional constituent structure and a functional structure, in which constituents are identified as subjects, objects, etc. Similarity structures provide a way of capturing the similarities between superficially rather different languages. For example, the following sentences from English, Welsh, and Turkish, which have rather different constituent structures because subject, verb, and object are ordered differently, could have essentially the same functional structure:

The cat saw the dog. (13)

Gwelodd y gath y ci (Welsh) (14)

saw the cat the dog

'The cat saw the dog.'

Kedi köpeğ-i gör-dü (Turkish) (15)

cat dog-ACC see-Past

'The cat saw the dog.'

See Bresnan (2001) for discussion.

An important question about the various forms of generative syntax is: what sorts of categories do they assume? Chomsky (1965, 1970) argued that various syntactic phenomena require the assumption that syntactic categories are complex entities made up of smaller elements normally known as features. This position has been generally accepted since then. However, different approaches differ in how complex their categories are. Categories are particularly complex within HPSG, where, as has been noted, the structure of a sentence is a single quite simple constituent structure. In contrast, they appear to be relatively simple within recent TG, which assumes much more complex structures.

4. Rules and Principles

Like any scientific research, work within generative syntax is concerned to capture whatever generalizations can be found in the data. A consequence of this is that broad general statements are preferred to statements of more limited scope.

In the 1960s and early 1970s TG assumed a variety of different PS and transformational rules. More recently, the various transformations have been replaced by a single rule known as Move Alpha or in recent work just Move. This is essentially a license to move anything anywhere. It interacts with various conditions on movement and the results of movement to allow just the right movement processes. Particularly important in recent work have been certain economy conditions, which prefer simpler to more complex derivations. In much the same way, the various PS rules have been replaced by a few general principles interacting with the categorial makeup of lexical items to give the observed variety of basic, pre-movement structures. In the minimalist version of TG, it is assumed that a single rule of Merge is operative here. A similar position is found in much work in HPSG, in which it is assumed that the variety of syntactic structures are the product of a small number of rules and principles interacting with the properties of lexical items.

Within TG, the move away from large numbers of PS and transformational rules has been associated with a denial that constructions as traditionally understood are of any theoretical significance. Thus, Chomsky remarks that '[t]he notion of construction, in the traditional sense, effectively disappears; it is perhaps useful for descriptive taxonomy but has no theoretical status' (1995, p. 25). However, since the late 1980s a number of syntacticians have argued that it is impossible to characterize the full range of syntactic structures with just broad general principles and individual lexical items, and that it is necessary to recognize constructions with certain properties of their own after all. For example, Sag (1997) argues that data

such as the following require the recognition of a relative clause construction with properties of its own which do not derive either from broad general principles or individual lexical items:

someone on whom I relied (16a)

someone who I relied on (16b)

someone on whom to rely (16c)

*someone who to rely on (16d)

(See Kay and Fillmore (1999) and Culicover and Jackendoff (1999) for similar arguments.)

Different approaches to syntax have taken different views about the nature of rules of syntax. Much work in TG and especially the minimalist framework has treated most rules as procedures for constructing well-formed structures. In contrast RG, HPSG, and LFG have interpreted rules as constraints to which well-formed structures must conform. Consequently these approaches are often described as constraint-based.

A further important issue is: how do grammatical constraints interact? The traditional view is that syntactic structures must conform to all relevant constraints. In contrast, proponents of what is known as Optimality Theory (OT) have argued that constraints are ranked and that a structure may violate a constraint if that is the only way to satisfy a higher ranked constraint. This view of constraints has been very influential within phonology but has also had some influence within syntax. It is combined with TG assumptions in e.g., Grimshaw (1997). Grimshaw argues, for example, that there is a constraint which rules out the movement of constituents. However, a higher ranked constraint requires movement of 'who' to sentence initial position from its underlying position as object of 'to' in the following:

Who did you talk to? (17)

The OT view of constraints has also been combined with LFG assumptions e.g. in Bresnan (2000).

5. Syntax and Other Components of Grammar

A central question for any theory of syntax is: how does the syntactic component of a grammar relate to other components? What is the relation between syntax on the one hand and semantics, morphology, phonology, and the lexicon on the other?

In the late 1960s and early 1970s, the relation between syntactic and semantic representations was a major topic of debate. Proponents of Generative Semantics argued that the deepest level of syntactic representation is a semantic representation. Proponents of Interpretive Semantics rejected this view and argued that semantic representation is derived by rules referring to both deep and surface structure. There are questions as to how different the two positions really were in the early stages of the debate. In the later stages they were clearly distinguished as generative semantics sought to bring more and more linguistic phenomena within the scope of grammar. Generative semantics collapsed in the mid 1970s, partly because of

a realization that some of the phenomena it had brought within the scope of grammar could be handled more satisfactorily within pragmatics. (See Newmeyer 1986, Chaps. 4 and 5 for discussion.) It is now generally accepted within generative syntax that syntactic and semantic representations are rather different, and that there is no one-to-one correspondence between them. The following examples are relevant here:

Have you seen a unicorn? (18a)

Never **have I** seen a unicorn. (18b)

Kim has seen a unicorn and so **has Lee**. (18c)

Had I seen a unicorn, I would have told you. (18d)

All these sentences exemplify what might be called the English auxiliary + subject construction. The key words are in bold. They show that this construction has a variety of different semantic interpretations. (18a) contains an interrogative, the kind of sentence that is normally used to ask a question. (18b) and (18c) contain declaratives, the type of sentence normally used to make a statement. Finally, (18d) contains a conditional clause. Here, instead of auxiliary + subject order, one might have 'if' and subject + auxiliary order.

Since the mid 1970s, most work in TG has assumed that semantic interpretation is based on a level of logical form (LF) derived from S structure by various movement processes. However, Jackendoff (1997) develops a detailed critique of this view. For HPSG, general principles provide a semantic representation for all expressions on the basis of the semantic representations of their constituents. For LFG, semantic interpretation is based on functional structure.

There are also important questions about the relation between syntax and morphology. At various times it has been proposed within TG that word stems may be combined with affixes in syntax. For example, it has sometimes been suggested that an English verb stem combines with the past tense suffix '-ed' through a movement process. This view largely fell from favor in the 1990s and it is now widely accepted even within TG that word-formation is confined to the lexicon.

There are also questions about the relation between syntax and phonology. It has sometimes been proposed within TG that certain movement processes apply within the phonological component. For example, Chomsky (1995, p. 368) proposes that verb-second order in German is a result of a rule applying within the phonological component. The following examples illustrate this phenomenon:

Ich las schon letztes Jahr diesen Roman. (19a)

I read already last year this book

Diesen Roman las ich schon letztes Jahr. (19b)

this book read I already last year

Schon letztes Jahr las ich diesen Roman. (19c)

already last year read I this book

'I read this book already last year.'

Each example has a different constituent in initial position, but all have the verb in second position. It

has also been assumed in much work in HPSG that linear order is solely a property of phonology in the sense that it is not words and phrases but only their phonological properties that are ordered.

Questions also arise about the relation between syntax and the lexicon. Within TG a central issue is where the insertion of lexical items into syntactic structures applies. It has often been assumed that it applies before any movement process. However, it has sometimes been proposed that some lexical insertion may take place after movement. Within HPSG, where lexical categories are quite complex, there are important questions about how generalizations about the categorial makeup of words can be captured. Considerable use has been made of lexical rules deriving lexical entries from lexical entries. More recently, it has been argued that most generalizations can be captured if words are classified in terms of a number of hierarchies of lexical types from which they inherit properties. Thus, 'chased' might be classified as both a transitive verb and a past tense verb and combine the properties of these two types (see Koenig 1999).

There are further issues about the scope of syntax. Should grammars include information about the contexts in which constructions are appropriate? For example, should a grammar include the information that (19) with no overt object following 'beat' is only appropriate in recipes?

Beat until stiff. (20)

It has generally been assumed that such information is something separate interacting with purely grammatical information (see, for example, Culy 1996). However, a number of researchers have argued that it is inseparable from purely grammatical information (see Hudson 1996, Paolillo 2000). Should grammars incorporate statistical information about how likely certain constructions are to occur? For example, should a grammar incorporate the information that (21a) with a clause in subject position is less likely to occur than (21b) with the clause in final position?

That Kim was there is odd. (21a)

It is odd that Kim was there. (21b)

Traditionally, statistical facts about syntactic structures have been seen as a matter of performance, but Abney (1996) argues that such information is part and parcel of grammars.

6. Some Syntactic Phenomena

There are many unresolved issues in generative syntax but a great deal has been learned about both the universal syntactic properties of natural languages and the ways in which they can vary in their syntax.

A major focus of research has been clause structure and a lot is known about the kinds of clause structure that a language may exhibit and about what kinds of analyses might be appropriate. This includes such

matters as the order of subject, verb, and object (see (13)–(15), the realization of tense, mood, and negation, agreement, and case-marking, and the relation between main and subordinate clauses. German, which has verb-final order in subordinate clauses, is relevant here. The following subordinate clause contrasts with the examples in (19):

dass ich schon letztes Jahr diesen Roman las (22a)
that I already last year this book read
'that I read this book last year'

There has also been a great deal of research on noun phrases and a lot is known about such matters as the realization of definiteness, the positioning of adjectives (which normally precede the noun in some languages, e.g., English, but normally follow in others, e.g., Welsh), and the positioning of relative clauses (which follow the noun in English and most European languages but precede in Turkish, Japanese, and Korean).

Perhaps the most important focus of research in generative syntax has been the various kinds of non-local dependencies which natural languages exhibit, which were largely unnoticed in pre-generative work. An important example is the dependency between a clause-initial interrogative element like 'who' and the associated gap which is a feature of many languages. A simple example is (21) in which there is a gap in object position.

Who did Kim annoy? (23)

It has been known since the 1960s and especially since Ross (1967) that these dependencies are subject to various constraints. For example, while (23) is fine, (24) with a gap following 'and' is ungrammatical.

* Who did Kim annoy Lee and? (24)

There are important differences between languages in this area. English allows a gap in prepositional object position. Hence both the following are possible:

Who did you talk to? (25a)

To whom did you talk? (25b)

However, most languages do not allow a gap in prepositional object position. The following Polish examples are typical.

* Kim rozmawiałeś z? (Polish) (26a)

whom you-talked with

Z kim rozmawiałeś? (Polish) (26b)

with whom you-talked

In contrast, English and many other languages do not allow a gap in subject position in a subordinate clause introduced by a complementizer. Hence, the following is ungrammatical:

* Who did you say that wrote this book? (27)

However, some languages do allow this. Thus, the following Italian example is grammatical:

Chi hai detto che ha scritto questo libro?
(Italian) (28)

who have-you said that has written this book

'Who did you say wrote this book?'

As the translation makes clear, an English example is grammatical if there is no complementizer. See Roberts (1997, Chap. 4) for discussion of the extensive work on these matters.

Another important dependency is that between a reflexive pronoun such as English 'himself' and its antecedent, the expression from which it derives its interpretation. In English, while (29a) is acceptable, (29b) is not.

John criticized himself. (29a)

*John said Mary criticized himself (29b)

Thus, the antecedent must be quite close to the reflexive. In other languages, however, a reflexive can be further from its antecedent than in English. For example, in the following Japanese sentence, the antecedent of the reflexive 'zibun' can be either 'Ziro' or 'Taro'.

Taroo-ga Hanako-ni [Ziroo-ga zibun-o
(Japanese) (30)

Taro-NOM Hanako-to Ziro-NOM self-ACC

hihan-sita]-to itta.

criticized-COMP said

'Taro said to Hanako that Ziro criticized self.'

Hanako is not a possible antecedent because it is not a subject. For discussion of these matters, see Roberts (1997, Chap. 3).

Non-local dependencies like these are likely to remain a major focus of research in generative syntax for the foreseeable future.

7. Acquisition and Processing

As has been noted, it is generally accepted within generative work that syntax is concerned with I-language, the cognitive system underlying ordinary use of language. It follows that a variety of data are relevant, for example, facts about acquisition, facts about processing, and facts about pathology.

Acquisition has been a major concern within TG. Since about 1980, it has been argued that the language faculty consists of a set of principles to which all languages conform and a set of parameters which allow a limited amount of variation between languages. For example, it is commonly assumed that there is a parameter distinguishing languages like English, in which interrogative elements appear in initial position, from those like Japanese, in which they appear in the same position as corresponding non-interrogative elements. The following examples illustrate:

John gave the book to Mary. (31a)

Who did John give the book to? (31b)

John-wa Mary-ni sono hon-o

John-TOP Mary-DAT that book-ACC

agemasi-ta-ka. (Japanese) (32a)

give-PAST-Q

'John gave the book to Mary.'

John-wa dare-ni sono hon-o

John-TOP who-DAT that book-ACC

agemasi-ta-ka? (Japanese) (32b)

give-PAST-Q

'Who did John give the book to?'

Within this framework, language acquisition is largely a matter of setting parameters. A great deal of work on the acquisition of syntax has drawn on these ideas. However, some are skeptical. There are two reasons for this. First, relatively little progress has been made towards a plausible set of parameters even for the central or 'core' phenomena of syntax, and languages have a wide range of noncore phenomena, which are also acquired (Culicover 1999). Second, it is clear that setting a parameter is a much more complex matter than it was originally thought to be (Fodor 1998). Nevertheless the idea remains a potentially important one.

Other frameworks have been more concerned with language use. It has been argued that HPSG receives some support from what is known about human language processing. HPSG is a constraint-based framework, in which all rules and principles are on a par and none takes precedence over any other. Pollard and Sag (1994, pp. 11–12) argue that this permits an explanation for some important features of human language processing. They highlight four features. First, it is highly incremental in the sense that 'speakers are able to assign partial interpretations to partial utterances.' Second, it is highly integrative in the sense that 'information about the world, the context, and the topic at hand is skillfully woven together with linguistic information whenever utterances are successfully decoded.' Third, '... there is no one order in which information is consulted which can be fixed for all language use situations.' Finally, it consists of various kinds of activity: 'comprehension, production, translation, playing language games, and the like.' They argue that these features are unproblematic for constraint-based grammars but pose serious problems for other sorts of grammars.

Facts about acquisition and processing are likely to be of continuing importance for the evaluation of proposals within generative syntax.

See also: Generative Grammar; Movement Theory and Constraints in Syntax; Syntax; Syntax–Phonology Interface; Syntax–Semantics Interface

Bibliography

- Abney S 1996 Statistical methods and linguistics. In: Klavans J L, Resnick P (eds.) *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. MIT Press, Cambridge, MA, pp. 1–26
- Blake B J 1990 *Relational Grammar*. Routledge, London
- Bresnan J 2000 Optimal syntax. In: Dekkers J, van der Leeuw F, van de Weijer J (eds.) *Optimality Theory: Phonology, Syntax and Acquisition*, Oxford University Press, Oxford, UK, pp. 334–85
- Bresnan J 2001 *Lexical-Functional Grammar*. Blackwell, Oxford, UK

- Chomsky N A 1957 *Syntactic Structures*. Mouton, The Hague, The Netherlands
- Chomsky N A 1965 *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA
- Chomsky N A 1970 Remarks on nominalization. In: Jacobs R, Rosenbaum P S (eds.) *Readings in English Transformational Grammar*. Ginn, Waltham, MA
- Chomsky N A 1995 *The Minimalist Program*. MIT Press, Cambridge, MA
- Culicover P W 1999 *Syntactic Nuts: Hard Cases, Syntactic Theory and Language Acquisition*. Oxford University Press, Oxford, UK
- Culicover P W, Jackendoff R S 1999 The view from the periphery: The English comparative correlative. *Linguistic Inquiry* 30: 543–71
- Culy C 1996 Null objects in English recipes. *Language Variation and Change* 8: 91–124
- Fodor J D 1998 Unambiguous triggers. *Linguistic Inquiry* 36: 1–36
- Grimshaw J 1997 Projection, heads and optimality. *Linguistic Inquiry* 28: 373–422
- Hudson R A 1996 *Sociolinguistics*, 2nd edn. Cambridge University Press, Cambridge, UK
- Jackendoff R S 1997 *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA
- Kay P, Fillmore C 1999 Grammatical constructions and linguistic generalizations: The *What's X doing Y?* construction. *Language* 75: 1–33
- Koenig J-P 1999 *Lexical Relations*. CSLI Publications, Stanford, CA
- Newmeyer F J 1986 *Linguistic Theory in America*, 2nd edn. MIT Press, Cambridge, MA
- Paolillo J C 2000 Formalizing formality: An analysis of register variation in Sinhala. *Journal of Linguistics* 36: 215–59
- Pollard C, Sag I A 1994 *Head-driven Phrase Structure Grammar*. University of Chicago Press, Chicago, IL
- Radford A 1997 *Syntactic Theory and the Structure of English: A Minimalist Approach*. Cambridge University Press, Cambridge, UK
- Roberts I 1997 *Comparative Syntax*. Arnold, London
- Ross J R 1967 *Constraints on Variables in Syntax*. PhD dissertation, MIT
- Sag I A 1997 English relative clauses. *Journal of Linguistics* 33: 431–83

R. Borsley

Genes and Behavior: Animal Models

Explaining behavior in genetic terms is no small task. Behavior draws on the activities of a wide range of genes acting at many different times in an individual's life. Model organisms have provided insight into the role of genes in behavior that have major implications for humans. Many of the mutations and genetic variants (alleles) that affect behavior in *Drosophila* and mice have proved to be subtle changes in genes that, when more severely mutated, produce wider ranging effects. Studies of mutants induced in the laboratory have identified a variety of genes that affect behavior:

genes that regulate other genes (transcription factors), regulatory enzymes (protein kinases), and signaling molecules in the brain (neuropeptides and receptors). Their study has made major contributions to our understanding of biological rhythms, learning, memory, and social behavior, and has helped explain newer findings on naturally occurring variants in behavior.

1. Introduction

Unitary explanations for human behavior have a dismal history in the modern world. Whether economic, cultural, psychological, or genetic, attempts to reduce human behavior to a single influence inevitably fail to give a full account and have often produced distortions with unfortunate political consequences. Charting the relationship between genes and behavior can therefore be a perilous activity. The challenges are many: the difficulty of defining genetic and nongenetic factors, the difficulty of understanding the interactions among these many factors, the many steps that intervene between the expression of a gene and the manifestation of a behavior, and the inescapable fact that each individual is the unique product of a series of historical accidents. Historical uniqueness also applies to nonhuman biological individuals. It is a hallmark of the biological world due to the fact that no particular combination of genes and experiences is ever replicated exactly in nature.

Studies of genes and behavior in model organisms offer one way out of this conundrum. To begin with, it is far easier to control environmental conditions and genetic heterogeneity in the laboratory, and such studies are generally undertaken keeping environmental conditions as uniform as possible and genetic makeup as well defined as possible. But this would be no help if the phenomena and principles are fundamentally different from those of humans. Fortunately, they appear to be similar by an ever-increasing number of criteria. One of the most striking outcomes of modern biology and the sequencing of genomes is the similarity of genes between distantly related creatures. The fruit fly *Drosophila melanogaster* and humans share a great many of the same genes in common. Similarly, they share considerable conservation of metabolic and signaling pathways at the cellular level. Most strikingly, and most relevant for this discussion, there is growing evidence for conservation at the level of behavior and its molecular mechanisms. This has been shown most convincingly, thus far, for several behaviors in *Drosophila*: circadian rhythms (Young 1998), learning and memory (Dubnau and Tully 1998), and most recently sleep (Greenspan and Ferveur 2000). Thus, there is good reason to believe that the characteristics, interactions, and contributions of genes in the behavior of model organisms will provide important guidelines for how to think about the role of genes in human behavior.

Traditionally, there have been two separate, and often hostile, approaches to behavior genetics in model organisms: the measurement and manipulation of naturally occurring variation in laboratory or wild-caught strains, and the isolation and manipulation of newly induced mutations. The former assayed the extent of relatively mild mutational variation that survives in the world and exerts effects on behavior, without being able to know the identity of any of the genes or their mechanism of action. The latter concentrated exclusively on identifying genes and mechanisms subserving particular behaviors, but did so by producing relatively drastic genetic lesions that did not reflect the permissible variation that would have been worked on by natural selection. Both have been relevant to humans, but a synthesis between the two, made possible by recent advances in molecular biology, holds the most promise for useful insights. In this article, a few salient examples of gene effects on behavior are described as illustrations of general principles that are emerging.

2. Severe Single-gene Mutants

2.1 Circadian Rhythms

The clearest contribution single-gene mutants have made is in the realm of identifying individual genes that are central to behavioral mechanisms. The unraveling of the cellular mechanism of the circadian clock began with the discovery of the *period* gene through the isolation of long-day, short-day, and arrhythmic mutants in *Drosophila* (Fig. 1; Konopka and Benzer 1971) and proceeded with the isolation and cloning of additional mutants in flies, fungi, and mice (Young 1998). What has emerged is a picture of the circadian clock as a feedback loop in which each cell counts out its own 24 hour period by means of an oscillating cycle of gene readout in which some of the products accumulate up to a point, then begin to dampen their own synthesis through a series of time delays between the various stages of gene expression: transcription, translation, and import back into the nucleus (Fig. 1). This serves as the central regulator for specific implementation of rhythmic activities. The cellular mechanism of the circadian clock, as well as the clock's phenomenology, appears to be universal in the biological world and many of the actual genes involved, such as *period*, are well conserved between flies and mammals. The human relevance is thus transparent.

2.2 Learning and Memory

Learning and memory have many features common to flies, rodents, and humans. Behaviorally, all are

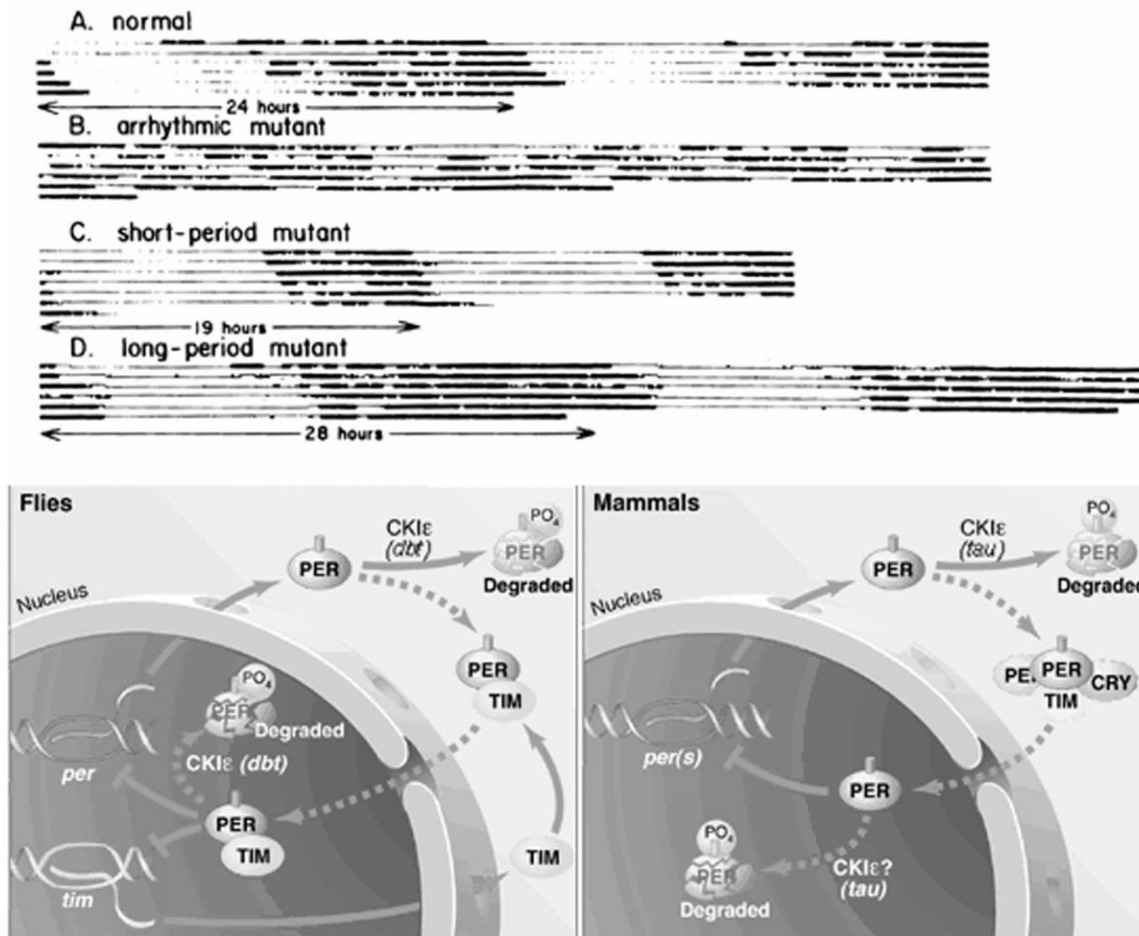


Figure 1

Top: circadian locomotor activity in normal *Drosophila* and in mutants of the *period* gene as measured in an activity monitor in the absence of a day/night cycle of light and darkness. Heavy traces represent activity, with the clock's day length indicated underneath (Source: *Proceedings of the National Academy of Sciences USA* 68: 2112–6. Reproduced with permission).

Bottom: conservation of molecular mechanisms of the circadian clock between flies and mammals. PER = *period* gene, TIM = *timeless* gene, CKIε = casein kinase Iε gene (Source: *Science* 2000 288: 451–3. Reproduced with permission)

capable of associative conditioning and the memory that is induced shows multiple phases that can be distinguished by pharmacology and by training regimen (Dubnau and Tully 1998). In the fly, learning takes the form of distinguishing between two odors and avoiding the one that has previously been paired with electric shock. Long-term memory is produced by repeated training trials separated by suitable time intervals. The original *Drosophila* learning mutant, *dunce* (Dudai et al. 1976), is defective in initial acquisition as well as in subsequent retention of the learned behavior (Fig. 2). Care was taken to ascertain that these mutant flies were normal in all other relevant

faculties that are required for the learning task: olfaction, locomotor activity, and shock reactivity.

The product of the *dunce* gene was subsequently identified as an enzyme that is part of the signaling cascade in nerve cells that utilizes the intracellular messenger molecule cyclic AMP (Fig. 2). The gene encodes a cAMP-phosphodiesterase, the enzyme that degrades cyclic AMP. Mutations in other components of this signaling system, *rutabaga* in the synthetic enzyme (adenylate cyclase), and *Dco* and *RI* in the key enzyme regulated by cyclic AMP (cAMP-dependent protein kinase), were subsequently shown to also affect learning and memory in the fly (Dubnau and Tully

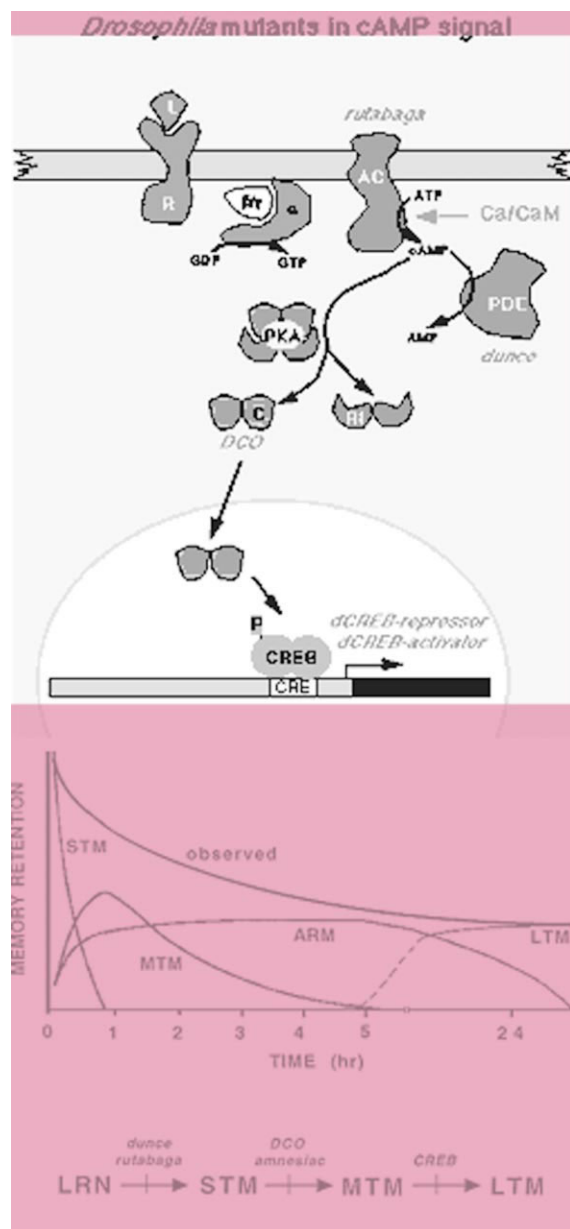


Figure 2

Top: the cAMP signaling pathway in memory formation. AC = adenylyl cyclase (*rutabaga* gene), PDE = cAMP phosphodiesterase (*dunce* gene), PKA = cAMP-dependent protein kinase (*DCO* and *RI* genes), CREB = cAMP-response-element-binding protein (*dCREB* gene) (source: *Nature Neuroscience* 1998 1: 543–5. Reproduced with permission). Middle: memory is consolidated via distinct phases that are genetically and pharmacologically dissectable in *Drosophila* and in mammals: short-term memory (STM), middle-term

1998). The cyclic AMP signaling system had already been implicated in neuronal plasticity as central to the mechanism that modifies efficacy at synapses in the circuitry underlying learning. The isolation of mutants affecting the system from unbiased, open-ended mutant screens and the demonstration that they affected classical conditioning in the intact animal gave universality to the finding. Mouse mutants in the synthetic enzyme (adenylate cyclase) show analogous deficits in learning (Silva et al. 1998). The relevance to humans has been borne out by the fact that there are currently several memory enhancement drugs in clinical trial that target cAMP phosphodiesterase.

As additional mutants accumulated, comparison of the time course of memory decay revealed some interesting differences between them (Fig. 2), suggesting that they differentially affected memory retention. This suggestion received major support from studies of another component of the cAMP signaling system: the *Creb* (cAMP-response-element-binding) protein that regulates genes induced by cAMP. *Creb* exists in both repressor and activator forms. A repressor form of the gene, when engineered to be expressed in the fly just prior to training, completely blocks the formation of long-term memory while permitting normal short- and medium-term memory (Dubnau and Tully 1998). In contrast, an activator form of the gene, when engineered to be expressed just prior to training, produces long-term memory after a single trial, in contrast to the usual 10 trials with intervals that are required to induce long-term memory (Fig. 2). These findings not only placed the differential genetic effects on memory phases on firmer ground, they were also subsequently extrapolated to mice and rats (Silva et al. 1998). A *Creb* mutant in the mouse that reduces, but does not eliminate, repressor activity fails to form long-term memory in a fear-avoidance paradigm. Analogous results have been obtained in the rat after injection of a *Creb* gene blocker (anti-sense RNA) to specifically reduce levels of the protein.

A variety of other mutants disrupting learning that also affect synaptic efficacy have also been obtained in the fly and mouse, including those affecting other regulatory enzymes (calcium/calmodulin-dependent protein kinase, *fyn* kinase) and receptors (NMDA and acetylcholine). The studies of the cAMP cascade stand out, however, for the depth to which they have taken the analysis through their insight into the mechanism of differential memory phases.

memory (MTM), anesthesia-resistant memory (ARM), and long-term memory (LTM) appear sequentially, and each has a progressively slower rate of decay (source: *Trends Neuroscience* 1995 18: 212–7). Bottom: a genetic model of information flow showing where in the pathway single gene mutant and pharmacological disruptions have their primary effects (source: *Cell* 1994 79: 35–47. Reproduced with permission)

2.3 Social Behavior

Genetic studies of social behavior have great relevance for humans and for considerations of human evolution. A productive avenue into this question has been taken in rodents, based on neuroendocrinological studies of the behavioral effects of oxytocin and vasopressin. Administration of these hormones produce complementary effects on a variety of behaviors including affiliative, sexual, maternal, territorial, and aggressive behavior as well as on social memory (Young et al. 1998). The detailed neural mechanisms for the hormones' actions have yet to be elaborated. A mouse mutant that lacks oxytocin altogether (*Oxt*^{-/-}) shows a distinct deficit in social memory as reflected by male olfactory investigation of a female in repeated exposures. The effect is more selective than predicted from the range of effects when oxytocin is administered. Moreover, *Oxt*^{-/-} mutant mice are normal for olfactory responses to other, nonsocial odors, and normal for learning tasks unrelated to this social behavior.

An increase in aggression between males was also observed in the oxytocin-deficient mice. This observation falls into line with a growing number of other mouse mutant studies reporting increases in aggression. These range from mutants in receptors for serotonin or adenosine, to synthetic enzymes for the neurotransmitters GABA or nitric oxide, to regulatory enzymes (calcium/calmodulin-dependent protein kinase II) to cell adhesion molecules. In contrast to mutants of circadian rhythms or of learning and memory, there seem to be a great many genetic lesions that produce aggression in mice. When one encounters such promiscuity in genetics, it is usually diagnostic of a system that is subject to many nonspecific kinds of disruption, with all of the caveats to interpretation that accompany effects on such a system, as opposed to a sensitive indicator of key mechanistic components.

2.4 Range of Action of Behavioral Genes

The foregoing examples highlight the mechanistic insights that have been gained from behavioral studies of laboratory-induced mutants. An important criterion for these studies was the relative refractoriness of the behavior to mutational disruption: mutations in most genes do not affect these behaviors. The question then arises of the relative specificity of these genes to the behaviors they disrupt. Are these 'genes for' behaviors such as rhythms, memory, or social life, as have often been alluded to in evolutionary discussions of behavior (e.g., Wilson 1978)? For the mutants described above, as well as for the vast majority of behavioral mutants studied in both flies and mice, the answer appears to be negative. By and large, the genes that produce behavioral mutants are not limited to affecting only one behavior or only one aspect of the

animal. They exert influences over many facets of the organism's biology (the genetic term for this is 'pleiotropy').

The archetypal circadian rhythms gene, *period*, regulates the song produced by a male fly during courtship as well as its 24-hour rhythms (Greenspan and Ferveur 2000). The male's song, produced by vibrating the wing, consists of a series of 'pulses' which sound like bumps occurring at the rate of about 30 per second. The rhythm lies in the slight but regular variation in the interval between pulses (Fig. 3) as the male gradually increases then decreases this interval to produce a smoothly wavering song over a period of approximately 60 seconds. Female flies respond to the male's song by becoming more receptive to their courtship advances. Mutants in the *period* gene that affect the circadian rhythm have corresponding effects on song rhythm such that long-day mutants have an 80-second rhythm, short-day mutants have a 40-second rhythm, and arrhythmic mutants have no regular rhythm (Fig. 3).

The *period* mutant also affects timing on scale greater than the 24-hour period: short-day and long-day mutants have a shorter and longer developmental time course, respectively, from egg to adult compared to normal controls. These effects of *period* occur on timescales an order of magnitude larger (10 days) or several orders of magnitude smaller (60 seconds) than the 24-hour clock. Distant as they are from the circadian timescale, they are timing mechanisms nonetheless and therefore not outside of the realm already affected by the mutants. It remains a mystery, however, as to how the relatively slow mechanism of feedback onto gene expression (Fig. 1) can also regulate a 60-second rhythm. One clue comes from a more divergent consequence of *per* and several other rhythm mutants that was revealed in pharmacological studies of cocaine responsiveness, in which they fail to become sensitized. There is no circadian or timing-dependent aspect to this phenomenon. Instead, it appears to involve a different use of their function as gene regulators: induction of the enzyme responsible for sensitization fails to occur in the mutants.

Mutants affecting learning and memory are no less wide-ranging. The *dnc* mutation produces sterility in female flies, defects in embryonic development, and shorter lifespan. The most severe *Creb* mutants are lethal in both flies and mice. Finally, *Oxt*^{-/-} female mice are defective in lactation, as well as in infant vocalization.

The conclusion that genes producing behavioral mutants are wide-ranging in their effects (pleiotropic) is based on a large enough sampling of mutants from two different organisms to consider it likely to be the rule rather than the exception (Hall 1994, Pflugfelder 1998, Mouse Knockout & Mutation Database, <http://research.bmn.com/mkmd/>). It would be a mistake, therefore, to ignore it in analyses of how genes affect behavior, or in considerations of what

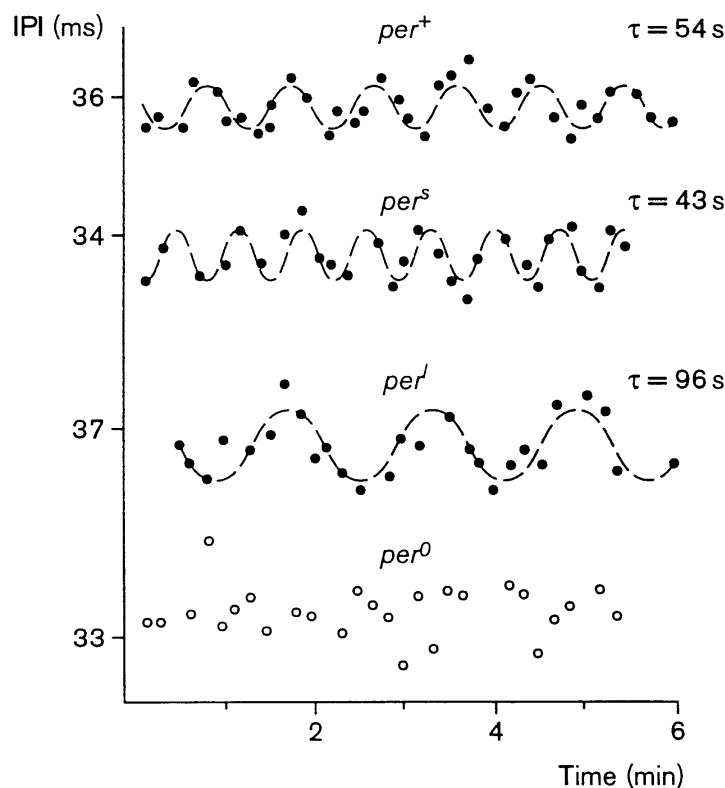


Figure 3

Rhythmic oscillations of the courship song of normal (per^+) male *Drosophila melanogaster* and short-day (per^s), long-day (per^l), and arrhythmic (per^0) mutants of the *period* gene. The song consists of pulses of sound lasting 33–7 milliseconds (ms) produced by vibration of the wing. The interval between these pulses (interpulse interval, IPI) oscillates regularly with a periodicity (τ) indicated in seconds (s) (source: *Proceedings of the National Academy of Sciences USA* 1980 77: 6729–33. Reproduced with permission)

kinds of behavioral variation are evolutionarily permissible. A sobering conclusion that emerges from this recognition is that the path from gene to behavior is extraordinarily complex, winding its way through all possible levels of biological organization, from gene regulation through cellular biochemistry, through communication between cells, through large-scale patterns of activity in the nervous system, finally emerging as the manifest behavioral output itself.

3. Mild Natural Variants

Natural variation in behavior exists between and within species, and provides a major substrate for natural selection. Divergence in behavior between species is thought to be a critical component of the speciation process itself, whereas behavioral variants within a species provide it with the flexibility needed to weather the unpredictability of external conditions. Until recently, the identity and mechanistic nature of

these behavioral variants has been a black box. Analyses were confined to statistical treatments of overall behavioral variation (e.g., Ehrman and Parsons 1981). A growing list of these variants has now been characterized molecularly and functionally. As foreshadowed by the experience with laboratory mutants, many have proved to be subtle variations of genes that are capable of more severe and wider ranging effects.

3.1 Differences Between Species: Biological Rhythms

The courship song produced by *Drosophila* males is species-specific and plays a role in species recognition for mating (Greenspan and Ferveur 2000). The courship song of *Drosophila melanogaster* differs from that of the closely related species *Drosophila simulans* in several parameters, the most important of which is its rhythmic oscillation (see Sect. 2.4). The *D. melano-*

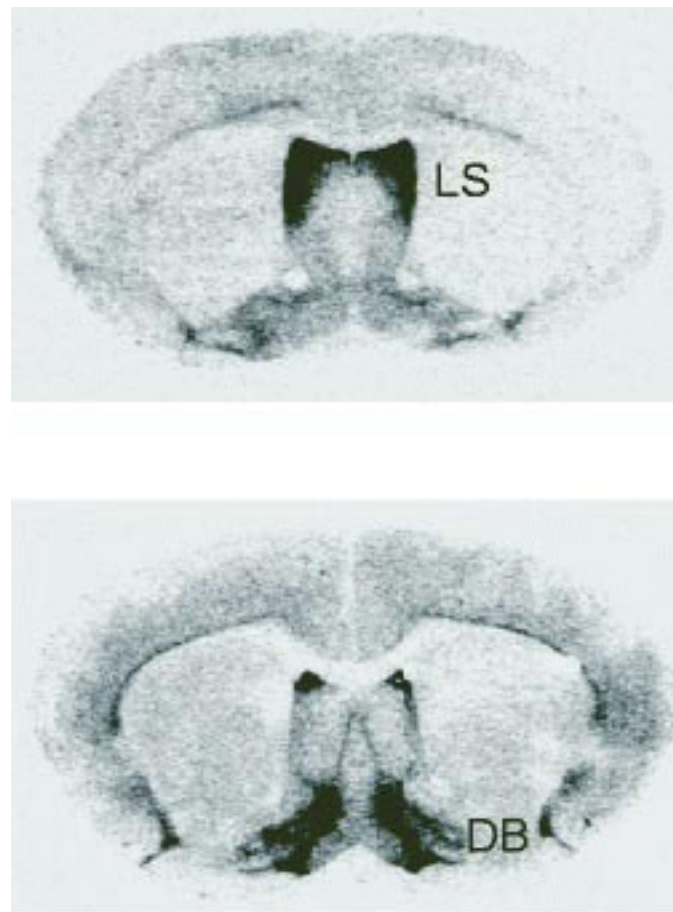


Figure 4

Anatomical distribution of the V_{1a} vasopressin receptor in the brain of the montane vole (top) and the prairie vole (bottom). Binding is more intense in the lateral septum (LS) of the montane vole but not the prairie vole, and in the diagonal band (DB) of the prairie vole but not the montane vole. Similar differences exist throughout the brain (source: *Nature* 1999 **400**: 766–8. Reproduced with permission)

gaster oscillation has a 60-second period, whereas the *D. simulans* oscillation has a 35-second period. This rhythmic difference is due to a subtle difference in the sequence of the *period* gene between *D. simulans* and *D. melanogaster*: a short segment of the protein coding region that contains repeats of the amino acid pair threonine and glycine (Fig. 3). A *D. simulans period* gene transferred into *D. melanogaster* that lack a *period* gene confers a characteristic *D. simulans* song rhythm (Wheeler et al. 1991). More specifically, a synthetic gene consisting of the entire *D. melanogaster period* gene, except for a segment of *period* from *D. simulans* consisting of a small region that is highly variable among many *Drosophila* species, also gave the same characteristic *D. simulans* song (Fig. 3). The remarkable feature of this finding is that one portion of the sequence was capable of diverging in a species-

specific manner, while the rest of the gene stayed constant enough to retain the normal 24 hour circadian rhythmicity.

3.2 Differences Between Species: Social Behavior

Differences in social behavior are marked in the animal kingdom even between closely related species. Two species of North American vole illustrate this point. The prairie vole (*Microtus ochrogaster*) is monogamous, biparental, and highly social. They contrast with the montane vole (*Microtus montanus*), which is promiscuous, maternal, and minimally social (Young et al. 1998). The hormones oxytocin and vasopressin exert complementary effects on these behaviors in the

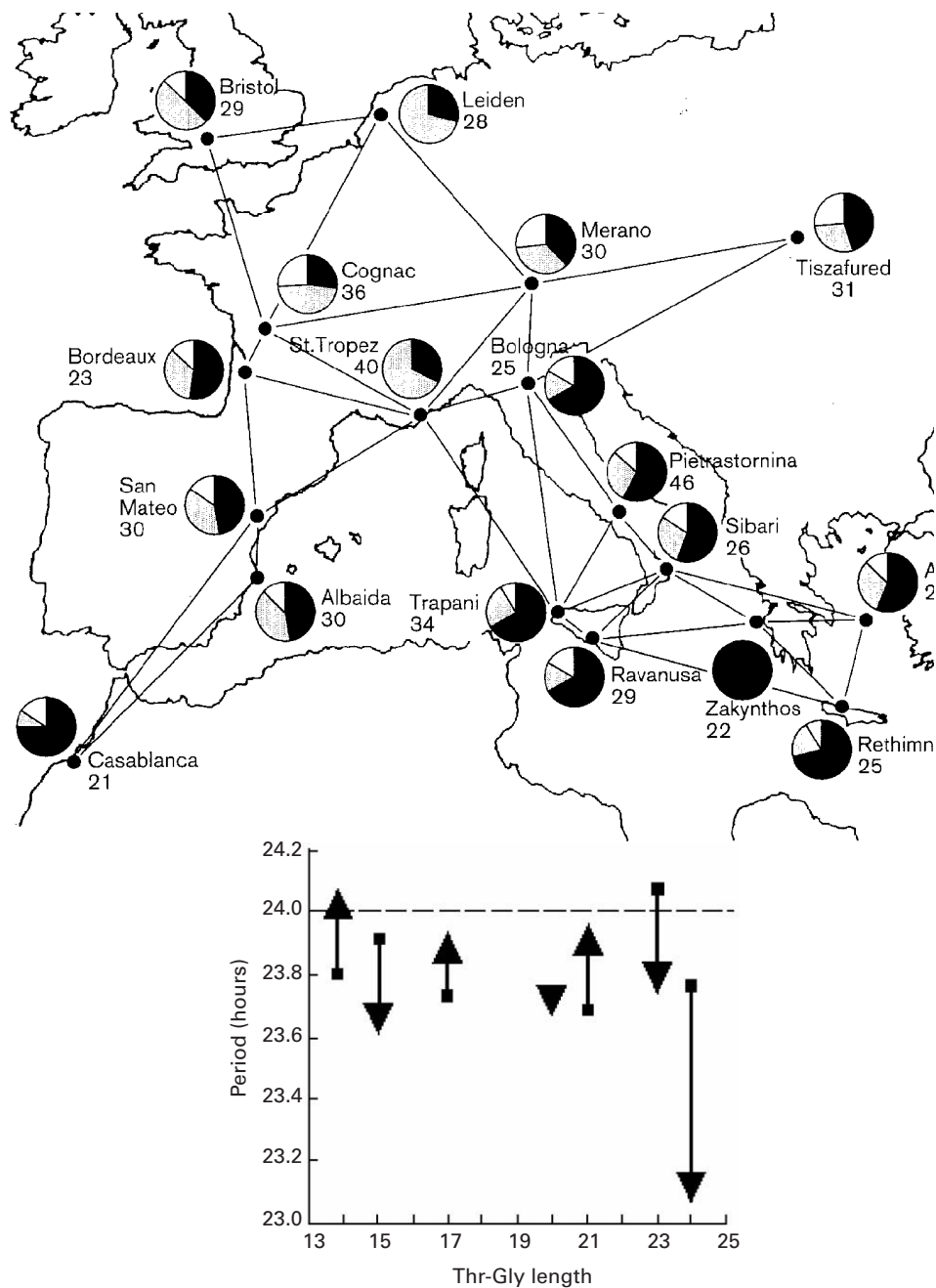


Figure 5

Top: latitudinal cline in distribution of the (Thr-Gly)₁₇ (dark gray) and (Thr-Gly)₂₀ (light gray) alleles of the *period* gene. The (Thr-Gly)₂₀ allele, which is more prevalent in the north, is more robust in its temperature compensation. Numbers indicate chromosomes tested (source: *Proceedings of the Royal Society of Biological Science* 1992 **250**: 43–9. Reproduced with permission). Bottom: temperature compensating ability at 18°C (box) and 29°C (arrowhead) of alleles of the *period* gene differing in the length of their Thr-Gly repeats. The arrow reflects the change in direction of the period at 29°C. The (Thr-Gly)₂₀ variants show almost identical periods at the two temperatures (source: *Science* 1997 **278**: 2117–20. Reproduced with permission)

prairie vole, similar to their effects in rats and mice (see Sect. 2.3). Moreover, the anatomical distribution of oxytocin and vasopressin receptors in the brain differs markedly between the two vole species (Young et al. 1999).

Vasopressin administration stimulates affiliative behavior in prairie voles but not in montane voles, nor in mice. When a prairie vole vasopressin receptor gene (V_{1a}) was transferred into mice, thereby creating an anatomical distribution similar to that of the prairie vole, the mice began to exhibit affiliative behavior in response to vasopressin (Fig. 4). This result argues that changing the pattern of V_{1a} distribution is sufficient to alter this behavior. This interpretation is further supported by the fact that the V_{1a} receptor genes are virtually identical between the two vole species that differ in affiliative behavior, except for the presence in prairie voles of a small segment in the region of the gene that is likely to regulate its expression. The presence of this DNA segment correlates well with behavioral and anatomical features of two other species, the pine vole (*Microtus pinetorum*), which is like the prairie vole behaviorally as well as in its anatomical and molecular characteristics of V_{1a} , and the meadow vole (*Microtus pennsylvanicus*), which is like the montane vole in all of these respects. Thus, it appears that subtle evolutionary tinkering with the expression pattern of the vasopressin receptor V_{1a} may account for some of the differences in social behavior between these rodent species.

3.3 Differences Within Species: Circadian Rhythms

If the genetic differences underlying behavioral variation between species are subtle, then those underlying variation within species are downright esoteric. The segment of the *period* gene that was sufficient to transfer a *D. simulans* courtship song rhythm into *D. melanogaster* also varies within the *melanogaster* species. Two versions (alleles) of the gene predominate: one containing 17 repeats of the pair of amino acids threonine and glycine (Thr-Gly)₁₇, and one containing 20 repeats (Thr-Gly)₂₀ (Costa and Kyriacou 1998). These two alleles are nonrandomly distributed in the wild, following a north-south gradient (known as a 'latitudinal cline') from northern Europe down to North Africa with (Thr-Gly)₂₀ predominating in the north and (Thr-Gly)₁₇ predominating in the south (Fig. 5). Distributions of this sort are traditionally explained as an adaptive response to climatic variation.

If these variants are undergoing selection, then the functional differences between them become relevant. A clue to such a functional difference was seen in tests of temperature compensation, one of the distinctive features of circadian rhythms which allow cold-blooded organisms to maintain their 24 hour cycle at

different temperatures. Flies carrying the northerly (Thr-Gly)₂₀ allele exhibit more robust temperature compensation than those with the more southerly (Thr-Gly)₁₇ allele (Fig. 5). The temperature range is much greater in the north and presumably imposes greater demands on the temperature-compensating ability of the clock. The size of the effect is small, perhaps reflecting a tradeoff between the temperature compensation and conservation of the overall 24 hour rhythm.

3.4 Differences Within Species: Foraging Behavior

Most natural variation in behavior is not primarily attributable to any single gene. Where multiple genes contribute, the effect often disappears when they are separated from each other in the crosses required for genetic mapping. An exception to this trend is the *foraging* gene that affects food search behavior by *Drosophila* larvae.

Natural populations of *D. melanogaster* harbor two larval types: rovers, who search widely for food, and sitters, who do not (Sokolowski 1998). The difference is not merely locomotor because it is expressed only in the presence of food. Genetic mapping revealed it to be due primarily to a single gene, *foraging*, and molecular analysis showed it to be an enzyme involved in regulating many physiological processes, cGMP-dependent protein kinase (Fig. 6). Severe mutations of the *foraging* gene are lethal. The detailed mechanism through which the enzyme alters neuronal and behavioral activity is not known, but it appears to regulate neuronal thresholds.

A gene whose product has as wide-ranging functions as the enzyme encoded by *foraging* must have severe constraints on the kinds of variation that are viable and permissible. The explanation, in this case, appears to be that the magnitude of the natural enzyme variation is relatively small: rovers have 12 percent more activity of the enzyme cGMP-dependent protein kinase than sitters and the levels of its protein and mRNA are correspondingly higher in rovers vs. sitters (Osborne et al. 1997). This small increase thus appears to be sufficient to make a very distinctive difference in behavior, while not seriously impairing viability or obligate functions of the gene.

3.5 Genetic Basis for Selected Behaviors

Artificial selection in the laboratory has been the conventional method for detecting the presence of naturally occurring genetic variation in behavior (Ehrman and Parsons 1981). While such variation has been clearly demonstrable in virtually every behavioral selection attempted, from maze learning in rats to

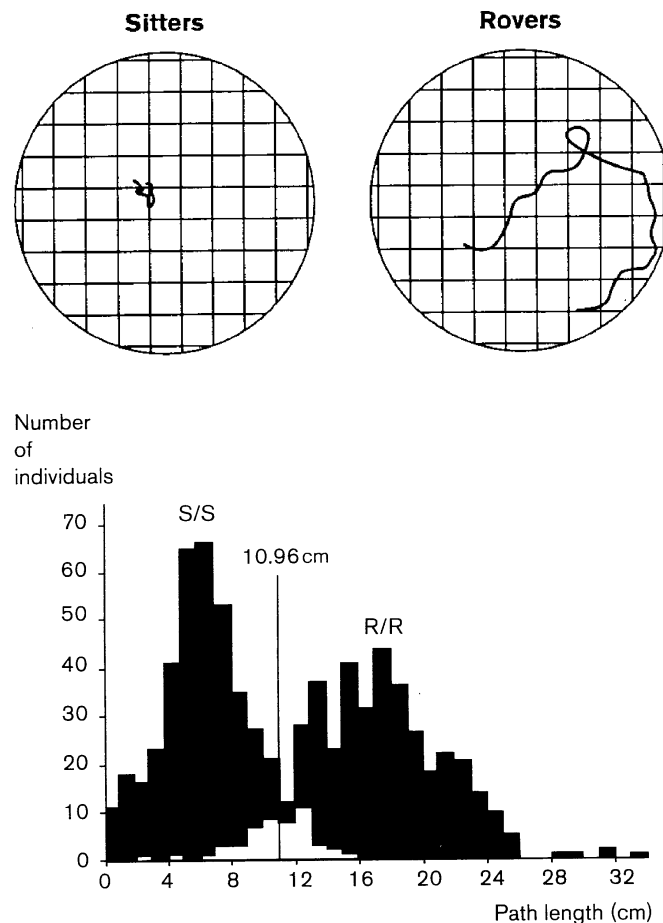


Figure 6

Top: larval foraging behavior of sitter and rover variants in *Drosophila*. Records of the trail left by larvae on a petri dish covered with yeast. Bottom: Bimodal distribution of sitter (S/S) and rover (R/R) larvae in a natural population of *Drosophila* (source: *Behavioral Genetics* 1980 **10**: 291–302. Reproduced with permission)

gravity response in *Drosophila*, there has been no way to identify any of the relevant genes. The behavioral differences are generally due to the contribution of many genes, usually of rather small effect, and thus refractory to conventional mapping techniques. Only their approximate number and putative interactions could be estimated from statistical analyses of selected lines and of F1, F2, and backcross progeny. The advent of QTL mapping (quantitative trait loci) using a widely distributed set of DNA markers has improved the prospects for mapping of such polygenic traits, but significant difficulties still attend the identification of the contributing gene itself.

The identification of natural variants in the *foraging* locus created the possibility of testing directly its response to selection under conditions likely to be relevant to its role in the wild: differences in population

density. In two different experiments, strains of flies mixed for rovers and sitters were cultured at high or low larval density for over 250 generations. In both cases, the resulting high-density strains had become predominantly rover in behavior and the low-density strains predominantly sitter (Sokowloski 1998). That is, selection had eliminated the vast majority of one *foraging* allele from each population.

These findings contrast with those on intraspecies variants of the *period* gene, where the possibility of selection between alleles was inferred from their geographical distribution and made more plausible by the subtle functional differences between alleles. The conditions necessary to impose selection on the *foraging* locus are easier to imagine and implement, and the outcome leads to the conclusion that being genetically rover is advantageous under crowded

conditions, whereas being genetically sitter is advantageous under conditions of low density.

The most common sort of selection experiments have also been the most refractory to analysis. The ability to analyze the genetic basis of the strains whose behavior results from selection for variation in many genes is likely to improve with the advent of genome-wide capabilities to measure differences in gene activity (White et al. 1999). This will likely make it feasible to identify genetic differences between selected strains or natural variants based on detecting asymmetries in the activity patterns of specific genes.

4. Implications for Humans

The foregoing discussion highlights explicit and implicit aspects of the relevance to humans of gene/behavior studies in model organisms. Explicit is the conservation of function (e.g., *period* for circadian rhythms, *Creb* for long-term memory), as well as the wide-ranging nature of the genes involved. Implicit is the importance of specific context in any given gene's effect on behavior and the humility with which the question of genetic influences on behavior must be approached.

4.1 Promiscuity of Behavioral Genes

The breadth of action of the genes influencing behavior has significant implications for speculation about human behavior and its evolutionary origins. The extent to which a gene can vary is constrained by how widely its effects will radiate through the organism. For genes with wide-ranging (pleiotropic) function, this means that variants with severe effects on the gene's function are likely to have deleterious consequences. In this light, evolution looks like a balancing act, selecting for genetic variants that do the most to help the organism without doing too much harm. Sometimes this may take the form of a highly restricted alteration in a gene (e.g., anatomical distribution of V_{1a} in prairie vole vs. montane vole) and other times as a tradeoff between utility and loss of fitness (e.g., adult sluggishness in sitter). An important implication of this point is that some traits will be carried along for no reason other than their connection to some other effect of the same gene. A further corollary is that not all existing traits have been selected for themselves, and thus cannot be automatically assumed to have adaptive value.

4.2 Context

The nature/nurture dichotomy is the traditional way to describe the putative role of context in affecting gene action, but it only conveys one axis of the

problem. Context is as much genetic, consisting of the particular constellation of alleles that an individual possesses, as it is nongenetic, consisting of the myriad environmental factors, both physical and social, that an individual experiences in a lifetime. Interactions between the two further complicate the matter. This means that different alleles of a given gene will not have the same effect in all contexts. A clear example of this was described earlier in the differential ability of two naturally occurring *period* alleles to compensate for temperature fluctuations (see Sect. 3.3). The (Thr-Gly)₁₇ allele makes the individual more sensitive to its external environment than (Thr-Gly)₂₀. This has important implications for analysis of the genetic contributions to particular behaviors (e.g., heritability).

The effect of an allelic variant on a behavior is also heavily dependent on the genetic context. This is known throughout the genetics literature as the sensitivity of a mutation to 'genetic background.' Genetic background is simply another way of describing the particular constellation of alleles at other loci present in the individual. This is exemplified in the interaction between alleles of the *foraging* locus and those of the *Chaser* locus, a gene identified based on the ability of certain of its alleles to convert sitters into rovers. In other words, mild alleles of the sort that occur naturally are even more susceptible to variations in genetic context than severe alleles.

4.3 Testability

Human genetics is primarily a correlative science. This applies as much to behavioral genetics as to any other variety. Hypotheses can be tested against observational data, but actual manipulations or 'test crosses' cannot, of course, be carried out. One of the principal virtues of model organisms, therefore, is experimental testability of gene function and interaction. Given the increasing evidence for conservation of genetic function between species as distant as *Drosophila* and mammals, they represent an important contribution to understanding and interpreting possible genetic influences on human behavior. One of these contributions should be a sense of humility before the problem. As the foregoing discussion elaborates, the road from gene to behavior is neither simple nor direct, and is difficult enough where experimental testing is possible. The difficulties are multiplied for natural variants because of the subtlety of their effects and their consequent elusive nature.

The universal importance of context in the action of genes on behavior presents further obstacles to a simple formulation. If combinations of alleles will give different effects due to interactions with each other and with environmental variables, then the problem takes on the aspect of an infinitely sliding scale. Whenever one variable is changed, the relative effect of many of the others may also change. This mallea-

bility is no doubt a source of robustness in biological systems, as well as a source for evolutionary change. Understanding it remains one of the central questions in the behavioral sciences.

See also: Adaptation, Fitness, and Evolution; Behavioral Genetics; Psychological Perspectives; Body, Evolution of; Brain, Evolution of; Comparative Method, in Evolutionary Studies; Darwin, Charles Robert (1809–82); Evolution, History of; Evolutionary Theory, Structure of; Human Evolutionary Genetics; Neurogenetics and Behavior

Bibliography

- Costa R, Kyriacou C P 1998 Functional and evolutionary implications of natural variation in clock genes. *Current Opinions in Neurobiology* **8**: 659–64
- Dubnau J, Tully T 1998 Gene discovery in *Drosophila*: new insights for learning and memory. *Annual Reviews in Neuroscience* **21**: 407–44
- Dudai Y, Jan Y N, Byers D, Quinn W G, Benzer S 1976 *dunce*, a mutant of *Drosophila* deficient in learning. *Proceedings of the National Academy of Sciences USA* **73**: 1684–8
- Ehrman L, Parsons P A 1981 *Behavior Genetics and Evolution*. McGraw-Hill Inc., New York
- Greenspan R J, Ferveur J-F 2000 Courtship in *Drosophila*. *Annual Reviews in Genetics* **34**: 205–32
- Greenspan R J, Tononi G, Cirelli C, Shaw P J 2001 Sleep and the fruit fly. *Trends Neuroscience* **24**: 142–5
- Hall J C 1994 Pleiotropy of behavioral genes. In: Greenspan R J, Kyriacou C P (eds.) *Flexibility and Constraint in Behavioral Systems*. Wiley, New York
- Konopka R J, Benzer S 1971 Clock mutants of *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences USA* **68**: 2112–6
- Osborne K A, Robichon A, Burgess E, Butland S, Shaw R A, Coulthard A, Pererira S, Greenspan R J, Sokolowski M B 1997 Natural behavior polymorphism due to a cGMP-dependent protein kinase of *Drosophila*. *Science* **277**: 834–36
- Pflugfelder G O 1998 Genetic lesions in *Drosophila* behavioural mutants. *Behavioral Brain Research* **95**: 3–15
- Silva A J, Kogan J H, Frankland P W, Kida S 1998 CREB and memory. *Annual Reviews in Neuroscience* **21**: 127–48
- Sokolowski M 1998 Genes for normal behavioral variation: recent clues from flies and worms. *Neuron* **21**: 463–6
- Wheeler D A, Kyriacou C P, Greenacre M L, Yu Q, Rutila J E, Rosbash M, Hall J C 1991 Molecular transfer of a species-specific behavior from *Drosophila simulans* to *Drosophila melanogaster*. *Science* **251**: 1082–5
- White K P, Rifkin S A, Hurban P, Hogness D S 1999 Microarray analysis of *Drosophila* development during metamorphosis. *Science* **286**: 2179–84
- Wilson E O 1978 *On Human Nature*. Harvard University Press, Cambridge, MA
- Young L J, Nilsen R, Waymire K G, MacGregor G R, Insel T R 1999 Increased affiliative response to vasopressin in mice expressing the V1a receptor from a monogamous vole. *Nature* **400**: 766–8
- Young L J, Wang Z, Insel T R 1998 Neuroendocrine bases of monogamy. *Trends Neuroscience* **21**: 71–5

Young M W 1998 The molecular control of circadian behavioral rhythms and their entrainment in *Drosophila*. *Annual Reviews in Biochemistry* **67**: 135–52

R. J. Greenspan

Genes and Culture, Coevolution of

Cultural transmission can be roughly defined as the transfer of information between individuals by social learning. Genes used to be—and for most species still are—the only means available for the accurate transfer of information across generations. In species where cultural transmission has developed, notably the human, interactions can occur between the two inheritance systems. Gene-culture coevolution refers to the evolutionary phenomena that arise from these interactions.

1. Quantifying Cultural Evolution

The principles of genetic evolution are well established. By contrast, cultural evolution is more contentious. The information that forms the basis of culture is transmitted by social learning—instruction by, or imitation of, other individuals. This information must somehow reside in the brain, but it is unclear how this is achieved. Nevertheless, just as significant insights into genetic evolution were obtained before DNA was identified as the genetic material, cultural evolution can be studied without waiting for definitive progress in cognitive science.

Central to the modern treatment of cultural evolution is the concept of a ‘meme’ (Dawkins 1976; also ‘culturgen’ of Lumsden and Wilson 1981). The meme is defined, in analogy with the gene, as a unit of information that is transmitted by social learning. Surnames are a good example of a meme that is highly congruent to a gene. Surnames are well-defined units that in many societies are transmitted from father to son, just like a gene on the Y chromosome. However, most memes cannot be delimited as precisely, so that copies of the same meme in two individuals may differ in content. Consider the ambiguity of ‘how to use a personal computer.’ In addition, although the meme is defined as information, it is usual to regard as a meme the more readily observable behavior that the information regularly induces.

Memes also differ from genes in that there is a multiplicity of transmission pathways. ‘Vertical’ transmission from parents to their children and ‘horizontal’ transmission between individuals of the same generation have received the most attention.

The classical, although by no means the first, quantitative study of horizontal transmission is due

two rural sociologists, Ryan and Gross (1943), who investigated the spread of hybrid seed corn. This striking technical advance in agriculture was introduced to Iowa farming communities in the late 1920s. Ryan and Gross learned through retrospective interviews when each farmer had adopted this innovation. By the time these interviews were conducted in 1941, adoption was almost complete. The cumulative number of adopters follows an S-shaped curve, which resembles the logistic curve. It is not known whether this resemblance is more than superficial.

The study of vertical transmission begins with Feldman and Cavalli-Sforza (1976). Assume two alternative memes, such as whether or not salty foods are eaten habitually. In general, denote the alternative memes by H and h . Then, the four possible combinations for husband and wife are $H \times H$, $H \times h$, $h \times H$, or $h \times h$. If children acquire their memes by vertical transmission—by social learning from their parents—the four types of families should produce different frequencies of H and h children. More specifically, if the probabilities of an H child in the four types of families are b_3, b_2, b_1 , and b_0 , respectively, then the inequalities $b_3 \geq b_2$, $b_3 \geq b_1$, $b_2 \geq b_0$, and $b_1 \geq b_0$ should hold. A survey of Stanford University students and their parents showed that, for the example of habitual salt use, $b_3 = 60$ percent, $b_2 = 36$ percent, $b_1 = 65$ percent, and $b_0 = 26$ percent, suggesting a fairly strong maternal influence (Cavalli-Sforza and Feldman 1981).

Cultural evolution is said to occur when the frequency (proportion) of a meme in a population changes. Memes are in principle subject to the same evolutionary forces as genes—the analog of mutation, random drift, natural and sexual selection. For example, the meme for habitual salt use may raise mortality and as a result be selected against. On the other hand, an important force acting on memes to change their frequencies, but one not normally applicable to genes, is transmission bias. The spread of hybrid seed corn in the Iowa communities is an example of cultural evolution by transmission bias, since nonadopters are more likely to be converted to adopters than vice versa.

Having briefly described cultural evolution, let us now consider in detail two well-documented instances of an interaction between genetic and cultural evolution. Many more case studies of gene-culture coevolution are discussed in books by Lumsden and Wilson (1981), Boyd and Richerson (1985), and Durham (1991).

2. Adult Lactose Absorption and Drinking the Milk of Domesticated Animals

Perhaps the paradigmatic example of gene-culture coevolution is the process whereby the genetic ability of adult humans to digest lactose (adult lactose

absorption) became associated with their consumption of animal milk.

The fresh milk of domesticated animals, such as goats, camels, and cows, has been an important food in northern Europe and in pastoral societies for about 5,000 to 6,000 years. By contrast, animal milk is not traditionally used by hunter-gatherer and nondairying agricultural groups.

Milk contains a sugar, lactose, which must be digested by the enzyme, lactase, before it can be absorbed in the small intestine. All normal infants have high lactase activity, but during development the majority of individuals experience a drop in lactase activity, which is fully manifest by school age. Hence, adults are of two types—lactose absorbers who have high lactase activity, and lactose malabsorbers who have low lactase activity. Data from families and individuals of mixed ancestry indicate that adult lactose malabsorption is inherited as a simple Mendelian recessive.

When milk is consumed in quantity by a lactose malabsorber, the lactose passes undigested into the colon. Symptoms of intolerance such as severe diarrhea can result. Malabsorbers may also be less efficient at absorbing the calcium in milk.

Many researchers have noted the following pattern. First, populations with high frequencies of absorbers were traditionally dependent on animal milk. Second, populations not known to have used milk have low frequencies of absorbers. The highest frequencies are observed in northern Europe among the Finns, Danes, and Germans. Comparable frequencies are seen among the nomadic tribes of the north African and Arabian deserts, such as the Beja, Tuareg, and Jordanian Bedouin. On the low end are groups like the Japanese, the Pima Indians of Arizona, and the Yoruba of central Africa, none of which have a history of dairying. Although the covariation is impressive, it is not the case that all dairying populations have a high frequency of absorbers, and numerous 'exceptions' can be found around the Mediterranean. Hence, populations with a high frequency of absorbers apparently form a subset of those with a history of milk use.

Lactase activity drops at weaning in nonhuman mammals, suggesting that our ancestors were also malabsorbers. This inference and the above observations led to the proposal of the 'culture-historical' hypothesis, which posits that lactose absorbers gained a selective advantage when children and adults began to drink the milk of domesticated animals (Simoons 1970, McCracken 1971, Kretchmer 1972). Hence, the frequency of absorbers increased by natural selection in populations that had access to animal milk—or at least in some of them.

A variant of the culture-historical hypothesis is the 'calcium absorption' hypothesis (Flatz and Rotthauwe 1973). Whereas the former attributes a general nutritional advantage to the use of milk, the latter claims

that calcium was all important. Recall that high frequencies of absorbers are observed in northern Europe. Here, solar radiation is weak, and the calcium in milk may have provided protection from rickets.

A different hypothesis called the 'reverse cause' argument suggests that dairying was adopted by precisely those populations that could tolerate lactose. The genetic differentiation of human groups is assumed to have occurred before the invention of dairying by some process unrelated to milk use. In particular, drinking milk is not assumed to have conferred a selective advantage.

The two major proposals, the culture-historical hypothesis and the reverse cause argument, thus differ in the temporal priority given to cultural or to genetic change. (A third possibility is that they occur together, in which case the process would be truly coevolutionary.) However, the prevailing view in modern anthropology is that the genetic differences (if any) among human populations are not the cause of the cultural differences among those populations. In particular, the claim that a preexisting genetic contrast, such as in the prevalence of lactose absorbers, would have produced divergent cultural evolution is anathema. Hence, the reverse cause argument, in straying from dogma, has already acquired one strike against it.

More substantively, Holden and Mace (1997) have recently used phylogenetic comparative methods to show that the cultural evolution of milking probably preceded the genetic evolution of lactose absorption. In a sense, this outcome is anticipated by the fact noted above that a high frequency of absorbers is observed in only some of the dairying populations and in none of the nondairying ones. In addition, this analysis reveals that statistical evidence for the calcium absorption hypothesis is weak. Instead, the parsimonious explanation is that northern European populations are similar in having high frequencies of absorbers because they share a relatively recent common ancestral population. On the other hand, the culture-historical hypothesis is supported by a global phylogenetic analysis.

In the previous section, it was explained how a useful innovation may spread by horizontal transmission. Presumably, drinking the milk of domesticated animals can be regarded as a useful innovation, at least for absorbers. Hence, in societies where domesticated animals were available, as short a time as one generation may have sufficed for dairying to become established. If so, the claim that milk use spread first and lactose absorption evolved later may be valid.

Against this, however, the possibility must be entertained that drinking milk was an unattractive option to a society dominated by lactose malabsorbers. In other words, there may be a genetically determined transmission bias against milk use among malabsorbers. Studies conducted in modern industrial

societies reveal little evidence of a clearcut difference in preference for milk between absorbers and malabsorbers. This may be because the availability of many foods, and especially the presence of lactose as an additive, makes it difficult for malabsorbers to associate the symptoms of intolerance with the ingestion of milk. By contrast, in traditional societies where food choices are limited, a difference in preference may be easier to demonstrate.

3. Recessive Hereditary Deafness and Sign Language

The incidence of profound childhood deafness is about one in 1,000. About one half of these cases is hereditary, with recessive genes in particular being responsible for about one third of the total incidence. Estimates of the number of recessive genes range from five to more than 30 depending on the population. Homozygosity for (having two copies of) any one of these genes results in deafness. Among the nongenetic causes are maternal rubella, neonatal jaundice due to rhesus incompatibility, and otitis media.

Sign languages are the naturally occurring forms of linguistic communication among the profoundly deaf. A sign language is not a collection of gestures. Rather, grammatical rules govern the production of words and the expression of thoughts. There are many sign languages just as there are many spoken languages. Sign languages are natural languages, whose origins cannot be determined and which have existed in the deaf communities possibly for many generations.

When deafness is recessive, an affected child may be born to normal heterozygous (having one copy of a gene for deafness) parents. In fact, about 90 percent of deaf children have hearing parents. A hearing child of hearing parents will learn the spoken language used by the parents. Vertical transmission plays an important role in the transmission of a spoken language across generations. By contrast, a deaf child of hearing parents usually cannot learn the local sign language from the parents, simply because the latter do not know it. Hence, the transmission of a sign language across generations is complicated by the fact that a significant fraction of profound childhood deafness is caused by recessive genes. Aoki and Feldman (1991), therefore, suggest that a sign language is a good example of a meme whose transmission and social distribution are subject to strong, if indirect, genetic constraints.

Another factor that interferes with the vertical transmission of a sign language is natural selection against the deaf. Selection is apparently due, not to higher mortality, but to lower marriage and fertility rates among the deaf. In the 1950s a deaf person in Northern Ireland or Japan could expect to have about 30 percent as many children as a hearing person. These observations suggest that a sign language cannot

persist under vertical transmission alone. This prediction is supported by a simple model, which is presented below to illustrate some of the principles of gene-culture coevolution.

Assume the following: (a) One gene causing deafness segregates in the population. There are two alleles (alternative states of a gene), the normal allele, and the mutant allele that causes deafness when homozygous; (b) Nongenetic causes of deafness are ignored; (c) The deaf suffer a higher mortality than the hearing, which is contrary to fact, but has approximately the same evolutionary effect as lower nuptiality or fertility. The relative viability of a deaf person is v ; (d) The recessive allele causing deafness arises by mutation in each generation; (e) Deaf persons tend to marry other deaf persons. The rate of assortative mating is m ; (f) Only the deaf use sign language. Sign language is transmitted vertically from signing parent(s) to deaf children. A deaf child acquires sign language with probability c , kc , or 0 depending on whether there are two, one, or no signing parents. The correspondence with the parameters of Sect. 1 is $b_3 = c$, $b_2 = b_1 = kc$, $b_0 = 0$. When the child is hearing, $b_3 = b_2 = b_1 = b_0 = 0$.

Assumptions (a)–(e) and assumption (f) determine the dynamics of the gene for deafness and the meme for sign language, respectively. Together, they constitute a model for the coevolution of genes and culture. When the strong force of natural selection and the weak force of mutation balance each other, the allele causing deafness will exist at a low equilibrium frequency. The dynamics of the frequency of signers, x , can then be approximately described by the following recursion equation, where the prime indicates the value in the next generation.

$$x' = 2kcvmx \quad (1)$$

Equation (1) is valid when the mutation rate and the frequency of signers are low, which are realistic assumptions.

Clearly, the frequency of signers will decrease and sign language will be lost if $2kcv m < 1$. On the other hand, sign language can persist if $2kcv m > 1$. Estimates of the strength of natural selection and of the rate of assortative mating are available for Japan and Northern Ireland. For example, in Japan in the 1950s, $m = 62$ percent and $v = 30$ percent. Substitution gives $2kcv m = 0.37 \times kc$. This is smaller than one, because the vertical transmission parameter, kc , cannot exceed one.

Thus, this simple coevolutionary model suggests that a sign language cannot be maintained in a deaf community by vertical transmission alone. Some linguists invoke horizontal transmission at schools for the deaf, where naive deaf children can learn from classmates who were fortunate enough to have deaf signing parents.

See also: Cultural Evolution: Theory and Models; Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Memes and Cultural Viruses; Social Evolution: Overview

Bibliography

- Aoki K, Feldman M W 1991 Recessive hereditary deafness, assortative mating, and persistence of a sign language. *Theoretical Population Biology* 39: 358–72
- Boyd R, Richerson P J 1985 *Culture and the Evolutionary Process*. University of Chicago Press, Chicago
- Cavalli-Sforza L L, Feldman M W 1981 *Cultural Transmission and Evolution*. Princeton University Press, Princeton, NJ
- Dawkins R 1976 *The Selfish Gene*. Oxford University Press, New York
- Durham W H 1991 *Coevolution*. Stanford University Press, Stanford, CA
- Feldman M W, Cavalli-Sforza L L 1976 Cultural and biological evolutionary processes, selection for a trait under complex transmission. *Theoretical Population Biology* 9: 238–59
- Flatz G, Rotthauwe H W 1973 Lactose nutrition and natural selection. *Lancet* 2: 76–7
- Holden C, Mace R 1997 Phylogenetic analysis of the evolution of lactose digestion in adults. *Human Biology* 69: 605–28
- Kretchmer N 1972 Lactose and lactase. *Scientific American* 277(April): 71–8
- Lumsden C J, Wilson E O 1981 *Genes, Mind, and Culture*. Harvard University Press, Cambridge, MA
- McCracken R D 1971 Lactase deficiency: An example of dietary evolution. *Current Anthropology* 12: 479–517
- Ryan B, Gross N C 1943 The diffusion of hybrid seed corn in two Iowa communities. *Rural Sociology* 8: 15–24
- Simoons F J 1970 Primary adult lactose intolerance and the milking habit: A problem in biologic and cultural interrelations. II. A culture historical hypothesis. *American Journal of Digestive Disease* 15: 695–710

K. Aoki

Genes and Languages

In a study by Cavalli-Sforza et al. (1988), the spread of anatomically modern man was reconstructed on the basis of genetic and linguistic pieces of evidence: the main conclusion was that these two approaches reflect a common underlying history, the history of our past still frozen in the genes of modern populations. The expression ‘genetic history’ was introduced (Piazza et al. 1988) to point out that if today we find many genes showing the same geographical patterns in terms of their frequencies, this may be due to the common history of our species. A deeper exploration of the whole problem can be found in Cavalli-Sforza et al. (1994). In the following, some examples will be shown of structural analogies between linguistic and genetic geographical patterns that supply complementary information. It is important to emphasize at the outset

that evidence for coevolution of genes and languages in human populations does not suggest that some genes of our species determine the way we speak; this coevolution may simply be due to a common mode of transmission and mutation of genetic and linguistic units of information and common constraints of demographic factors.

1. *The Genetic Analysis of a Linguistic Isolate: The Basques*

The case of the Basques, a European population living in the area of the Pyrenees on the border of Spain and France who still speak a non-Indo-European language, is paradigmatic. What are the genetic relations between the Basques and their surrounding modern populations, all of whom are Indo-European speakers?

Almost half a century ago it was suggested (Bosch-Gimpera 1943) that the Basques are the descendants of the populations who lived in Western Europe during the late Paleolithic period. Their withdrawal to the area of the Pyrenees, probably caused by different waves of invasion, left the Basques untouched by the Eastern European invasions of the Iron Age. In their study of the geographic distribution of Rh blood groups, Chalmers et al. (1948) pointed out that the Rh negative allele, which is found almost exclusively in Europe, has its highest frequency among the Basques. Chalmers et al. hypothesized that modern Basques may consist of a Palaeolithic population with an extremely high Rh negative frequency, who later mixed with people from the Mediterranean area. In more recent times it was suggested (Bertranpetit and Cavalli-Sforza 1991) that the immigrants might have been the Neolithic farmers who came from the Near East and whose modern descendants have a very low frequency of the Rh negative allele.

In summary, genetic analyses have produced the following conclusions:

(a) If Basques originally settled in a large area including the Pyrenean mountains, we can postulate that the genetic differentiation of modern Basques from the neighboring French Béarn population occurred, at the latest, during the Roman period, about 2,000 years BP.

(b) On the assumption that genetic differences are proportional to the times when populations separated (which is strictly valid only when one mechanism of genetic change, technically called genetic drift, is the major cause of differentiation), it seems very plausible that genetic differentiation of Basques from European samples occurred before Celtic people settled in France and originated the French population (about 2,600 years BP). In fact recent analyses of mitochondrial DNA support the idea that the Basques are the descendants of a Paleolithic population (Richards et al. 2000).

(c) The linguistic hypothesis originally put forward by Trombetti (1926) that Basques share a common ancestry with the modern Caucasian speaking people living in the northern Caucasus (see Ruhlen 1991) is in agreement with the genetic evidence: the ancestors of modern Basques could have shared a common genetic origin with populations from the Middle East at the latest before the Neolithic introduction of agriculture in Europe (ca. 6,000–5,500 years ago).

2. *From Basque to Ligurian and Etruscan*

The population and language name ‘Basque’ is interesting by itself. The Basques call themselves *Euskaldunak*: the term ‘Basque’ derives from *Vascos*, a Spanish word. The geographical distribution of the place names ending with the suffix *asc* includes a large region where such place names are significantly represented, from the southwest of France to the Po river delta in Italy. The number of place names whose origin is documented in historical records before 1850 is about 250, and there is general agreement among linguists to associate this suffix with a pre-Indo-European *Ligurian* substratum. Place names are generally very conservative, being notoriously resistant to ethnic stratification; conquerors may impose a language on invaded populations, but usually if they modify place names, they do not change them entirely. The presence of these place names in the Alpine region between Italy and the Swiss Italian-speaking region but not in the region north to the Po River is peculiar.

The linguistic interpretation of this kind of geographical distribution—one linguistically homogeneous area surrounded by two marginal areas with the same linguistic characters—was proposed by the Italian scholar Bartoli (who together with G. I. Ascoli, Trombetti, and Terracini pioneered the new discipline called ‘linguistic geography’) at the beginning of the twentieth century. One of the so-called Bartoli rules says: if a central area is linguistically homogeneous and is surrounded by marginal areas which are linguistically different from the central one but equal between themselves, the language of the marginal areas is almost always older than that of the central area. In fact, the distribution of place names of Ligurian origin surrounds a region of Celtic substratum that flooded any pre-Indo-European relic. Claimed to be among the most ancient ethnic groups of Western Europe, the Ligurians’ original domain far exceeded the boundaries of the modern Italian region bearing their name.

The case of Ligurian is different from that of Basque because the latter is still spoken today while the former can possibly be detected as a linguistic relic (substratum) in the modern dialect spoken by the inhabitants of the relevant region of Italy (‘Liguria’). Although the non-Indo-European nature of the Ligurian language has been debated because of the

very small number of glosses dating at least before the Celts occupied most of the Po river valley in the early sixth century BC, nobody doubts the non-Indo-European origin of the Etruscans (Pallottino 1975), and a natural question is whether we are able to recognize in today's Tuscans the fingerprints of their ancestors' genes. Apart from the number of genetic samples which have to be tested to supply enough information, the Tuscan dialect spoken today certainly cannot be defined as a non-Indo-European or a pre-Roman language: quite the reverse, it was adopted as the 'true' Italian language. What happened is that the transition from Etruscan to Latin was so difficult, the innovation so remarkable, that the process of nearly complete language substitution produced deep subconscious reactions by the speakers, and the effect of these reactions to innovation can be detected as 'fossils' in the language substratum of their descendants: the more different the languages of the conquerors and the conquered populations, the more recognizable these fossils are.

Genetics can be of great help in recognizing these fossils and possibly the population where they originated for at least two reasons. The first is that the transmission of genes is a much slower process than the transmission of a cultural trait such as a linguistic item; therefore genes can keep their identity for a longer time. Second, genetic evolution is very much dependent on population sizes. While it is perfectly legitimate to suppose that a few people can impose their language on large populations—as the Romans in fact did over the 'Romance' Europe—it is much more difficult to think of the same process in genetic terms. A few well-organized people can subdue a whole population, but it is unlikely that they would overwhelm the resident gene pool. It is reasonable to suppose that the Romans subdued the Etrurian inhabitants, but a part of the latter's probable descendants still preserve Etrurian genes. With new DNA markers, and testing of the relevant bone fossils, a better picture of the relationship between Romans and Etruscans might emerge.

3. *Co-evolution of Genes and Languages: The Origin of Indo-European*

Barbujani and Sokal (1990) found a correlation between linguistic and genetic boundaries in Europe. In the majority of cases (22 out of 33) there were also physical barriers that may have caused both genetic and linguistic boundaries. In nine cases there were only linguistic and genetic boundaries but not physical ones: three of them (northern Finland vs. Sweden, Finland vs. the Kola peninsula, Hungary vs. Austria) separate Uralic from Indo-European languages. It remains to be determined whether in these cases linguistic boundaries have generated or

enhanced genetic boundaries, or if both are the consequence of political, cultural, and social boundaries that have played a role similar to that of physical barriers.

The problem of the origin of the Indo-European linguistic family and of the people speaking its languages has roused much more interest over the last years than in earlier times partly owing to the book by Renfrew (1987), who suggested that farmers, beginning to spread from Anatolia around 9,000 years ago, spoke Indo-European languages. His hypothesis was based on the suggestion originally put forward by Ammerman and Cavalli-Sforza (1984) that the spread of Neolithic farming from the Fertile Crescent was due to the spread of the farmers themselves and not only of the farming technology, and on the consideration that migrating people retain their language, if at all possible. Renfrew's hypothesis was criticized by most Indo-European linguists (for a review, see Mallory 1989, Lehmann 1993, pp. 283–8) and did not fare well when contrasted with earlier hypotheses, now identified with the name of another archaeologist, Marija Gimbutas (1985), that Indo-Europeans migrated to Europe from the Pontic steppe area of south Russia from Dniepr to the Volga (which she called 'Kurgan' from the Russian name of mounds covering the graves), beginning with the early Bronze Age, that is, around 5,500 years ago.

Genetic data cannot give strong evidence on dates of migration, especially since the 'Kurgan' area, one of the largest pre-historic complexes in Europe, probably remained very active in generating population expansions for a long time after the Bronze Age. In that area we find at c. 6,000 years ago the Sredni-Stog culture, later (5,500–4,500 years ago) the Yamnaya cultures (formerly called pit-grave cultures) which stretched from the Southern Bug River over the Ural River and which dates from 5,600 to 4,200 years ago. From about 5,000 years ago we begin to find evidence for the presence in this culture of two and four-wheeled wagons (Anthony 1995).

Genetic data on European populations using blood typing (Piazza et al. 1995) and Y-chromosome DNA markers (Semino et al. 2000) have strongly supported a center of radiation in the Ukraine. It has been suggested (Cavalli-Sforza et al. 1994, Piazza et al. 1995) that the hypotheses of Renfrew and Gimbutas should not be treated as mutually exclusive; they may be compatible, as Schrader anticipated as long ago as 1890: 'the Indo-Europeans practiced agriculture at a site between the Dniepr and the Danube where the agricultural language of the European branch was developed' (quoted from Lehmann 1993, p. 279). The settling of the steppe by Neolithic farmers must have occurred after the beginning of their migration from Anatolia, and if the expansions began at 9,500 years ago from Anatolia and at 6,000 years ago from the Yamnaya culture region, then a 3,500-year period elapsed during their migration to the Volga–Don

region from Anatolia, probably through the Balkans. There a completely new, mostly pastoral culture developed under the stimulus of an environment unfavorable to standard agriculture, but offering new attractive possibilities. Our hypothesis is, therefore, that Indo-European languages derived from a secondary expansion from the Yamnaya culture region after the Neolithic farmers, possibly coming from Anatolia and settled there, developing pastoral nomadism.

A new analysis of the problem has been given in an unpublished analysis (Piazza et al. n.d.) of a set of lexical data (200 words) in 63 Indo-European languages published by Dyen et al. (1992). From a linguistic distance matrix whose elements are the fraction of words with the same lexical root for any pair of languages and its transformation to make the matrix elements proportional to time of differentiation, we were able to reconstruct a linguistic tree.

The root of the tree separates Albanians from the others, with a reproducibility rate (the error in reconstructing the tree) of 71 percent. The next oldest branch is Armenian. The simplest interpretation is that the language of the first migrant Anatolian farmers survives today in two direct descendants, Albanian and Armenian, which diverged from the oldest pre-Indo-European languages in different directions but remained relatively close to the point of origin.

If we give to the first split the time depth of the beginning of the expansion of the pre-Indo-European Anatolian farmers, about 9,000 years ago, we can then calculate that the origin of the European branch dates to about 6,000 years ago. The four major branches (pre-Celtic, pre-Balto-Slavic, pre-Italic, pre-Germanic) may correspond to some extent to different migratory waves, but archaeological dating is too scanty to provide unambiguous associations. It is reasonable to suggest that a first migration corresponds to the first branch, the pre-Celts (6,000 years ago, according to the tree), who settled first and went further west. Their only linguistic remnants are still alive today at the extreme of their original range. They profited from being among the first to develop an Iron Age culture, and were able to develop a wide community that spoke their language. Before Roman rule they spread to half of Europe, extending from Spain to France, most of the British Isles, northern Italy, and central Europe.

A detailed description of the other possible expansions will be found in the original paper (for a summary, see also Cavalli-Sforza 2000). In this context it is more interesting to explore how the linguistic lexical data analyzed above correlate with genetic data. The gene frequency geographical maps of the 95 blood type gene frequencies published and analyzed in Cavalli-Sforza et al. (1994) are interpolated at the geographical centers of gravity of the areas where each of the 63 languages is spoken. By taking into account

the spatial autocorrelation of both—linguistic and genetic—sets of data, the following results are obtained: (a) the correlation between genetic and linguistic variables is 69 percent and is statistically significant; and (b) since there is a possible effect of the geographical distance in explaining this correlation, the partial correlation between linguistic and genetic data, given constant geographical distances, has been calculated. The resulting value of 27 percent is still statistically significant and gives evidence for a correlation between genes and languages in Europe that is not only due to the effect of the geographical distance.

4. *Towards a Global Perspective*

More than 5,000 languages are spoken today in the world, and it does not take a linguist to recognize that some languages are more closely related than others due to history. The official origin of historical linguistics can be dated to 1786, when the English judge Sir William Jones advanced the idea that Sanskrit, a classical language in India, Greek, Latin, and possibly Celtic and Gothic (the ancestor of Germanic languages) shared a common origin. These old languages were the first members of a family of languages that would become known as the 'Indo-European' family (or 'phylum'). As Indo-European is the earliest and best studied linguistic family, coevolution of genes and languages has been documented. Since the eighteenth century, however, many other linguistic families or superfamilies have been recognized. The most complete classification on a world basis was proposed by Ruhlen (1991) who lists 17 linguistic families (Khoisan, Niger-Kordofanian, Nilo-Saharan, Afro-Asiatic, Caucasian, Indo-European, Uralic-Yukaghir, Altaic, Chukchi-Kamchatkan, Dravidian, Sino-Tibetan, Austric, Indo-Pacific, Australian, Eskimo-Aleut, Na-Dene, Amerind).

The reconstruction of the relationships above the family level is hotly debated among historical linguists who have yet to agree on the existence of a single tree linking all the existing language families, that is on the possible differentiation of modern languages from a single ancestor language. Even unification at a lower level such as that of the (pre-Columbian) American languages proposed by Greenberg (1987), who grouped them into just three macro-families (Eskimo-Aleut, Na-Dene, and Amerindian), has been strongly opposed by the majority of American linguists. Interestingly, Greenberg's proposal seems to agree with the analysis of genetic markers in extant Native Americans (Cavalli-Sforza et al. 1994) and these three families seem to identify three major migrations suggested by archaeological data. Amerindian speakers appear to have come first (between 30,000 and 15,000 years ago according to genetic data), followed by Na-Dene

speakers and finally Eskimo-Aleut (both in a period between 15,000 and 10,000 years ago). It must be said, however, that at a finer level of classification contemporary Amerindian speakers show high genetic variability, and this is not easy to reconcile with linguistic taxonomy.

Even without an agreed genealogy of the linguistic families covering all tongues spoken today, it is relevant to note the impressive one-to-one correspondence of the genetic phylogeny of the world populations with the classification into the 17 large linguistic families listed above (Cavalli-Sforza et al. 1988). This correspondence is expected because there are important similarities between the evolution of genes and languages. In either case: (a) a change which first appears in a single individual can subsequently spread throughout the entire population (for genes they are called mutations; they are rare, are passed from one generation to the next and can, over many generations, eventually replace the ancestral type; linguistic innovations are much more frequent and can also pass between unrelated individuals); and (b) the dynamics of change is affected by the same demographic pressures, isolation, and migration. Two isolated populations differentiate both genetically and linguistically because isolation, which could result from geographic, ecological, or social barriers, reduces the likelihood both of marriages and cultural exchanges and, as a common result, reciprocally isolated populations will evolve independently and gradually become different. Both genes and languages will drift apart regularly over time, the former slowly, the latter much more quickly.

In principle, therefore, the linguistic tree and the genetic tree of human populations should agree since they reflect the same history of population splitting and subsequent independent evolution. The different rate of change, however, is a major source of divergence: one language can be replaced by another in a relatively short time. In Europe, for example, Hungarian is spoken in a land surrounded by Indo-European speakers but it belongs to the Finno-Ugric subdivision of Uralic. At the end of the ninth century AD, the nomadic Magyars left their land in Russia and invaded Hungary. The number of conquerors was probably less than 30 percent of the conquered population so that their genetic contribution was limited, but they imposed their language on the local Romance-speaking population. Today all Hungarians speak a Uralic language, but barely 10 percent of their genes can be attributed to the Uralic conquerors. Generally it is intuitive that the total substitution of one language for another occurs more easily under the pressure of a strong political power of the newcomers, as witnessed in the Americas. The case of Basques, on the other hand, shows that separate languages spoken in nearby countries can remain relatively unaffected for thousands of years, even when their genes experience a partial substitution.

It is remarkable that, despite the above sources of confusion, the correlation between genes and languages has been maintained through the centuries until today and is still statistically significant. The ties between biology and linguistics were already evident since the times of Darwin, who in chapter XIV of his *The Origin of Species* wrote:

If we possessed a perfect pedigree of the mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world; and if all extinct languages, and all intermediate and slowly changing dialects, were to be included, such an arrangement would be the only possible one. Yet it might be that some ancient language had altered very little and had given rise to few new languages, whilst others had altered much owing to the spreading, isolation, and state of civilization of the several co-descended races, and had thus given rise to many new dialects and languages. The various degrees of difference between the languages of the same stock, would have to be expressed by groups subordinate to groups; but the proper or even the only possible arrangement would still be genealogical; and this would be strictly natural, as it would connect together all languages, extinct and recent, by the closest affinities, and would give the filiation and origin of each tongue.

The increasing resolving power of modern genetic data makes it possible to follow Darwin and to use the genetic phylogeny of our species to infer the earliest branches of a hypothetical linguistic tree. The most comprehensive genetic phylogeny reconstructed in Cavalli-Sforza et al. (1988) was used by Ruhlen (1994) to draw the tree of origin of human languages (some reference dates from genetic and archaeological evidence have been added). The oldest linguistic families must be African: Khoisan is probably the oldest and Afro-Asiatic the most recent, while Niger-Kordofanian and Nilo-Saharan believed by some linguists to descend from an ancestor tongue, and the Congo-Saharan, were probably spoken at an intermediate time. A more exhaustive discussion of this hypothetical tree can be found in Cavalli-Sforza (2000). As the genetic data improves with the inclusion of more representatives from those geographical areas of the world where the sampling is still scanty, the tree will be more complex but it is likely that its main features will remain unchanged.

In conclusion, our present genome keeps the record of its past evolution with an impressive richness of detail that is also reflected by our languages. Genes and languages contribute to the understanding of human history by highlighting human diversity; both are instrumental in giving some of the silent voices of our past a chance to be heard.

See also: Cultural Evolution: Overview; Evolution and Language: Overview; Languages: Genetic Classification; Linguistics: Overview

Bibliography

- Ammerman A J, Cavalli-Sforza L L 1984 *Neolithic Transition and the Genetics of Populations in Europe*. Princeton University press, Princeton, NJ
- Anthony D W 1995 Horse, wagon & chariot: Indo-European languages and archaeology. *Antiquity* **69**: 554–65
- Barbujani G, Sokal R R 1990 Zones of sharp genetic change in Europe are also linguistic boundaries. *Proceedings of the National Academy of Sciences* **87**: 1816–9
- Bertranpetit J, Cavalli-Sforza L L 1991 A genetic reconstruction of the history of the population of the Iberian peninsula. *Annals of Human Genetics* **55**: 51–67
- Bosch-Gimpera A 1943 El problema de los orígenes vascos. *Eusko-Jakintza* **3**: 39
- Cavalli-Sforza L L 2000 *Genes, Peoples, and Languages*. North Point Press, New York
- Cavalli-Sforza L L, Menozzi P, Piazza A 1994 *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ
- Cavalli-Sforza L L, Piazza A, Menozzi P, Mountain J 1988 Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences* **85**: 6002–6
- Chalmers J N M, Ikin E W, Mourant A E 1948 Basque blood groups. *Nature* **162**: 27
- Gimbutas M 1985 Primary and secondary homeland of the Indo-Europeans. *Journal of Indo-European Studies* **13**: 185–202
- Greenberg J H 1987 *Language in the Americas*. Stanford University Press, Stanford, CA
- Lehmann W P 1993 *Theoretical Bases of Indo-European Linguistics*. Routledge, London
- Mallory J P 1989 *In Search of the Indo-Europeans: Language, archaeology and myth*. Thames and Hudson, London
- Pallottino M 1975 *The Etruscans*. Penguin Books, New York
- Piazza A, Cappello N, Olivetti E, Rendine S 1988 A genetic history of Italy. *Annals of Human Genetics* **52**: 203–13
- Piazza A, Rendine S, Minch E, Menozzi P, Mountain J, Cavalli-Sforza L L 1995 Genetics and the origin of the European languages. *Proceedings of the National Academy of Sciences* **92**: 5836–40
- Piazza A, Minch E, Cavalli-Sforza L L n.d. The Indo-Europeans: Linguistic tree and genetic relationships. *Antiquity*
- Renfrew C 1987 *Archaeology and Language: The Puzzle of Indo-European Origins*. Cambridge University Press, New York
- Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Y, Gölge M, Dimitrov D, Hill E, Bradley D, Romano V, Calì F, Vona G, Demaine A, Papiha S, Triantaphyllidis C, Stefanescu G, Hatina J, Belledi M, Di Rienzo A, Oppenheim A, Norrelletto A, Nørby S, Al-Zaheri N, Santachiara-Benerecetti S, Scozzari R, Torroni A, Bandelt H-J 2000 Tracing European founder lineages in the near eastern mtDNA pool. *The American Journal of Human Genetics* **67**: 1251–76
- Ruhlen M 1991 *A Guide to the World's Languages*. Stanford University Press, Stanford, CA
- Semino O, Passarino G, Oefner P J, Lin A A, Arbuzova S, Beckman L E, De Benedictis G, Francalacci P, Kouvatzi A, Limborska S, Marcikiai M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti A S, Cavalli-Sforza L L, Underhill P A 2000 The genetic legacy of paleolithic *Homo sapiens sapiens* in extant Europeans: A Y chromosome perspective. *Science* **290**: 1155–9

Trombetti A 1926 *Le origini della lingua Basca*. Bologna, Italy

A. Piazza

Genetic Counseling: Historical, Ethical, and Practical Aspects

Genetic counseling is one of the most important direct applications of human genetic knowledge. Genetic counseling refers to the totality of activities that (a) establish the diagnosis of genetic disease, (b) assess the occurrence risk, (c) communicate to the patient and family the chance of recurrence, (d) provide full and unbiased information within a caring relationship regarding the many problems raised by the disease and its natural history, including the potential ethical conflicts as well as the medical, economic, psychological, and social burdens, and (e) provide information regarding the reproductive options to be taken, including prenatal diagnosis, and refer the patients to the appropriate specialists. Genetic counseling should offer guidance but will allow those being counseled to arrive at their own choices, both before and after genetic procedures (Vogel and Motulsky 1997; World Health Organisation [WHO] 1998).

1. Paradigm Development and Concepts of Genetic Counseling

Pre-scientific knowledge of heritability in humans dates back to the early Greek physicians and philosophers and it can be assumed that there have been practical measures, including advice to individuals and couples as to how to use such knowledge (Vogel and Motulsky 1997). The idea of an improvement of mankind by rationalized mate selection was a major aspect of ancient and Renaissance state and social utopias (Plato, *Politeia* (about 374 BC); Thomas More, *Utopia* (1516); Fra Tomaso Campanella, *The City of the Sun* (1623)). In the medical literature of the eighteenth and nineteenth centuries, reports on the familial recurrence of particular disorders begin to accumulate. An early and noteworthy example is the book of the English physician Joseph Adams (1756–1818) *A Treatise on the Supposed Hereditary Properties of Diseases* (1814), where an attempt is made to lay out the fundamentals of genetic counseling in affected families.

Towards the end of the nineteenth century, two paradigms established human genetics as a science. Gregor Mendel (1822–1884) founded the concept of the gene as the basic factor of inheritance (1866), while Francis Galton (1822–1911) introduced statistical

Table 1
Paradigm shifts in applied human genetics

Paradigm	Basis	Goal	Measure
Eugenics	Emotionally carried social utopia	Improvement of the gene pool	Directiveness, compulsion, coercion
Prevention	Rational calculation	Reduction of individual suffering, economic distribution of resources	Paternalism
Psycho(socio)-biology	Complex interaction between counselor (doctor) and client (patient)	Help in individual decision making	Empathy, nondirectiveness

methods to investigate and interpret the familial clustering of traits and thus founded the biometrical paradigm. Until the present, biometrics and Mendelian genetics have remained the basis of risk determination as the naturalistic component of genetic counseling. Galton coined the term 'eugenics' and favored measures to enhance propagation of individuals with supposedly high genetic qualities (positive eugenics). In Germany, around the turn of the nineteenth to the twentieth century, Wilhelm Schallmeyer (1857–1919), Alfred Hegar (1830–1914), and Alfred Ploetz (1860–1940) developed the idea of negative eugenics, a concept of decreasing the number of offspring of individuals with supposedly below average hereditary qualities, or of preventing congenitally diseased individuals altogether, with the aim of improving the gene pool of future generations and especially that of the presumed superior Nordic race (see Ploetz 1895).

In many countries during the first half of the twentieth century, positive and negative eugenics formed the ideological basis of state-enforced measures to exert reproductive control over individuals. Between 1911 and 1930, 24 US states proclaimed sterilization laws and eugenic thought also played a role in the enactment of restrictive immigration laws. In Europe, Denmark was the first country to proclaim a Law on Genetic Hygiene (1929), in which voluntary sterilization for genetic reasons was regulated. In Germany, the Law for the Prevention of Congenitally Diseased Progeny (1933), followed by a series of other laws and decrees, mandated and organized the sterilization of individuals with certain diseases, which were thought to be genetically determined. In Germany, eugenics, racial hygiene, and National Socialist ideology formed a gloomy alliance. The involvement of German geneticists in coercive measures and eventually the Holocaust to sections of the population that were declared to be inferior has been well documented by Müller-Hill (1997). It is beyond doubt that this historical burden strongly influences the current debate on ethics and genetics in Germany and elsewhere.

From about 1930 onwards, the eugenic movement outside Germany began to lose power, but it was not

until the 1950s that the population-directed eugenic orientation was slowly replaced by a medically oriented practice with genetic disease prevention as a major focus. Nevertheless, eugenic thinking, whether acknowledged or not, has continued to play a role until the present time, especially outside the Western world (Wertz 1997). When, in 1947, Sheldon Reed coined the term 'genetic counseling,' it became clear that eugenics and what was then understood to be genetic counseling, were no longer identical. According to Reed, 'genetic counseling is a type of social work entirely for the benefit of the whole family without direct concern for its effect upon the state or politics' (1974). This definition marks a change of paradigm in applied human genetics, away from the concern about the genetic make-up of the population and towards the attitude that decision-making with regard to genetic risks should be left to the individual, even if this may create or increase social costs or worsen the gene pool. Thus, one of the fundamental ethical principles in medicine, the respect for the autonomy and self-determination of persons, is also the cornerstone of genetic counseling.

Genetic counseling has thus gone through three overlapping historical phases (Table 1). First, the eugenics phase, mainly rooted in England, the United States of America, Scandinavia, and Germany, was based on the social and biological utopia of general mental and physical health and the optimization of humankind. In order to attain these goals, depending on political circumstances, directive measures including direct and indirect coercion were applied. Eugenics was scientifically ill-founded, and the definition of inferior or superior genetic qualities a matter of ideology. Second, the preventive-medical phase laid its emphasis on the reduction of individual suffering through the prevention of genetic diseases and disabilities, in order to economize rare medical resources. It was founded on the rational calculation that any means towards that end would find acceptance as being in the immediate interest of the individual. Strategies were those of benevolence, authority, and paternalism. Third, in the psycho-(sociobio-)logical phase, the focus of genetic counseling is to help the individual or the family to manage a problematic

situation caused by genetic disease. The basis of this help is a complex interaction between the counselor (a specialized physician or a non-medically trained expert) and the client (patient) in a nondirective process of communication. Help is provided with the understanding that genetic diseases are often associated with special psycho-social problems and ethical conflicts which deserve to be managed individually. Thus, any direct or indirect coercive influence is to be avoided. Nondirectiveness has become a basic concept of genetic counseling, to which a majority of counselors worldwide feel obliged (Wertz 1997).

The history of genetic counseling shows that its purpose has always been value-determined and has never been seen outside the context of social politics in which it was embedded. In countries that acknowledge human rights as defined in the democratic state theories of the eighteenth century, the only defensible position is that of the autonomy of the individual to make decisions regarding genetic counseling and genetic diagnosis, and to draw consequences for life and family planning. The healthcare system, on the other hand, carries the ethical obligation to provide full information and the tools that may enable the individual, depending on the state of knowledge, the emotional situation, and moral positioning, to make use of that information.

1.1 Excursus: Ethics and Genetic Counseling: The Respect for Personal Autonomy

Today the term 'eugenics' is sometimes used to describe the cumulative result of individual choices after genetic counseling, especially in view of reproductive choices. The decreasing prevalence of births of children with Down Syndrome in Western Europe, especially in the UK, may be regarded as a result of individual choices to undergo prenatal diagnosis and subsequently to terminate a pregnancy after a positive test result. Whether this type of individual eugenics is socially desirable and a benign development has been discussed with some controversy by Wertz (1998). Critics argue that free, autonomous choices (after genetic counseling) not only depend on surrounding social, cultural, and other factors, but also on the full and honest presentation of information in genetic counseling, i.e., on access to valid information. In the absence of correct and unbiased information, free choices are not possible and are compromised, and information and genetic counseling becomes a means of furthering attitudes and goals of health professionals. A worldwide survey of 2,091 genetics professionals in 36 nations conducted by Wertz (Nippert and Wolff in Germany) suggests that eugenic thought underlies the perception of the goals of genetics and genetic counseling in many countries. In developing nations of Asia and Eastern Europe especially, directive counseling is practiced to influence individual

decisions. Intentionally pessimistically biased information is routinely provided after positive results from prenatal diagnosis in order to avoid the birth of affected children (Wertz 1998, Cohen et al. 1997). The data from this survey suggests that the concept of non-directive counseling found in English-speaking nations and Western Europe is an aberration from the rest of the world.

2. Indications for Genetic Counseling

In medicine in general, a measure is 'indicated,' when, after weighing up the risks and benefits, it appears medically required. This definition is hardly applicable to genetic counseling, because the recognition of risks and benefits of knowledge, as well as any actions that may result from such knowledge, will emerge only in the counseling process itself, and their appraisal is often more strongly determined by individual life circumstances than by medical judgment. On the other hand, the professionalization of genetic counseling, as a specialization within medicine, and jurisdiction in regard to correctness and completeness of any medical advice, have brought about a core service that every counselee may duly expect to receive: A scientifically founded statement regarding the etiology and the (recurrence) risk of a genetic condition, comprehensive information and counseling regarding diagnostic options, and referral to preventive, therapeutic, and/or socially alleviating measures.

Generally speaking, an indication for genetic counseling is given whenever an individual regards it as a means to help solving a personal problem related to an existing or suspected genetic risk. Table 2 summarizes the most frequently encountered situations. With increasing awareness of old and novel pre- and post-natal methods of genetic diagnosis, genetic counseling is increasingly sought by individuals with genetic risks not immediately apparent to themselves. Concerns,

Table 2
Indications for genetic counseling

-
1. Birth of a child with congenital disease or disability
 2. One or more relatives suffering from (suspectedly) genetic disease or disability
 3. Affected counselee
 4. Consanguineous couple
 5. Recurrent miscarriage
 6. Infertility
 7. Mutagenic influences
 - 7.1 Advanced maternal age
 - 7.2 Pre- or postconceptional mutagen exposure
 8. Teratogenic exposure
 - 8.1 Infections
 - 8.2 Drugs
 - 8.3 Alcohol
 - 8.4 Radiation
-

including moral doubts, may have been raised, for example, by offers of prenatal screening (maternal serum marker and ultrasound screening), or carrier detection, within or outside the framework of a defined screening programme, because such offers may bring up the question of termination of the pregnancy.

3. The Counseling Process

According to a statement issued by an Ad hoc Committee on Genetic Counseling of the American Society of Human Genetics, genetic counseling is a communication process which:

deals with the human problems associated with the occurrence or risk of recurrence of a genetic disorder in a family. This process involves an attempt by one or more appropriately trained persons to help the individual or the family to 1. comprehend the medical facts, including the diagnosis, the probable course of the disorder, and the available management; 2. appreciate the way heredity contributes to the disorder and the risk of recurrence in specified relatives; 3. understand the options for dealing with the risk of recurrence; 4. choose the course of action which seems appropriate to them in view of their risk and their family goals and act in accordance with that decision; 5. make the best possible adjustment to the disorder in an affected family member and/or to the risk of recurrence of that disorder (Fraser 1974; Ad hoc Committee on Genetic Counseling 1975).

The counseling process is usually initiated by clarifying the counselee's motivation for consultation. What is the background to the problem, and what are the expectations? The counselor should explain, in general terms, what genetic counseling can achieve, and where its limits lie. This is important, because expectations are frequently unrealistic. It needs to be explained, for example, that it is technically not possible to exclude any kind of genetic risk, or to guarantee a healthy child. The next step is to record the personal and family medical history in the form of a genealogical tree, encompassing at least all first and second degree relatives. This is followed by providing information regarding the most important medical facts concerning the disease in question (natural history of the disease, the likely degree of suffering it may confer, its treatability) as well as the likelihood of its occurrence or recurrence. It is important to give unbiased, objective information. During the later phases of the counseling process, the counselor will often return to this information in order to relate it to the personal views of the counselee. The 'prior' risks for a genetic condition as deduced from its pattern of inheritance (monogenic dominant or recessive, autosomal or sex-linked, co-determined by several genes and/or the environment) may be modifiable by genetic diagnostic tests so that a 'posterior,' usually more precise, risk figure emerges. Ever more often, with the Human Genome Project nearing its completion and

concomitant technical developments becoming routine practice, firm confirmation or exclusion of carrier status for any particular genetic disposition will be the rule, rather than the exception. Nevertheless, the phenotypic outcome will often be variable, due to environmental factors and complex interactions between genes. In the counseling process, the resulting uncertainty needs to be explained and worked upon with great care. Without thoroughly understanding the chances and limits of genetic testing, an informed consent to any such procedure cannot be attained. Informed consent is generally held to be the prerequisite for any genetic testing. The decision whether to make use or not of diagnostic offers can only be made sensibly in a highly personalized context, so that genetic counseling is required to be nondirective (Wolff and Jung 1995).

During the counseling process it may be necessary to make a physical examination of the counselee and/or of family members in order to search, for example, for early signs of manifestation of a late-onset disorder or for microsymptoms of a disease. The impact of such examination needs to be explained beforehand, because counselees and their family members may wish to help clarify genetic risks for others, but not necessarily for themselves. It is sometimes difficult to decide to what extent information from the family history which may confer a genetic risk to the counselee needs to be discussed, if this information has not been central to the counselee's motivation for consultation. This is particularly delicate when no therapeutic or preventive options exist. If they do, however, the counselee has a right to comprehensive information and counseling.

4. The Determination of Genetic Risks

The determination of genetic risks in any individual case is the essential naturalistic component of genetic counseling. It requires expertise in general and human genetics, including formal (mathematical) genetics, population genetics, and clinical genetics, as well as an awareness (attainable in this fast-moving field only through continuous further self-education) of novel developments, especially in the differential genetic diagnosis. In the following discussion, we provide a summary of the most important groups of conditions where a risk assessment is often sought and a brief outline of risk provision. Often, the prior genetic risks need to be modified by using information from the pedigree or genetic tests (conditional probabilities). By employing Bayesian methodology this information can be used to arrive at posterior (final) risks (Harper 1998).

If a disease is (mainly) due to the effect of a single gene, the phenotype may be inherited in a recessive or a dominant fashion. The disease may manifest itself in a sex-dependent way, if the causative gene is not

located on the autosomes (the chromosomes shared by both sexes in equal number) but rather on the sex chromosomes (X and Y). In an autosomal-recessive condition, the recurrence risk in a sibship is 1/4, but usually low (less than 1 percent) for second-degree relatives. If a parent has an autosomal-dominant condition, there is a 50 percent chance that each child will also carry the disease-causing mutation. If neither parent has a condition that is known to be autosomal-dominant, the recurrence risk in a sibship is usually negligible, because the presumption is that a spontaneous mutation has occurred in the germ-line leading to the affected individual. In X-linked recessive disease, if the mother is a carrier, half of all her sons will develop the full clinical picture, and half of her daughters will be carriers. Mothers and daughters are usually unaffected or exhibit symptoms to a lesser degree (due to non-randomness of X-inactivation during early development). In Y-linked disease (such as sub-fertility), the trait is passed on from fathers to sons. In both recessive and, most notably, dominant conditions, the inheritance patterns may be obscured by variable penetrance (proportion of carriers manifesting the phenotype at all) and expressivity (degree of disease severity). Greatly reduced penetrance and expressivity, i.e., where only a small fraction of carriers suffer from the condition, leads to the concept of genetic predisposition (susceptibility) for disease, which may be modulated by other genes or environmental factors. Polygenic and multifactorial inheritance are the etiology of most of the common diseases (such as coronary heart disease, obesity, diabetes mellitus) and the majority of cases of birth defects. In these categories, empirical risk figures apply. Numerical chromosomal aberrations are usually sporadic, with a low empirical recurrence risk in a sibship, but may also run in families, e.g., structural aberrations, such as translocations, which may be inherited as balanced or unbalanced in offspring in a more or less apparent Mendelian fashion.

5. Genetic Counseling in the Social Context

How have genetic counseling and its supposed high standards fared in common practice? Throughout the world and even in Europe, there is a huge variance in the professional identity and training of providers of genetic counseling services. In Germany, most counseling services are provided by physicians who are specialists in medical genetics (Nippert et al 1997) and in the UK, genetic counseling is provided primarily by nurses (Raeburn et al 1997), whereas in the US and Canada a substantial amount of genetic counseling is provided by counselors trained to the master's degree level. This implies that practice differs between countries. Surveys which compare the attitudes of genetic service and counseling providers have revealed a remarkable difference in attitudes and values. An

international survey carried out in 37 nations in 1994–1995 found that, in most of the surveyed countries, genetic counseling, especially in the context of prenatal diagnosis (PD), is directive in that the counselor provides the client with intentionally slanted information in order to influence the client's decision (Wertz 1998). The survey shows that, depending on the diagnosed conditions of a fetus, on the culture's view of the condition, and on the counselor's personal view about the morality of abortion, he or she deliberately provides more optimistically or pessimistically biased information.

Although published policies and guidelines exist (Fraser 1974, German Society of Human Genetics 1996, Andrews et al 1994, WHO 1998, Fryer and Lister Cheese 1998, Holtzman and Watson 1997), the type of information and the way information is communicated vary within countries, and standard practices can hardly be found. Obviously genetic counseling practices are shaped by the characteristics of the various healthcare systems, as well as by the personal characteristics of the counselor, such as sex, professional expertise, and type of employment. The process and outcomes of genetic counseling have been documented by a few empirical studies (Sorenson 1992). The outcomes have been defined primarily on the basis of patient knowledge and patient satisfaction, but so far have hardly paid any attention to the social aspects of the information (Clarke et al 1996). Concern has been voiced that in the future more counseling will be provided by professionals not trained in genetics. Although there is insufficient empirical evidence that genetic counselors successfully achieve autonomous, informed decision-making, research from Europe and the US suggests that non-geneticists do not set a high value on providing information that allows informed decision-making before undergoing tests such as prenatal diagnosis (Dodds 1997).

With the development of more new molecular tests, implementation of these tests in various fields of medicine will drive the diversification of genetic counseling practice. Neither general healthcare providers nor non-genetics specialists have sufficient professional education and expertise in genetics, molecular testing, and genetic counseling to perform this task (Holtzman and Watson 1997). How to target the education for providers via medical curricula and continuing education remains an unsolved challenge. So far, no country has developed any strategy or has made any far-reaching efforts. Genetic specialists are arguably the most valid source for providing the needed expertise and information for their medical colleagues to facilitate their understanding of basic genetic principles and the application of molecular testing, including pre- and post-test counseling, in clinical practice. However, the more genetics becomes part of general medicine and the more genetic testing is introduced into the various specialty areas, the more the task to provide information and to interpret test

results will fall on primary care physicians and non-genetics specialists. Probably the future of genetic services will be provided in the way envisaged by Wertz:

It is perhaps more likely that most genetic services will be provided in the same manner as other types of medicine. In managed care systems this may mean hurried, assembly-like encounters that focus on tests and treatments rather than on providing information and counseling. Most primary care providers may be ill-equipped to provide information and counseling needed in any case ... Many will wonder why genetics was ever expected to produce superior provider-patient relationships, to uphold superior ethical standards, or to allow greater patient autonomy than other fields of medicine (1997, p. 343).

Most likely the future of genetic testing and counseling will be shaped by economic interests, because of the need for financial returns on heavy investments made in the field of molecular testing, by reimbursement practices, and by the extent to which national healthcare policies will encourage genetic testing. It remains unclear whether the main ethical concepts attached to genetic counseling, such as nondirectiveness, individual autonomy, and greater equality in the doctor-patient relationship, will survive or will wither away during the integration of genetic tests into routine medical practice. Hopefully genetic counseling will retain some of its best elements, namely respect for patients' rights to valid, non-slanted information and for patient autonomy in reproductive decision-making.

See also: Ethical Dilemmas: Research and Treatment Priorities; Ethical Issues in the 'New' Genetics; Ethics and Psychiatry; Eugenics, History of; Genetic Screening for Disease-related Characteristics; Genetic Testing and Counseling: Educational and Professional Aspects; Risk Screening, Testing, and Diagnosis: Ethical Aspects

Bibliography

- Ad hoc Committee on Genetic Counseling of the American Society of Human Genetics 1975 Genetic counseling. *American Journal of Human Genetics* 27: 240–2
- Andrews L B, Fullarton J E, Holtzman N A, Motulsky A G (eds.) 1994 *Assessing Genetic Risks. Implications for Health and Social Policy*. National Academy Press, Washington, DC
- Clarke A, Parsons E, Williams A 1996 Outcomes and process in genetic counseling. *Clinical Genetics* 50: 462–9
- Cohen P E, Wertz D C, Nippert I, Wolff G 1997 Genetic counseling practices in Germany: A comparison between East German and West German geneticists. *Journal of Genetic Counseling* 6: 61–80
- Dodds R 1997 *The Stress of Tests in Pregnancy: Summary of a National Childbirth Trust Screening Survey*. National Childbirth Trust, London
- Fraser F C 1974 Genetic counseling. *American Journal of Human Genetics* 26: 636–59
- Fryer A E, Lister Cheese I A F 1998 *Clinical Genetic Services: Activity, Outcome, Effectiveness and Quality*. Report from the Clinical Genetics Committee of the Royal College of Physicians of London, The Royal College of Physicians of London, London
- German Society of Human Genetics 1996 Position paper, <http://www.gfhev.de>
- Harper P 1998 *Practical Genetic Counseling*. 5th ed. Butterworth-Heinemann, Oxford, UK
- Holtzman N A, Watson M S 1997 *Promoting Safe and Effective Genetic Testing in the United States. Final Report of the Task Force on Genetic Testing*. Johns Hopkins University Press, Baltimore. <http://www.med.jhu.edu/tfgtelsi/>
- Human Genome Project, http://www.ornl.gov/TechResources/Human_Genome/home.html
- Müller-Hill B 1997 *Murderous Science: Elimination by Scientific Selection of Jews, Gypsies and Others in Germany 1933–1945*. Oxford University Press, Oxford, UK
- Nippert I, Horst J, Schmidtke J 1997 Genetic services in Germany. *European Journal of Human Genetics* 5 (suppl 2): 81–8
- Ploetz A 1895 Racial hygiene. In: Weingart P, Knoll J, Bayertz K (eds.) *Rasse, Blut und Gene*. Suhrkamp, Frankfurt/Main, Germany
- Raeburn S, Kent A, Gillot J 1997 Genetic services in the United Kingdom. *European Journal of Human Genetics* 5 (suppl 2): 188–95
- Reed S C 1974 A short history of genetic counseling. *Social Biology* 21: 332–9
- Sorenson J R 1992 What we still don't know about genetic screening and counseling. In: Annas G J, Elias S (eds.) *Gene Mapping: Using Law and Ethics as Guides*, Oxford University Press, New York, pp. 203–14
- Vogel F, Motulsky A G 1997 *Human Genetics*. Springer, Berlin Heidelberg, New York
- Wertz D C 1997 Society and the not-so-new genetics: What are we afraid of? Some future predictions from a social scientist. *Journal of Contemporary Health Law Policy* 13: 299–346
- Wertz D C 1998 Eugenics is alive and well: A survey of genetic professionals around the world. *Science in Context* 11: 493–510
- Wolff G, Jung C 1995 Nondirectiveness and genetic counseling. *Journal of Genetic Counseling* 4: 3–25
- World Health Organisation 1998 *Proposed International Guidelines on Ethical Issues in Medical Genetics and Genetic Services*. World Health Organisation Human Genetics Programme, Geneva. <http://www.who.int/ncd/hgn/hgnethic.htm>

J. Schmidtke, I. Nippert, and G. Wolff

Genetic Engineering

The term 'genetic engineering' stands for human alteration of the genetic code of an organism, so that its biosynthetic properties are changed. The major applications are for the industrial production of desired peptides or proteins, or to alter the biological capabilities of the organism. These techniques have

been used to develop crops with agronomically useful changes, such as pest resistance and ripening properties that allow for shipment. Surprising results have been obtained by silencing genes in experimental organisms, as well as the production of animal models of human disease by deriving strains of animals with mutated human genes. Much attention is being given to the production of products and methods to modify an individual's cells to treat human disease. Although the majority of activity at the present time is directed toward traits governed by a single gene, commercial and academic investigators are hopeful that complex diseases and behavior can also be modified.

1. Biological Background

The technology has developed in the context of our understanding of genetics. A vital property is the production and maintenance of variability in populations. One of the mechanisms that produces such variability is recombination. Information for a number of peptides as well as the controlling mechanisms for their synthesis are inherited together on a single chromosome. In diploid organisms, exchanges between homologous chromosomes is an important source of variation. This process occurs in the context of DNA replication, including the enzymatic unwinding of the double helix and synthesis of the new, complementary chain of nucleic acids.

Cellular properties may be altered by external mechanisms. One of the first systems that was extensively studied, and provided the tools that are currently used, is the phenomenon of infection of bacteria by bacteriophages, the bacterial viruses. The usual direction of events is followed. The message is coded in the DNA sequence. It is then transcribed into the sequence of the intermediate messenger RNA molecules, and then translated into protein. In the lytic cycle, the viral DNA instructions are used to synthesize new viral DNA and protein that can assemble into new viral particles that destroy the cell and are then released into the environment to infect other host cells. In the lysogenic cycle, the DNA of the virus integrates into the host DNA. In addition, there are a number of other particles called plasmids, which can transfer genetic information from one bacterium to another. They are small, circular DNA molecules that replicate independently of the DNA of the host bacterium. Their role is particularly prominent in the transmission of antibiotic resistance. Most animal viruses are similar to the lytic cycle of bacteriophage, however, there are a number of viruses that can become latent. There is also agents like the retroviruses, whose genome gets transcribed into the host DNA, thereby changing cellular properties. Some of these agents will also incorporate host DNA into their sequences, and then capable of transmitting this sequence into a new host.

The tools that allow us to transfer genes from one organism to another are the *restriction* endonucleases. These enzymes are found naturally in bacteria and protect them from infection by breaking down foreign DNA. They are chemically very specific, only working on specific palindromic sequences. Consequently, the kind of infection discussed in the last paragraph can only occur if the bacteriophage sequence does not contain elements that will be broken down by the host bacterial system of endonucleases. A number of different enzymes are found in different bacterial species, and the enzymes referred to by an abbreviation of the bacterial species name and a number. Similarly, there are enzymes that are used to connect fragments, the ligases (Griffiths et al. 1996).

2. Cloning Genes

All of the manipulations of genetic engineering require multiple copies of the DNA sequence or gene of interest. The original methods of getting multiple copies relied on bacteriophage or plasmid vectors to introduce the foreign DNA into bacteria to produce these copies, as each modified cell produces multiple copies, and the bacterial culture itself increases. This is done by first physically isolating the vector, opening its DNA with a restriction enzyme and binding in DNA from the organism being studied that has also been cleaved with a restriction endonuclease. A new population of bacteria is then infected with the altered vector. Given an appropriate way of selecting the population of bacteria so that it uniformly has the DNA of interest multiplying within, one can isolate a large population of vector molecules with the desired sequence, which is then freed by enzymatic cleavage once again.

Fragments of DNA are identified by physically separating them by electrical charge and molecular weight through gels. The DNA of the vectors and bacteria are generally in the range of one to ten thousand base pairs, and there are a sufficiently small number so that the fragments can be identified with a simple staining technique, usually a compound that binds to DNA and fluoresces under ultraviolet light. The larger quantity of fragments that would be isolated from more complex organisms produces a smear with such dyes, so the base-pairing property of DNA, the obligate pairing of adenine with cytosine and guanine with cytosine that allows for both recognition and synthesis of the linear sequence, is used to identify the same sequence on the gel by labeling a known fragment with an isotope or fluorescent dye. The labeled molecules are called probes. This is also the basis for identifying genetic variation in organisms, either for basic studies or identification of mutations associated with disease.

Isolation of fragments produced by digestion with several enzymes, used both singly and in combination,

allows for the construction of a physical, restriction-fragment map. Smaller fragments may be replicated, followed by the chemical analysis of the base sequence within fragments which are then assembled into the final base sequence of the gene. Once the sequence is known, production of useful amounts of a region of DNA may now be done enzymatically *in vitro* with the polymerase chain reaction (PCR). In this technique, the region between two primers, one from each strand of the final DNA molecule is copied in a logarithmic fashion by a heat-resistant DNA polymerase from a small amount of genomic DNA (it has been done with single cells), using multiple heating and cooling cycles. This technique is also used in diagnostic work (Strachan and Read 1996).

3. Modifying Cells to Produce Proteins

A number of strategies have been used to produce large amounts of a peptide or protein. Multiple techniques have been necessary because of the different kinds of cells that have been used. The production of human insulin, one of the earlier products produced commercially, is a useful model that is an example of expression cloning. The sequences for the A and B chains were introduced into separate vectors in sequence with a bacterial promoter (a signal to turn a gene on) and the enzyme β -galactosidase that produces colored bacterial colonies. This arrangement is necessary because bacteria will not just produce the human protein if only the structural gene is introduced. The bacteria are tricked into producing a fusion protein: a bacterial β -galactosidase with a human insulin chain sequence at the end. After selecting the colonies by color, mass cultures are produced from which the fusion proteins can be purified. The normal insulin dimer is then produced chemically by removing the β -galactosidase, purifying the chains, combining the chains and allowing them to refold and form the sulfide bridges between the two chains.

Expression cloning in bacteria works well if the peptide or protein is composed entirely of the amino acid sequence. Many functional proteins contain specific carbohydrate groups at precise locations on the chain that are attached after translation. These modifications are different in different classes of organisms, so techniques using yeast, plant, cultured insect, other mammalian, or human cells have been used to make the complete molecule. Another problem that is addressed in this approach is the elimination of bacterial toxins that cannot be removed from the desired protein product mixture. The engineered DNA can be introduced into these cells by a variety of methods. The physical techniques like electroporation, which uses an current to move the DNA inside, or coating various kinds of particles, like gold, with the DNA and then bombarding the cells to get the DNA inside has a chemical parallel in the use of calcium salts

and heat. Artificial yeast chromosomes have been produced to move gene constructs. For the cultured insect cells, modified Bacloviruses are used (these are viruses that have invertebrate hosts). In mammalian systems, several viruses have been used, including retroviruses (the class of agents involved in acquired immunodeficiency and tumor transformation in experimental organisms), adenoviruses (agents that typically produce respiratory disease, however, some strains are capable of producing more severe systemic disease), and herpes simplex.

The goal of expression cloning is to produce large amounts of protein that can be purified more easily, and in much larger amounts than is possible using chemical techniques starting with cells from the normal tissue. On the research laboratory scale, the three-dimensional structure of the whole protein or the active region can be determined by a combination of nuclear magnetic resonance and x-ray crystallography. This information is useful for understanding function or designing small molecules that are expected to have pharmacological actions. This is then one element of the strategy in rational drug design. If the protein is to be used as a product, the growth of the modified cells has to be done on an industrial scale that is an engineering problem similar to that used in the production of a number of antibiotics (Glick and Pasternak 1998).

4. Modifying Organisms

An organism that has been engineered so that it has a gene from another species integrated into its genome is called a transgenic organism. These techniques have been used to produce plants and animals expressing mutated genes of that species to study function, with mutated or genes introduced from another species with desired agronomic properties, as models of human diseases, and to produce therapeutic products. There is a large and continuously growing literature of studies in the mouse.

Two procedures have been widely performed in mice: the interruption of a mouse gene to eliminate its function, for which the term knockout has been used, and introduction of a transgene. The 'standard' procedure employs classical, cellular, and molecular genetic techniques. The gene to be studied is first cloned, and then introduced into a targeting vector in such a way that the coding sequence of the gene is interrupted by the neomycin resistance gene. The vector also has the thymidine kinase gene from herpes. Embryonic stem cells are obtained from a brown mouse and placed in culture. The cells are then exposed to the vector. There are three possible outcomes. The vector does not insert at all, the whole construct inserts at random in a chromosome, or only the desired coding sequence and the neomycin resistance gene insert into the normal location on the chromosome.

Treating the cells with a neomycin analogue and the antiviral agent ganciclovir will kill all of the cells except the ones with the integrated sequence. The surviving stem cells are injected into the blastocyst stage embryo of a black mouse and placed in a surrogate mother. If a chimera is formed, it will have a coat of two colors. The chimeras are mated with black mice, and their brown offspring tested for the modified gene. Those that are heterozygous are mated, and one of four will be homozygous for the modification.

In addition to the introduction of genes into stem cells by vectors, injection of the desired sequence has been done by some of the physical methods discussed earlier such as electroporation, coated gold particles, and salt precipitation. Some transgenics have been derived from direct injection of the plasmid into the egg.

The production of genetically identical organisms, cloning in the original sense, was first done in amphibians by nuclear transplantation into fertilized ova from which the original nucleus was removed. This was done widely in mammals by using embryonic cells as the nuclear donor. The production of the sheep named Dolly was noteworthy because the donor nuclei were obtained from adult mammary cells that were maintained in cell culture. This precipitated a round of discussion about the feasibility and the desirability of doing this in human beings. The claim was made that this would be an alternative for infertile couples, and therefore acceptable. It is necessary to point out to people who would propagate themselves for narcissistic reasons that the derived individual would not be absolutely the same because of the developmental environment, both before and after birth (Hubbard and Wald 1997).

Another possible use of this technology that has been proposed is the therapeutic cloning of replacement tissues or organs. Cells would be biopsied from the eventual recipient and placed in culture. Nuclei from these cells would be transplanted into eggs from which the nuclei have been removed; these eggs will be allowed to develop to the blastocyst embryonic stage for the isolation of embryonic stem cells that are treated to produce the desired cell types. This avoids the problem of using fertilized eggs, but still involves the use of human eggs and early embryonic stages (Gurdon and Colman 1999).

5. Gene Therapy

The goal of gene therapy is to provide a functional protein to an individual whose genotype leads to a disease because the protein is missing or modified so that it is not functional. There are two potential modes. In germ-line therapy, the modification would be heritable, and the individual's offspring potentially would not be affected. In somatic cell therapy, only the treated individual would have the modification. There

have been many arguments against germ-line therapy including our inability to know whether potentially harmful changes in the sequence have been made inadvertently. All of the trials to date have been in somatic cells for the treatment of inherited disease and cancer by control of cell death.

Two strategies are being attempted. For those disorders for which bone marrow transplantation has been effective, the gene is introduced into marrow cells, the population of which is then expanded and infused into the patient. The techniques for gene introduction are similar to the ones previously discussed: viral vectors and chemical or physical techniques. An immune deficiency disorder, adenosine deaminase deficiency, was the first attempt reported. Although the modified cells express the enzyme, repeated infusions of treated cells have been necessary, and the patients have continued to receive chemically modified enzyme, so it is not clear whether or not gene therapy alone is sufficient.

Other trials have used viral vectors to treat target organs, such as the lung in cystic fibrosis. The therapeutic trial of an adenovirus-mediated transfer of a gene to the liver has received general media attention because of the death of a participant. Evidently, other participants in gene therapy trials have died, although it is not clear how many of those deaths are a result of the procedure (Strachan and Read 1996).

6. Applications to Behavior

The application of these techniques in behavior is loaded with scientific, ethical, and legal complications. Part of the problem stems from the very human desire to have a simple answer to a complex problem. Genetic, environmental, and developmental factors interact to produce behavior, and it is difficult to distinguish the role of each component. One should ask if it is even desirable to do this, because of the mixture of potential benefit and risk. It is clear that each behavioral trait is heterogeneous. There may be families where a single major gene is responsible, and a particular drug, expressed protein or even gene therapy could work, but in the more usual instance of several genes and environmental factors, it is unlikely that such a unitary approach would work. The promise of a solution to disease problems by definition of molecular mechanisms has been difficult to achieve. We still do not have an answer in the case of sickle cell anemia, the first disorder to be described in this manner. Common behavioral traits would be more difficult because of the interactions.

Because of the Human Genome Project, it is likely that a number of genes will be discovered that contribute to the development of a number of behavioral characteristics. The technology therefore will create the possibility of diagnostic tests for the presence of variants at these genetic loci that might be involved in diagnostic confirmation, or even asymp-

matic screening without necessarily having the ability to safely and ethically treat those in whom the genetic variant is identified. Furthermore, since it is well known that the genotype does not always predict the phenotype even for such things as emphysema associated with a variant α -1-antitrypsin, one should question this approach with respect to behavioral characteristics (Hubbard and Wald 1997).

Using genetic information to design pharmacologically active agents will probably be useful for drugs interacting with receptors influencing neural function. A similar problem arises here. Will the intervention be beneficial to someone who is suffering or will it be used to control individuals who are designated by those in authority to be deviant? Modifying the environment to enhance development is effective and poses less of a biological risk than drugs or gene therapy (Duyne et al. 1999). The relative economic costs are unknown, but the human ramifications are immense.

See also: Ethical Issues in the 'New' Genetics; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Testing and Counseling: Educational and Professional Aspects; Genetics and Development; Genetics and the Media; Genetics of Complex Traits Through the Life Cycle; Human Genome Project: German Perspective; Human Genome Project: Japanese Perspective; 'New' Eugenics, The: Role in Society

Bibliography

- Duyne M, Dumaret A-C, Tomkiewicz S 1999 How can we boost IQs of 'dull children'? A late adoption study. *Proceedings of the National Academy of Science USA* **96**: 8790-4
- Glick B R, Pasternak J J 1998 *Molecular Biotechnology: Principles and Applications of Recombinant DNA*. 2nd ed. ASM Press, Washington, DC
- Griffiths A J F, Miller J H, Suzuki D T, Lewontin R C, Gelbart W M 1996 *An Introduction to Genetic Analysis*. 6th ed. W H Freeman, New York
- Gurdon J B, Colman A 1999 The future of cloning. *Nature* **402**: 743-6
- Hubbard R, Wald E 1997 *Exploding the Gene Myth: How Genetic Information is Produced and Manipulated by Scientists, Physicians, Employers, Insurance Companies, Educators and Law Enforcers*. Beacon Press, Boston
- Strachan T, Read A P 1996 *Human Molecular Genetics*. Wiley-Liss, New York

D. J. Harris

Genetic Factors in Cognition/Intelligence

Research has not only established the importance of both genetic and environmental influences on human intelligence, but has also begun to examine how genes

shape intellectual development and the relationship among different cognitive skills (e.g., verbal and performance ability). Additionally, researchers have begun to search for DNA markers associated with human cognitive ability. This article will not only examine these new research directions, but will also discuss the implications of these findings in the context of the environment.

1. Behavioral Genetics, Genes, and Intelligence

One of the major misconceptions about behavioral genetics is that the purpose of this research is to examine whether genes *or* the environment is operating. Instead, behavioral genetics assumes that differences in intelligence, for example, are due to differences in genes, the shared family environment (the environment that family members share, such as socioeconomic status), the nonshared environment (the environment that makes family members different, such as differential parenting and different experiences), and error. Genetic influences on intelligence are statistically significant, averaging around 50 percent when collapsing across all available twin and adoption studies (see Brody 1992 for a review). In other words, genes account for roughly half of the measured variability in intelligence.

2. Beyond Heritability

While a heritability of 50 percent is important in terms of the nature/nurture debate, finding that intelligence 'is both' genetic and environmental (the other 50 percent has to be due to nongenetic influences) in and of itself does not elucidate *how* genes influence the development of human intelligence. Behavioral genetics has begun to move beyond heritability and, as a result, has reported findings that fundamentally change how we think about the development of intelligence.

2.1 Genes Become More, Not Less Important With Age

A common assumption is that genes become less important with age, as environmental influences 'add up' to shape cognitive performance. Behavioral genetic results suggest that the opposite is the case. Although genes account for 50 percent of individual difference in intelligence collapsed across all studies, heritability appears to vary with age, starting at around 40 percent in early childhood, rising to 60 percent in early adulthood, and to 80 percent in later life (see Plomin et al. 1997). More recent evidence suggests that the genetic influences on intelligence may decline to around 60 percent in old age (Finkel et al. 1998, McClearn et al. 1997).

Genes in childhood Genes in adulthood

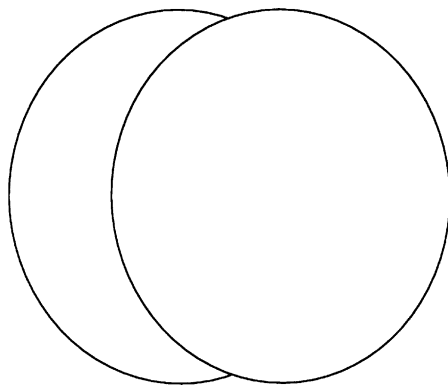


Figure 1

Genes are largely shared across development, but there is also genetic specificity across different ages

2.2 Genes Drive Stability, But Are Not Invariant Across Development

Another new research question addressed by genetic methods is the extent to which the genes that influence cognitive abilities at one age are the same genes that influence cognitive abilities at later ages. Are genes relating to intelligence immutably 'on' from birth, or do different genes turn on and off across development? Results (see Plomin et al. 1997) suggest that genes are largely stable, with some genetic change across development (see Fig. 1). In contrast, the environment is largely responsible for the changes in intelligence across development.

2.3 Genes Also Drive The Correlation Among Different Aspects of Cognitive Ability

One of the most interesting 'new directions' in behavioral genetics research involves the examination of how genes influence different aspects of cognitive ability. Are the genes that influence verbal skills the same as those that influence spatial ability or are verbal and spatial skill each influenced by different genetic processes? This debate, often called the molarity vs. modularity debate, has important implications for cognitive neuroscience. A molar system is one in which different aspects of cognitive ability, such as verbal ability, spatial ability, perceptual speed, and memory, are all influenced by a common set of genes. A modular system, in contrast, holds that genes function independently across different cognitive skills. In multivariate behavioral genetics, this degree of overlap is often measured by a genetic correlation. A high genetic correlation suggests almost complete overlap while a low genetic correlation suggests genetic independence. Behavioral genetic studies suggest that the genetic correlation approaches 1.0, suggesting

Genes for verbal ability Genes for spatial ability

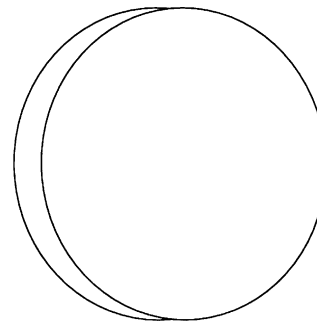


Figure 2

Genes are largely shared across different measures of cognitive ability

almost complete genetic overlap (see Petrill 1997). The genes that influence spatial ability, for example, are largely the same genes that influence verbal ability (see Fig. 2). Genes are the 'tie that binds' different cognitive skills while the discrepancy between different cognitive skills (e.g., certain people may be better at verbal as opposed to spatial skills) is due largely to the environment.

2.4 Molecular Genetics and Intelligence/Cognition

One of the most quickly developing areas of behavioral genetics involves the search for DNA markers associated with human cognitive ability. For example, the APOE-4 marker is associated with a four-fold increase in risk for developing late-onset Alzheimer's disease (Breitner et al. 1998). Another notable example is the continuing research on the association between DNA markers on Chromosome 6 and reading disability (see Gayan et al. 1999).

While these markers are important in that they show that differences in the human cognitive system are associated with specific DNA markers, these studies do not necessarily show that individual differences in healthy, normally distributed cognitive functioning are influenced by these same DNA markers. A study called the IQ-QTL Project is attempting to search for DNA markers associated with intelligence. In 1998 this study yielded the first replicated DNA marker associated with intelligence (Chorney et al. 1998) and is expected to complete a full genome scan, likely yielding other markers for intelligence.

3. Implications

Taken together, these data not only suggest that genes influence the development of intelligence, but that intelligence develops differently than once thought.

Genes become more, not less important as the age of the sample increases. Genes influence the correlation among measures of intelligence, both across time and across different measures of cognitive performance. Additionally these behavioral genetic results, although indirectly measured through twin and adoption studies, are beginning to be tied to specific DNA markers. What are the implications of these data?

3.1 DNA Markers Will Likely Be General In Effect

Although the search for DNA markers is in its infancy, it is likely that the first decade of the twenty-first century will yield a set of DNA markers that account for a significant proportion of the variance in intelligence. Given the fact that twin and adoption research suggests a great deal of genetic overlap, it should be expected that DNA markers will cut across different cognitive skills, and will be stable over time. This 'molar' focus will become increasingly important as researchers turn from simply trying to find markers associated with cognitive ability to attempting to delineate the mechanism by which genes influence complex cognitive skills.

3.2 The Environment Is Also Important

A significant heritability estimate, genetic correlation, or DNA marker associated with intelligence does not mean that intelligence cannot be modified by the environment. For one, the environment is an important determinant of individual differences in intelligence, especially in childhood (the heritability of intelligence is roughly 40 percent, which means that the other 60 percent has to be due to nongenetic influences). More importantly, although heritability describes why individuals differ within a group, it does not examine average increases in ability or average differences in ability between groups. Although genetic influences account for 90 percent of the individual differences in height, for example, the height of the average person has increased, presumably due to an average improvement in the environment. Similarly, although the heritability of IQ has remained stable across the 80+ years that twin studies have been conducted, the measures used to measure IQ have been restandardized numerous times. The average IQ, although set to be 100 on a particular test, increases over time. Put another way, if a random sample of children in 1999 were given an intelligence test standardized in 1920, the average IQ would be significantly greater than 100 (Flynn 1998). Thus, behavioral genetic methods tell us 'what is' for a specific population at a specific time, not 'what will be' or 'what should be' (Plomin and Petrill 1997).

In fact, some of the most important discoveries about the relationship between the environment and

intelligence have come through behavioral genetic research. As stated earlier, behavioral genetics separates the environment into that which is shared by family members and that which is not shared by family members. Thus, behavioral genetic methods allow the examination of environmental effects that operate between families (shared environment), and environmental effects that operate within families (nonshared environment). The same behavioral genetic studies that suggest increasing heritability and high genetic correlations also suggest that the most pervasive environmental influences across the lifespan are non-shared. In other words, the most important environmental effects are those that make family members different from one another. Behavioral genetic studies are currently attempting to identify these nonshared environmental influences.

4. Summary and Conclusions

Genetic influences *are* important to understanding intelligence. However, the environment is also important. Intelligence tests are not a veridical window into genetic potential nor are they indices of environmentally driven learning. Intelligence is defined by measures of behavior and is thus subject to both genetic and environmental effects. The key issue is understanding how genes and environment work together to yield cognitive development.

Bibliography

- Breitner J C S, Jarvik G P, Plassman B L, Saunders A M, Welsh K 1998 A risk of Alzheimer disease with the epsilon 4 allele for apolipoprotein E in a population-based study of men aged 62–73 years. *Alzheimer Disease and Associated Disorders* **12**: 40–4
- Brody N 1992 *Intelligence*, 2nd edn. Academic Press, San Diego
- Chorney M J, Chorney K, Seese N, Owen M J, Daniels J, McGuffin P, Thompson L A, Determan D K, Benbow C, Lubinski D, Eley T, Plomin R 1998 A quantitative trait locus associated with cognitive ability in children. *Psychological Science* **9**: 159–66
- Finkel D, Pedersen N L, Plomin R, McClearn G E 1998 Longitudinal and cross-sectional twin data on cognitive abilities in adulthood: The Swedish adoption/twin study of aging. *Developmental Psychology* **34**: 1400–13
- Flynn J R 1998. IQ gains over time: Toward finding the causes. In: Neisser U (ed.) *The Rising Curve*. American Psychological Association, Washington, DC
- Gayan J, Smith S D, Cherny S S, Cardon L R, Fulker D W, Brower A M, Olson R K, Pennington B F, DeFries J C 1999 Quantitative-trait locus for specific language and reading deficits on chromosome 6p. *American Journal of Human Genetics* **64**: 157–64
- McClearn G E, Johansson B, Berg S, Pedersen N L, Ahern F, Petrill S A, Plomin R 1997 Substantial genetic influence on cognitive abilities in twins 80 or more years old. *Science* **276**: 1560–3

- Petrill S A 1997 Molarity versus modularity of cognitive functioning? A behavioral genetic perspective. *Current Directions in Psychological Science* 6: 96–9
- Plomin R, DeFries J C, McClearn G E, Rutter M 1997 *Behavioral Genetics*, 3rd edn. W. H. Freeman, New York
- Plomin R, Petrill S A 1997 Genetics and intelligence: What's new? *Intelligence* 24: 53–78

S. A. Petrill

Genetic Screening for Disease-related Characteristics

1. Introduction

Although one could imagine genetic screening for human characteristics that are not disease-related, in what follows we will restrict ourselves to disease-related screening. Furthermore, although disease-related screening could serve other purposes than the benefit of the participants, e.g., scientific or economic motives, we consider here only screening that has as its main goal the interest of people. So we place genetic screening in the medical perspective, where it is a component of community genetics, the application of medical genetics to the benefit of large numbers of people.

2. Definition

Genetic screening is the application of a test on people for the systematic early detection or exclusion of a hereditary disease, a genetic predisposition to a disease, or to determine whether a person carries a predisposition which may produce a hereditary disease in their offspring (Health Council of the Netherlands 1994). Testing is not restricted to the study of DNA or chromosomes, but includes the use of any test which is designed to detect and exclude hereditary disease, predisposition, or carrier state. Early detection means that the testing is applied to those who have not (yet) sought medical aid because of physical signs, symptoms, or anxiety. So it is typically the care system that takes the initiative. The word 'systematic' in the definition implies that screening is offered to all people in a predefined target group, e.g., neonates or pregnant women.

2.1 Terminology

Genetic screening shares its terminology with epidemiology and genetics. A test result is called positive when it points to the presence of the feature sought

after. Similarly, a result is said to be negative when it indicates the absence of the feature. Both results, however, may be true or false. A test result, for instance, is called false positive if it indicates that the feature looked for is present while in reality it is absent. The probability that a certain test result is true or false depends on the prevalence of the feature and two test characteristics: sensitivity and specificity. Sensitivity refers to the proportion of cases in which the presence of the feature is identified by the test. Specificity is the proportion of cases in which the absence of the feature is correctly identified by the test. The predictive value of a test result, either positive or negative, indicates how often that test result is correct.

As to the terminology of genetics the reader is referred to other articles. An exception is made for the term 'penetrance.' Penetrance is the frequency of cases in which a particular genotype comes to expression.

3. Characteristics of Genetic Screening

As with other types of medical screening, diagnostic confirmation of a positive screening test is an integral part of the screening procedure. Unlike other types of screening, however, genetic screening looks for stable characteristics of the participants. So in general, once an individual is tested, repetition of the test is not needed. In screening for infectious diseases or cancer, the majority of participants, namely the true negatives, the false positives, and the cured, have to return for subsequent testing after a set time period.

The stability of the genetic characteristics also makes it possible to predict disease long before it would become manifest without intervention. In principle this lead time (time until disease becomes manifest) can be lifelong, or with prenatal diagnosis even longer. It is clear that this is an unwanted situation unless interventive measures are more effective when they are started early.

A further characteristic of genetic screening that differentiates it from other types of medical screening is that a positive test result may frequently have consequences for the family members. This does not apply to all types of genetic screening. For instance, finding a fetus with a regular trisomy 21 Down syndrome does not increase the risk for sisters of the pregnant woman, since this is a spontaneous mutation. But finding through screening that someone carries, for example, two mutations of the gene involved in hemochromatosis puts his or her siblings immediately at a 25 percent risk of having the same mutations (Worwood 2000). They may be grateful to know this since manifestations of this iron overloading disease can be prevented effectively by regular bleeding, but the point here is that one's decision to participate in a genetic screening program is to some extent a decision involving family members who might have preferred to make such choices themselves.

4. Target Groups

The target groups for genetic screening may consist of individuals or couples. Screening newborns for phenylketonuria (PKU) or young adults for hemochromatosis are examples where the individual is the target. In both cases the goal of identifying these persons is to offer timely intervention (dietary measures prevent the development of mental retardation in PKU, and bleeding prevents the iron overloading in hemochromatosis). Other examples are screening for other inborn errors of metabolism, congenital hypothyroidism (CHT), and familial hypercholesterolemia.

Screening for carrier status of autosomal recessive diseases such as cystic fibrosis, thalassemia, and Tay-Sachs disease can be offered to individuals, but because being a carrier is only significant when the partner is a carrier too, screening for carrier status can also wait until a partner has been found and reproductive issues arise. In this case the couple rather than the individual is the target. When both partners of a couple are carriers (for the same disease) their children are at 25 percent risk for the disorder.

Screening during pregnancy may be targeted at characteristics of one to three individuals. First, the fetus may be studied directly, for instance when ultrasound is used for the detection of congenital anomalies in general or nuchal translucency (an indication for the existence of Down's syndrome) in particular. Second, the screening may involve the pregnant woman first, such as is the case with maternal serum screening, e.g., triple test. In case of a positive test result the fetus becomes involved in a definitive diagnosis. Third, both the pregnant woman, her partner, and the fetus may become involved. This is the case with couple screening in pregnancy.

4.1 Timing

As hinted at above, apart from determining who should be screened, it is important to define the best moment for screening within the lifetime of the individual or couple. The earliest possibility is pre-implantation genetic screening (PGS) in case of *in vitro* fertilization (IVF). With pre-implantation genetic diagnosis (PGD), one is looking for a disorder for which this particular embryo has a particular high risk, while in PGS the embryo at stake has a not much higher risk than other pre-implantation or IVF embryos. PGS for chromosomal anomalies, for instance, would be very helpful to select the embryos with the highest chance of survival.

Prenatal screening, newborn screening, and adult screening are discussed above. Some people advocate screening in high school. Couple screening can be performed premaritally, preconceptionally, and prenatally. Premarital screening is perhaps the best choice in cultures where selection of marriage partners is

done by third parties, without involving the partners at all. Compulsory premarital screening, after the partners have matched, as is the case for thalassemia in Cyprus, infringes personal autonomy. Preconceptional screening is characterized by difficulties of reaching the target group in time. So the majority of couple-screening programs target pregnant women (and their partners).

4.2 Taking Advantage of Family Structure

Cascade screening is a special type of genetic screening. Starting from an individual with the characteristic of interest (the proband), one offers screening to his or her first-degree relatives. Once the relatives with the same characteristic have been identified, their first-degree relatives are also offered screening, and so on. This type of screening is suitable for dominant disorders, such as familial hypercholesterolemia, and for X-linked disorders with high penetrance and low mutation rate. It has, however, also been proposed for the detection of couples who are carriers of an autosomal recessive disease such as cystic fibrosis. The disadvantage there is that only a small proportion of at-risk couples can be identified in this way. So, screening without taking the family structure into account is a better option. A combination of non-cascade and cascade screening may, however, be very efficient.

4.3 Other Technicalities

In many cases it is not possible or feasible to have a test which is 100 percent sensitive or specific. The lack of perfect specificity can be countered by subsequent diagnostic workup. When the screening involves a DNA test, specificity is usually very high. If in the cascade approach family members are tested for the same mutation as has been found in the proband, sensitivity is also very high. In case of a test with imperfect sensitivity (e.g., 90 percent), the imperfection is multiplied in couple screening (81 percent).

Couple screening has many modalities (ten Kate et al. 1996). One may test one partner first, and only if this first partner tests positive, test the other partner (single entry or step-wise screening), or one may test both partners always (double entry). One may also decide to screen partners of carriers for more mutations than is feasible in the initial screen. One may sample both partners simultaneously or stepwise. One may give complete results, or only tell the couple whether they are both carriers or not.

5. Psychosocial Impact of Genetic Screening

The purpose of genetic screening is to provide people with information about their risks and the options available to them to avoid the birth of an affected child or to take preventive measures in case of a serious

disease. Genetic screening seems especially advantageous when intervention measures are available. For instance, for women with a high risk of hereditary breast cancer, confirmation that they are carriers makes it possible for them to take preventive measures such as participating in an intensive screening program for early detection of the disease. Because the potential impact of genetic screening is considerable, the advantages as well as the disadvantages of genetic screening should be considered carefully before deciding to offer a specific screening program routinely.

The availability of genetic screening may also have its drawbacks. Although the available choices would appear to increase people's control over their own destiny, they are actually experienced by many as an additional burden. An indication of the perceived difficulty of decisions is the finding that although most people at risk for a genetic disease say they intended to be tested in the future, many are not (e.g., Lerman et al. 1996). Some people might prefer the uncertainty of not knowing above the certainty that, for instance, they will develop a serious disease for which there is no cure, such as Huntington's disease. In addition, with the increasing possibilities and availability of genetic screening, more and more people, some with a very low risk for chromosomal abnormalities who first did not consider screening, are now confronted with the choice to undergo screening. In many countries young pregnant women with a low risk of having a child with Down's syndrome are offered prenatal screening. If they choose screening, they must cope with the possibility of a positive screening result and the implication that confirmation is needed with invasive diagnostics such as amniocentesis that itself carries a risk of miscarriage. When they decide against screening, they must deal with the feelings that they could have done otherwise in case things turn out badly and the child is affected.

The decision for or against screening is not only complicated emotionally but also cognitively. Genetic risk information is complex and replete with uncertainties, associated not only with the hereditary nature of the disease but also with (un)informativeness of the test results, the effectiveness of preventive measures, and the variability of expression of the disease (incomplete penetrance). For example, when a woman is diagnosed as 'at risk' for breast cancer based on family history, there is only a 30 percent chance that DNA testing will result in a conclusive test result. If a woman is a carrier of the breast cancer gene, she has an increased risk of getting breast cancer: 7 out of 10 women who are carriers will get breast cancer before the age of 75, and 3 of them even before 40. There is a 3 in 10 chance, however, that a woman who is a carrier will not get breast cancer. This is complicated information and difficult to integrate to make the decision about testing (e.g., Bottorff et al. 1998). Providing information about genetic risks of, e.g., breast cancer is considered an important component

of genetic education and genetic counseling. The importance attributed to risk information is illustrated by the fact that recall of risk information and understanding of probabilities are considered central measures of the effectiveness of genetic counseling. The underlying assumption that accurate perception of risk could help people to make informed and individual decisions seems to be based on the principle of the autonomous and rational decision maker. From this perspective the most important goal of genetic education and counseling is to provide people with all relevant information in an understandable way. It is therefore important to find the best format to communicate risks and to stimulate a rational tradeoff of pros and cons of several options. Research has shown, however, that people find it difficult to understand risks and to make rational tradeoffs. It has been shown that knowledge of actual population risks of, e.g., breast cancer is far from adequate. People either underestimate or overestimate the risk and often fail to make a distinction between the population risk and their own risk. Genetic counseling has been shown to improve accurate risk perception, although the majority of women still have an inaccurate perception after counseling (e.g., Cull et al. 1999).

In the case of screening for a hereditary disease, the decision to screen is complicated by the potential impact that test results could have on family members. As has been pointed out, genetic screening may have implications for family members and may put familial relationships under pressure. For example, one sister may want to know more about the hereditariness of breast cancer while another sister might prefer not to know. The same goes for couples when there is no agreement as to whether to test for carrier status for, e.g., cystic fibrosis.

If the decision is made and test results become available, these need to be interpreted. Mostly, test results also involve uncertainties. A positive test result after prenatal screening for Down's syndrome means that the risk of having a child with Down's syndrome is higher than 1 in 250. A decision then must be made whether or not to confirm this chance estimate with invasive diagnostics whereby the extra information must be traded off against the risk of a miscarriage. Screening and waiting for the results have an emotional impact on people. Receiving test results, whether they are positive or negative, is also emotionally stressful. Many low-risk people are likely to undergo testing for reassurance and hope for a good result. Most of them will receive a negative test result. For those who perceive themselves as being at risk this is reassuring, but the news will perhaps be learned at the expense of knowing that they were at risk in the first place. It has been shown that anxiety exists even after negative results of prenatal screening (Green and Statham 1996). Learning what could go wrong without the possibility of ruling out all possible defects makes women more aware of birth defects and affects their

sense of vulnerability. Testing for carrier status of cystic fibrosis involves a large number of mutations but it is practicable to test only for the most common. If tested negative, the carrier status of cystic fibrosis, therefore, cannot be ruled out completely and no complete reassurance can be given that one is not a carrier. Even if testing implies that the person is not at risk, some members of risk families for, e.g., Huntington's disease find it difficult to cope with a future that no longer includes the expectation of getting a serious disease. An explanation for these negative feelings after being tested negative is that being at risk for this serious disease was part of one's identity which then needs to be changed (Tibben et al. 1997).

A positive screening test is obviously upsetting even if it turns out to be false. Women having diagnostic tests as a result of a positive screening test have been shown to be especially anxious compared to women who chose amniocentesis directly. Most of these women will be shown not to carry a child with Down's syndrome, and their anxiety can be considered to be unnecessary. Testing positive as a carrier of a recessive disease such as cystic fibrosis without actually having the disease can affect one's view of and one's optimism about the future. Being identified as a carrier for a genetic disease may affect how an individual perceives him- or herself and how he or she is perceived by others. Being tested positive for a disease for which there is no cure (such as Huntington's disease) also affects one's view on the future. Because nothing can be done to prevent the disease, this knowledge will not be considered as useful by all people (Keenen 1996). As more genetic conditions become diagnosed before any cure is in sight, more and more individuals will be added to the at-risk pool many years before a cure is possible. Some people may even be at risk their whole life for a condition they will never develop. It is therefore important to determine for which conditions and for whom genetic testing should be available.

It could be argued that at the basis of many of the problems or possible negative effects of genetic screening described above is a poor understanding of genetics, the tests, and the meaning of test results. It is therefore very important that the public as well as people with an at-risk status learn more about genetics, what it means to have some abnormal genes, how best to make a decision, what it means to have a positive screening result, etc. Professionals should learn the best way to provide genetic information in such a way that people can make a balanced decision for or against genetic screening and be able to cope with the test results so that the advantages of screening will be maximized and the disadvantages minimized (Schwartz et al. 1999).

6. Criteria

There are many genetic, medically relevant characteristics for which, in theory, screening could be started

now. There will be many, many more in the future. It is quite clear that it is not desirable to screen for all these characteristics, and careful selection is needed. Apart from potential benefit, potential harm should be considered. Harm could be physical, psychological, social, or otherwise. Not only the effects of screening on the participants should be taken into account, but also the effects on nonparticipants.

Many individuals and groups have worked on the development of criteria that should be fulfilled for screening in general (EUROGAPPP PROJECT 1999–2000, Holtzman and Watson 1997, Nuffield Council on Bioethics 1993). Some countries have also developed specific legislation. It is widely recognized that every individual screening program should be tested to see if it reaches appropriate standards.

Basically, the information that is needed for this test can be grouped into three sets. The first set concerns what is known about the disease or characteristic and about the screening program (state of knowledge). The second set concerns the advantages and disadvantages arising from the screening, and the third set is about the guarantee of autonomous decision.

The disease should be well delineated. Its general course should be clear but also the amount of variability in its course. The effects of intervention should be known. A reliable test should be available with known sensitivity, specificity, and predictive value. The screening program should be described in detail, including the availability of intervention and follow-up. Privacy protection, and permanent quality control of all aspects of the program should be ensured.

Practical courses of action should be available for the participants, such as preventive measures or reproductive choices. The psychological and social effects should be known for the participants and nonparticipants in general, but also and in particular for those among the participants who are faced with a positive test and those among the nonparticipants who are confronted with the disease in themselves or in a child. Also, more remote effects on society should be considered, e.g., on the amount of support for handicapped people. Finally, the financial costs of the program should be taken into account and alternative strategies to combat the health problem should be considered. All this information should make it possible to compare and weigh the benefits and drawbacks of the program. The final balance should be clearly biased towards the benefits.

Participation in the screening program should be voluntary and based on good-quality information and informed consent. People should be given enough time to reach a balanced decision. This requirement may be difficult to realize in prenatal and neonatal screening, but every effort should be made to inform pregnant women about the decision to come as early in pregnancy as possible. People should have ample opportunities for pre-test consultation. People should

not be taken by surprise (as in opportunistic screening).

It is desirable to have an authority to supervise all (genetic) screening programs in a country or larger region. Criteria should be the same for screening programs offered by nonprofit organizations and commercial firms. As more and more tests become available, problems inherent in multiple testing arise. Multiple testing increases the frequency of false negative and false positive results and augment the difficulty of giving sufficient information to reach a balanced decision.

The complexity of genetic screening makes it a multidisciplinary area, involving medical doctors, laboratory specialists, epidemiologists, psychologists, sociologists, jurists, ethicists, patient organizations, and representatives of the community.

See also: Alcoholism: Genetic Aspects; Cancer Screening; Community Genetics; Dyslexia, Genetics of; Eating Disorders, Determinants of: Genetic Aspects; Ethical Issues in the 'New' Genetics; Familial Studies: Genetic Inferences; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Engineering; Genetic Testing and Counseling: Educational and Professional Aspects; Mental Illness, Genetics of; 'New' Eugenics, The: Role in Society; Obesity, Determinants of: Genetic Aspects; Risk Screening, Testing, and Diagnosis: Ethical Aspects; Schizophrenia and Bipolar Disorder: Genetic Aspects

Bibliography

- Aymc S, Bobrow M, Daviello D 2000 Population genetic screening programmes: Proposed recommendations of the European Society of Human Genetics. *European Journal of Human Genetics* **8**: 998–1000
- Bottorff J L, Ratner P A, Johnson J L, Lovato C Y, Joab S A 1998 Communicating cancer risk information: The challenges of uncertainty. *Patient Education Counselling* **33**: 67–81
- Cull A, Anderson E D C, Campbell S, Mackay J, Smyth E, Steel M 1999 The impact of genetic counselling about breast cancer risk on women's risk perceptions and levels of distress. *British Journal of Cancer* **79**: 501–8
- Green J, Statham H 1996 Psychosocial aspects of prenatal screening and diagnosis. In: Marteau T, Richards M (eds.) *The Troubled Helix: Social and Psychological Implications of the New Human Genetics*. Cambridge University Press, Cambridge, UK, pp. 140–64
- Health Council of the Netherlands: Committee Genetic Screening 1994 *Genetic Screening*. Health Council, The Hague, publication no. 1994/22¹⁵
- Holtzman N A, Watson M S (eds.) 1998 *Promoting Safe and Effective Genetic Testing in the United States*. Final report of the task force on genetic testing. Johns Hopkins University Press, Baltimore, MD
- Keenen R H 1996 The at-risk health status and technology: a diagnostic invitation and the 'gift' of knowing. *Social Science & Medicine* **42**: 1545–53
- Lerman C, Narod S, Schulman K, Hughes C, Gomez-Caminero A, Bonney G, Gold K, Trock B, Main D, Lynch J, Fulmore C,

Snyder C, Lemon S J, Conway T, Tonin P, Lenoir G, Lynch H 1996 BRCA1 testing in families with hereditary breast-ovarian cancer. A prospective study of patient decision making and outcomes. *Journal of the American Medical Association* **275**: 1885–92

Nuffield Council on Bioethics 1993 *Genetic Screening. Ethical Issues*. Nuffield Council on Bioethics, London

Schwartz L M, Woloshin S, Welch H G 1999 Risk communication in clinical practice: Putting cancer in context. *Journal of the National Cancer Institute of Monographs* **25**: 124–33

ten Kate L P, Verheij J B G M, Wildhagen M F, Hildering H B M, Kooij L, Verzijl J G, Habbema J D F 1996 Comparison of single entry and double entry two-step couple screening for cystic fibrosis carriers. *Human Heredity* **46**: 20–25

Tibben A, Timman R, Bannink E C, Duivenvoorden H J 1997 Three-year follow up after presymptomatic testing for Huntington's disease in tested individuals and partners. *Health Psychology* **16**: 20–35

Worwood M 2000 Early detection of genetic hemochromatosis: should all young adults be offered the genetic test? *Genetic Testing* **4**: 219–28

L. P. ten Kate and D. R. M. Timmermans

Genetic Studies of Behavior: Methodology

The phrase 'methodology in genetic studies of behavior' is intended to cover any scientific analysis of behavioral phenomena that includes biometric genetic information. Human beings have multiple thousands of genes—specific DNA sequences on specific areas of the chromosomes—and each gene represents a specific sequence of amino acids that creates RNA for the formation or coding of specific proteins. Different versions of a gene appear as a pair of *alleles* at a specific position or *locus*, and a specific collection of alleles are labeled a *genotype*. While over 99 percent of all genes in humans are identical to the corresponding genes in all other humans, the other 1 percent are important because they are thought to lead to individual differences in both genotypes. Human beings exhibit multiple thousands of observed behaviors and any well-defined subset of these behaviors is labeled a *phenotype*. In these terms, the methods of genetic analysis of behaviors are used to clarify and improve our understanding of the relationships among particular variations or combinations of genotypes and particular variations or combinations of phenotypes.

Genetic methods and analyses have roots in the nineteenth century, specifically in Charles Darwin's well-known ideas on evolution and Gregor Mendel's work on the mechanisms of heredity in plants. Many techniques used in genetic analyses of behavior were initiated in the agricultural sciences to better control the size and health of plants and animals, further developed in the medical sciences to understand and control sources of disease, and are now routinely used in large-scale family studies to better understand many

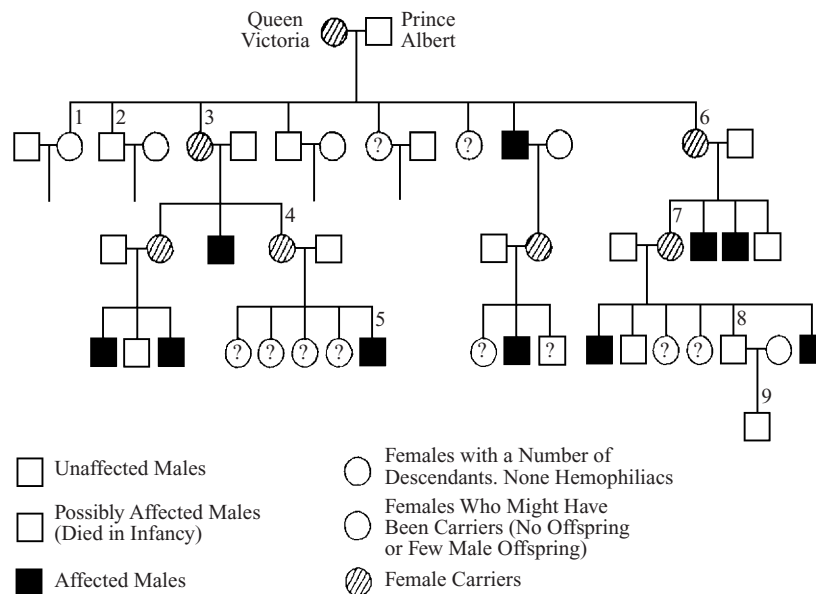


Figure 1

A pedigree chart for the inheritance of hemophilia in the family of the English Queen Victoria, Prince Albert, and their descendants (from Dixon and Johnson 1980, p. 24)

behavioral phenomena. Most of the latter methods described here stem from the classical nineteenth century work of Sir Francis Galton on the measurement of individual differences in physical and mental characteristics (Johnson et al. 1985).

In this article, we discuss some historical background, we focus on contemporary methodological issues, and we emphasize the methodological unity of most current genetic approaches. We do not deal with important contemporary issues about the use of genetic information (i.e., genetic testing and screening, ethical use of genetic information, human cloning, genetically altered food, etc.).

1. Classical Methods in Genetics and Behavioral Studies

From a historical point of view, human genotypes have been very difficult to measure and have often been treated as unobserved or latent attributes. Analyses of family trees or 'pedigrees' were among the first scientific evidence provided to clearly establish that aspects of human behavior are influenced by genetic variation. Figure 1 is a chart of a family tree or 'pedigree' from Galton of the inheritance of hemophilia in the family of the English Queen Victoria, Prince Albert, and their descendants. The pedigree method has been used to establish whether given disorders (e.g., hemophilia) run in families and, if so, what kind of inheritance (dominant or recessive, autosomal or sex-linked) best fits the observed pattern,

for example hemophilia was found to be a sex-linked characteristic carried on the X chromosome. In another well-known example, clinical research on mental retardation, especially research showing the effects of inbreeding and the genetic cause of Down syndrome, has established the role for genetic regulation of impaired intellectual functioning (for references see Dixon and Johnson 1980, McArdle and Prescott 1996).

Findings from clinical disorders led researchers to investigate whether genetic factors might regulate aspects of behavior within the normal range of variation. Early evidence for genetic effects on normal

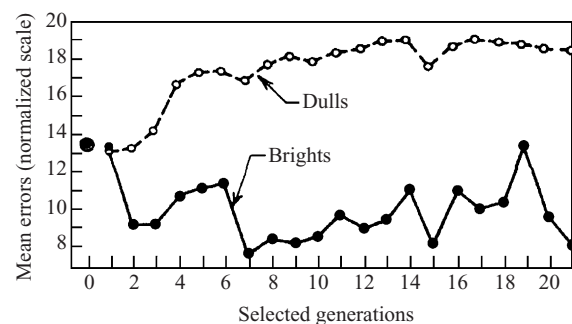


Figure 2

The results from Robert Tyron's (1940) selective breeding for maze learning over 20 generations of rats (from McGill 1965, p. 75)

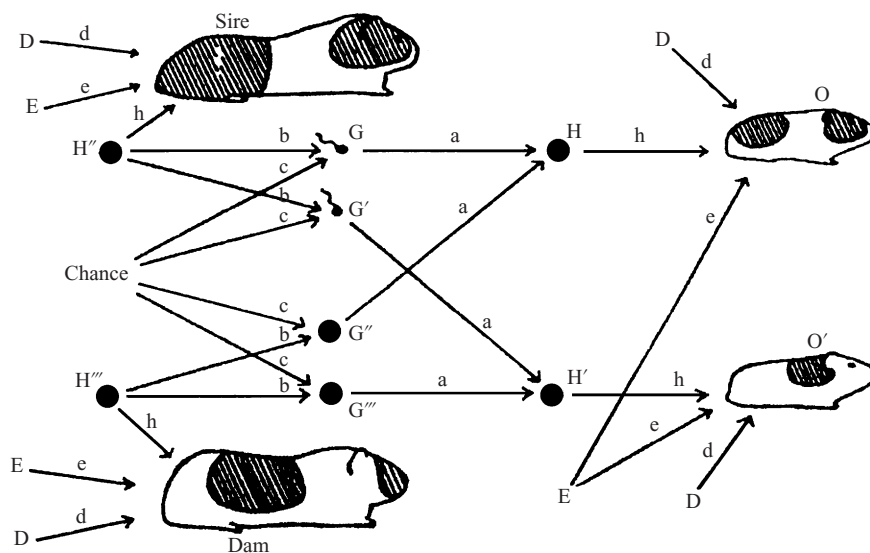


Figure 3

A path analysis model of inheritance of coat-pattern in guinea-pigs (from Sewall Wright 1920)

behaviors came from selective-breeding studies with non-human animals. Figure 2 is a plot of results from the well-known selective-breeding studies of Tolman, Jeffress, and Tryon (Tryon 1940 as reported by McGill 1965). In these studies a heterogeneous stock of rats was tested on learning tasks in an automatic maze. The observed performance in these learning tasks led the experimenters to classify each rat into one of two groups—'dull' or 'bright'—within which the rats were allowed to mate and produce offspring for the next generation. After accounting for some deleterious effects (e.g., inbreeding and regression), these investigators showed that after successive selection of relatively few generations this behavioral selection led to distinct strains of 'maze-bright' and 'maze-dull' rats. With no change in rearing environments, these experiments provided direct evidence for the role of genetic components in learning.

Galton also initiated large-scale studies based on family resemblance. He studied physical features (e.g., height) in parents and children and reported the law of 'filial regression (reversion) towards mediocrity' better known as 'regression to the population mean.' He also studied intellectual achievement and reported that individuals born to higher social classes attained higher educational and social standing than individuals born to lower classes. These family studies were not controlled experiments, so scientific inferences about the separate impacts of changing genotypes and changing environments remained obscure. Selective breeding experiments were impossible with human populations for ethical reasons. Nevertheless, experimental concepts derived from animal research gradually led to creative use of a novel class of methods

based on correlations observed among atypical groups of families—'twin studies,' 'adoption,' 'half-sibs,' and other 'genetically informative' family groupings were discussed in the early twentieth century by many researchers including Thorndike, Siemans, Merriman, and Fisher (see Spector et al. 2000).

Multivariate statistical methods were developed to accommodate unobserved or latent genotypes. Figure 3 is a path diagram from the classical work of geneticist Wright (1920) showing a complex system of influences. Here Wright suggested the environmental factors (*E*) were mixed with the heredity influences (*H*) together with 'Chance' factors (and those labeled *D*), and all were required to account for the 'coat-pattern' of a guinea-pig litter. More complex models of inheritance were defined by the quantitative rules of Mendelian inheritance and the factors of the correlation decomposition of the statistician R. A. Fisher and later by many other including R. B. Cattell, and J. Jinks and D. Fulker (see Falconer and Mackay 1996).

Advanced versions of these genetic methods are now widely used in behavioral research. The search for specific genetic abnormalities linked to behavioral disorders continues at a rapid pace in research from the Human Genome Project (e.g., Strachan and Read 1995). Genotypes are now directly related to behaviors using experimental techniques based on direct manipulation of the genome in model organisms, as well as statistical techniques of linkage and association (Elston 1998). In advanced statistical studies, substantial progress has been made in the use of mathematical and statistical techniques for studies of family resemblance based on latent genotypes and latent phenotypes (McArdle and Goldsmith 1990, Neale and

Cardon 1992). In the next four sections we highlight contemporary issues in these methodologies and present some issues for future research.

2. Contemporary Experimental Methods with Measured Genes

The last decade of the twentieth century has seen a revolution in molecular genetic methodology. Techniques based on recombinant DNA methods permit direct measurement and experimental manipulation of variations in genotype or gene expression. It is useful to divide these techniques on the basis of whether they primarily use information about a gene's position within the genome, a gene's function, a gene's DNA sequence, or a gene's protein product to develop our understanding about the genetic influences on behavioral characteristics.

Studies based on positional information are usually aimed at the long-term goal of positional cloning of the gene or genes that influence some trait (Burmeister 1999, Schork and Chakravarti 1996). In positional cloning, one capitalizes on the fact that Mendel's law of independent assortment does not hold for genes that are physically near each other on the same chromosome, so genes do not assort randomly. Rather they tend to co-segregate within families to a degree that is monotonically and directly related to their physical proximity to each other on a chromosome. A genetic marker is a specific physical location on a strand of DNA. If co-segregation is observed between a genetic marker and a phenotype, it is reasonable to conclude that the marker is physically close or linked to a gene influencing the phenotype. When this co-segregation is examined as an association between markers and phenotypes within families one is generally testing for linkage. When this co-segregation is examined as an association between markers and phenotypes across nominally unrelated individuals one is often testing for linkage disequilibrium (Elston 1998).

There are many types of studies that use positional information to find genes influencing traits in non-human animals (e.g., Lynch and Walsh 1998). Selective breeding studies with measured genes are the conceptually simplest method used to locate genes influencing traits. These studies are based on measurements of the frequencies of alleles at genetic loci in an outbred population. Any loci exhibiting statistically significant changes in allele frequency over generations are either locus influencing the trait or linked to loci influencing the trait.

A more complex but efficient method uses recombinant inbred strains. Once several inbred strains are produced, one animal from each line can be genotyped at a set of marker loci. Because all animals within a strain have the same genotype, it is only necessary to genotype one animal within each strain to

know all animals' genotypes. At each locus of interest, one simply tests whether the different lines (i.e., AA vs. BB genotypes) have different phenotypic outcomes. If significant differences are found, this is evidence that the locus under study either causes variation in the phenotype or is linked to a locus that causes variation in the phenotype. In certain situations, the use of inbred strains can be quite efficient in terms of both genotyping and phenotyping costs. More complex methods involve intercrosses between inbred strains, multipoint analyses with simultaneous uses of multiple markers, multivariate analysis of multiple phenotypes, selective sampling, and other enhancements.

One limitation of all approaches based on positional information is that they only work with existing variation within a gene, that is, they do not tell us what genes are fundamentally important to a phenotype when all individuals within a population have the same alleles at that gene. In one kind of experimental approach, molecular geneticists are now able to alter or remove particular genes and observe the behavioral effects. One of the most common types of induced mutation studies is the development of a 'knockout' animal where a specific targeted gene is disrupted or 'knocked-out.' Another type of study is based on the use of random insertion clones—expressed sequences tag (EST) acting as a molecular homing device—to disrupt genes throughout the genome. Another type of direct genetic manipulation is referred to as the generation of a 'transgenic' model where additional genetic material or a new gene is inserted into an animal. A final approach to studying gene function uses microarrays, small silicon chips, which have been set up to simultaneously measure the amount of expression of literally thousands of genes. This approach is permit study the circumstances under which certain genes come in to play and, in future research, may permit analyses of multiple interactions of measured genes with measured environments.

3. Contemporary Statistical Methods Based on Measured Genotypes in Humans

In human studies, one classical way to find genes that influence behavioral traits is to look for cytogenetic anomalies among rare individuals with respect to those traits (see Dixon and Johnson 1980). If a cytogenetic anomaly affects a large segment of the chromosome (i.e., missing, transposed, or present in extra copies), and it can be detected in the individuals, then it is plausible that the genomic region of the anomaly harbors a gene influencing the trait. While this method may provide a good starting point for further study, most behaviors are not marked by discontinuity, are not related to abnormality, and cannot be examined in this simple way.

In the method known as linkage analysis, associations (e.g., correlations) between transmission of a

measured allele and some measured phenotype are studied within families (Elston 1998). The variety of methods for testing for linkage in human studies is now quite large, but the prototypical study is the sib-pair study. Two classes of tests are commonly applied to quantitative genetic information. Neither relies on the assumption of a particular genetic model (Elston 1998) and both are based on allele sharing at marker loci (Lander and Schork 1994). Covariance-based tests are based on standard statistical principles and are commonly used (Allison et al. 2001). Under the null hypothesis of no linkage between a marker and a quantitative trait locus (QTL), there will be no association between the number of alleles a sib-pair shares identical by descent (IBD) and their degree of phenotypic similarity for the traits under study. In contrast, when there is linkage between the marker and the QTL, there will be a positive association between (a) the number of alleles a pair shares IBD and (b) the degree of phenotypic similarity of the pair. Thus, by evaluating whether there is evidence of positive covariance between IBD and phenotypic similarity one can test for linkage. Another class of tests is based on inferring such association by observing a departure of the 'marginal' distribution of IBD from its null expectation in samples of sib-pairs that have been selected on the basis of their intra-pair phenotypic similarity or dissimilarity.

One of the key issues in both human and non-human genetics, particularly linkage analysis with large numbers of candidate genes, is the determination of an appropriate threshold for statistical significance given the enormous amount of multiple testing conducted (Lander and Schork 1994). Although no definitive consensus now exists, probability or p -values greater than about 10^{-4} are rarely taken as compelling evidence for a linkage effect. Reports of p -values smaller than 10^{-4} are common and are often persuasive. However, as in any statistical model, these p -values are calculated under various untested assumptions (e.g., distributional assumptions, accuracy of genotyping, accuracy of phenotyping, independence of observations) and the test procedures may lead to empirical Type I error rates which exceed (or fall below) nominal alpha levels (e.g., Allison et al. 1999). It follows that small p -values alone should not be taken as convincing evidence for a linkage finding in the absence of additional evidence. Researchers look for careful documentation of the statistical robustness of the finding (e.g., bootstrap and sensitivity analyses), replication in an independent data set, or confirmation through independent methods such as gene knockouts in animal models.

Perhaps the chief limitation of linkage analysis in the context of human studies is low statistical power. For example, to detect a QTL explaining 5 percent of the variance in a phenotype would usually require that one study, or at least screen, thousands of sibling pairs. The reasons for this low power are manifold, but

chief among them is the fact that in linkage analysis in humans, one is generally not examining the direct effect (in terms of mean displacement) of different alleles at a QTL, but rather the variance explained by the QTL (i.e., a variance component). In addition, this is only an indirect estimate based on the observation of IBD scores at a marker allele, and these alleles are only partially correlated with IBD scores at the QTL. This being the case, initial analyses focus on a simple test of mean differences in phenotype by genotype at a putative QTL; such a test is referred to as an 'association test' in the human genetics literature. Under most circumstances, association tests will be enormously more powerful than linkage tests when one has the polymorphisms of the actual QTL in hand. The typically greater power of association tests has led some to suggest that linkage analyses be replaced by association tests, but this position has led to controversy (e.g., Risch and Merikangas 1996).

One special concern comes from the need to control for background correlations or other confounding, usually referred to as admixture or population stratification in the context of human genetic studies. These confounds arise from the fact that the effects of alleles at any one locus may be stochastically dependent upon the effects of alleles present at another locus even if these two loci are not physically linked on the same chromosome. To avoid the inappropriate conclusions that could stem from such confounding, Spielman et al. (1993) developed what is termed the transmission disequilibrium test (TDT). In the TDT, one counts the number of times a particular allele is transmitted to an offspring affected with the disease from a parent that is heterozygous at the marker locus. If a particular allele is transmitted from a heterozygous parent (i.e., a parent that has two different alleles at the candidate locus) more than 50 percent of the time (i.e., the chance expectation) this is evidence that, not only is that allele associated with the disease, but that it is also linked to a locus which causes the disease. This is because, under ordinary circumstances, each allele a parent has at a genetic locus has an equal (50 percent) chance of being transmitted to an offspring, so any difference from this expectation is evidence of both linkage and association. Since then, numerous extensions of the TDT approach have been developed for both qualitative and quantitative traits (e.g., Allison 1997, Allison et al. 1999).

4. Contemporary Statistical Methods Based on Latent Genotypes in Humans

The vast numbers of human genotypes are very difficult to measure and are necessarily treated as latent attributes. This is especially true when a complex of influences and a complex of outcomes are considered. In behavior genetics and genetic epidemiology

all statistical methods utilize classical biometric genetic theory to define and specify the relations among the underlying components thought to represent a continuously distributed trait (e.g., Falconer and Mackay 1996). Most models start by asserting that individual genes combine to produce the underlying genotype. Furthermore, these genes are assumed to be: (a) numerous or 'polygenic,' (b) additive in effect, (c) segregating independently (i.e., the probability of receiving one gene is independent of the probability of receiving others), (d) not sex-linked (i.e., located on the X or Y chromosomes), and (e) not influencing characteristics used for spousal selection.

These biometric assumptions are important because they lead to statistical expectations about the observed correlations among pairs of relatives. If all resemblance between relatives in a family is due to genetic sources, the correlations between their observed phenotypic scores are expected to be proportional to the correlations between their genotypic backgrounds, and unrelated individuals would be uncorrelated. More specifically, the genotypes of first-degree relatives correlate (on average in a large sample) at $1/2$. First-degree relatives include fraternal (also called dizygotic or DZ) twins, ordinary siblings, and parent-child relationships. The genotypes of second-degree relatives (grandparents-grandchildren, aunt-niece, half-siblings) are expected to correlate $1/4$, and the third-degree relatives (cousins) at $1/8$, etc. Unrelated individuals living together, such as stepsiblings or parents and their adopted children would be expected to have zero genetic correlation. In the special case of identical (monozygotic or MZ) twins, the genotypes are considered identical and correlate 1.0. Of course, the adequacy of inferences from the sampling of people that are related or unrelated is a major concern of behavior genetic researchers.

Nongenetic influences are often considered as well. In some models the similarity in behavior is expected to be correlated with the similarity in relevant environments (e.g., home of rearing or social background). If true, the highest correlations would be found in scores from twins and similar-aged siblings, lower correlations from parents and children, and still lower correlations from related individuals who have not lived together. Thus, under a purely environmental hypothesis, adoptees should be as similar to their adopting parents and siblings as any other children, and related individuals living in different environments would have scores with no correlation. A model that includes both genetic and environmental sources of familial resemblance would produce estimates intermediate between these other two. All such models can include measurement error and the influence of experiences unique to particular individuals, so even MZ twins living together are not expected to have scores that correlate perfectly.

Different family data permit different assessments of the theoretical components. Figure 4(a) illustrates

this kind of latent path model for a 'family' design. In this model, additive genetic and shared environmental sources of resemblance are confounded, that is, any interpretation of family resemblance cannot be partitioned into separate additive genetic scores (A) and common family environment scores (C). In the adoption design of Fig. 4(b) these A and C components are (theoretically) separable, and the only source of parent-child resemblance is through cultural transmission, resulting in an estimate of common family effects. By comparing differences in resemblance among family members in the adoptive versus biological family designs, an estimate of the additive genetic (A) contribution to the phenotypic deviation is obtained.

Figures 4(c) and 4(d) show models for resemblance between siblings or fraternal (DZ) twins and identical (MZ) twins. In this kind of model, pairs of DZ twins and pairs of full siblings are assumed to have the same genotypic correlation, even though the siblings can differ widely in chronological age. Unlike the family and adoption models, here we cannot estimate the effects of assortative mating based on the correlation between A and C . In most twin models these correlations are assumed to be zero although fixed estimates from other sources can be inserted. Using these assumptions, the twin design can provide estimates based on the difference between the resemblance between MZ and DZ pairs. The special case of MZ twins raised apart (MZA) is important because, in theory, these data permit a complete separation of these latent components.

Structural equation models developed in several areas of science are used in behavior genetics to organize, estimate, and test aspects of these theories of underlying influences (McArdle and Goldsmith 1990, Neale and Cardon 1992). In some recent models an observed phenotype (P) is often cast as a simple combination of true scores (T) and error of measurements (W). Behavioral genetics models go further and include unobserved variables representing the biometric hypotheses: Additive genetic scores (A), non-additive genetic scores (D), common family environments (C), other nongenetic (E) components, and possibly other complex sources. The goal of the structural analysis is independently estimate the total impact (contribution to the variance) of the separate components using the available data from relatives. The statistical basis of the models relies on features common to many scientific analyses—maximum likelihood estimation and likelihood ratio comparisons (e.g., chi-square tests). A sequence of nested statistical tests is routinely used to examine the necessity of each latent component and confidence boundaries can be created.

After specific model coefficients are estimated at a specific value, it is possible to standardize the unobserved variables, assume the latent scores (A , C , and E) are uncorrelated (orthogonal), and calculate the

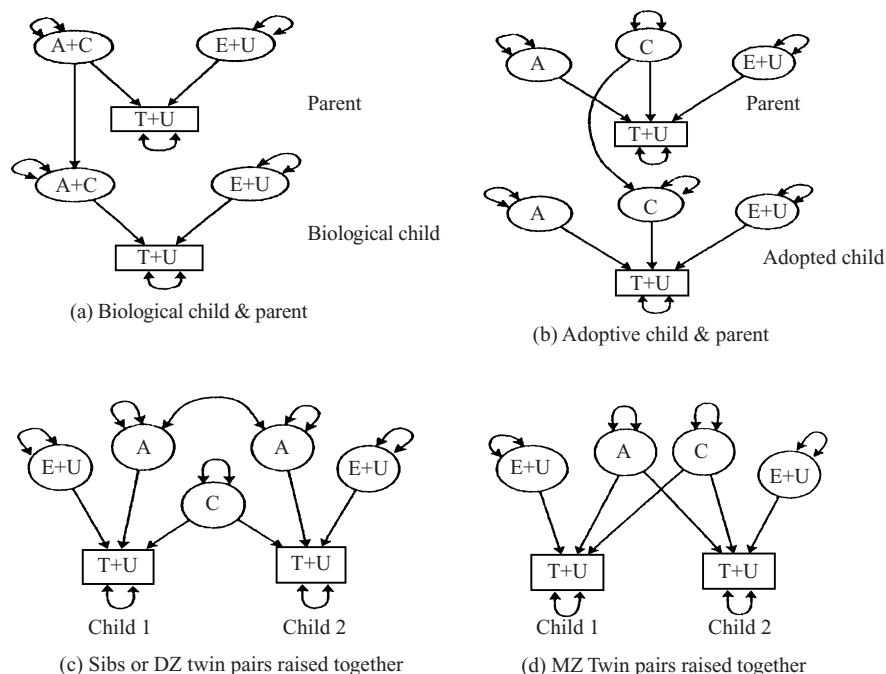


Figure 4

Latent variable path models for the analysis of family resemblance (from McArdle and Prescott 1996). (Note: In these diagrams: (1) Squares are used to represent observed or measured variables. (2) Circles are used to represent unobserved or theoretical variables (3) One-headed arrows are used to represent fixed regression coefficients. (4) Two-headed arrows are used to represent estimated variance or correlation coefficients.)

unobserved genetic variance. In order to compare results for phenotypes with different measurement scales, it has become common to calculate the 'heritability' ratio—the unobserved additive genetic variance divided by the observed phenotypic variance ($h^2 = V_a/V_p$). This kind of calculation may or may not take into account the common family environment, other genetic factors (dominance, epistasis), and is limited by the sample phenotypic variance. In most published research this sample statistic (h^2) is reported and the utility of this routine practice has been seriously questioned (e.g., Gottlieb 1997, Wachs 1999, Hooper in Spector et al. 2000).

There has been a great deal of recent progress towards addressing more complex problems and creating methods that permit the study of more complex patterns of influences. Contemporary behavioral genetic methods do include the possibility of evaluating these more complex patterns of influence, including: parent to child environmental transmission of shared effects, further decomposition of the non-shared environments, the correlation between spouses, and the separation of measurement error via multiple phenotypes. The possibility of a correlation between A and C variables is an important feature of contemporary models. These correlations can be meaningful

but they often cannot be estimated without more extensive data collections and more complex models. In this recent work, parameters for all effects are estimated by assuming the model parameters are invariant over all the multiple groups of Fig. 4. Other important model features include nonadditive genetic effects such as dominance (nonadditive effects of alleles at the same locus) and epistasis (nonadditive effects of genes at different loci) and the general utility of mean intercepts, subject sampling corrections, and multivariate phenotypes (see McArdle and Goldsmith 1990, Neale and Cardon 1992, McArdle and Prescott 1996).

5. Future Issues in Genetic Methods for Behavioral Studies

As with any methodology, there are both strengths and weakness of the genetic methods discussed here. Direct gene measurement permits usually strong experimental designs—in gene association analyses, the appropriate dependent variable is phenotype and the appropriate independent variable is genotype. However, in most studies the genotype is a nearby marker and not the gene of interest. This produces several

unresolved questions that generate skepticism about the value of association tests: (a) Can candidate loci and polymorphisms be identified with sufficient confidence to be useful? (b) Can enough separate polymorphic loci be studied to do genome-wide association studies when candidate loci cannot be confidently identified on an *a priori* basis? (c) Can confounding by population admixture/stratification be controlled? (d) Will applied investigators conduct association studies with adequate rigor even when the tools are available to them?

The challenge of creating hypotheses about candidate genes must be acknowledged, but this is more a limitation than a flaw of the method. The availability of sufficient markers when one wants to go beyond prespecified candidate genes, the ability to economically assess a large number, and the ability to determine how large that number needs to be are currently being addressed by many researchers. The occasional lack of rigor in the conduct of association studies should not be trivialized, but is a problem that potentially affects all areas of research.

There is ongoing research in trying to understand genes and their role in producing phenotypic effects by studying the more complete gene sequence and gene-protein structure. This latter field is often referred to as proteomics and is one of the most active fields in molecular biology at the beginning of the twenty-first century. These methods will become increasingly important in the future and, moreover, will be combined with studies involving information about gene position, gene function, and behavioral measurements to provide a much richer picture about the way genes work to produce their effects. In these more detailed analyses of measured genes we may also be able to study about how gene-protein functions are stimulated by specific environments.

In a similar vein, the statistical methods for latent genotypes rest on assumptions that may not be reasonable representations of the underlying mechanisms, and some of these assumptions are not testable. Statistical biases arise in adoption studies when the effects of selective placement lead to a correlation between the genotypes of the biological and adoptive parents. In twin data, positive assortative mating leads to underestimates of additive genetic effects and overestimates of downward and common family effects. In addition, the standard twin model assumes that both kinds of twins experience similar environments to an equal degree, but the similarity of MZ pairs is in part due to sharing a more similar common environment, the additive genetic effects will be raised spuriously.

But other modeling assumptions are more fundamental and problematic. For example, the classical models above routinely assume no interaction of the genetic and non-genetic components, initially because these additional components are hard to estimate without detailed measurements of both the genotypes

and the environments. Similarly, the concept of genotypes as immutable characteristics with asymmetric direct effects on phenotypes is often challenged. Researchers often suggest that both genes and the environments play an important role in creating constraints on developmental processes (e.g., Waddington 1962, Gottlieb 1997, Wachs 1999).

More complex genetic methodologies are needed to answer these complex questions about complex human characteristics. These are not questions that can be simply answered in terms such as, 'we have found the gene(s) *Q* for behavior(s) *P*,' or, 'recent twin research shows *P* is largely genetic.' Further answers for the questions raised here will, no doubt, combine the best aspects of both measured and latent genotype approaches. Recent research demonstrates the natural convergence of QTL analysis and latent variable structural models (e.g., Neale in Spector et al. 2000). Newer models are likely to include basic information about the phenotypic outcomes in terms of multivariate measurement models combined with longitudinal dynamic modeling.

A clearer understanding of the methods, models and results from these newer analyses may illuminate the complexity of genetic theory and behavioral theory, as well as the appropriateness of scientific inferences that can be made from such studies. We are given pause when we consider that, even after the entire human genome is accurately mapped, we will still be confronted with an enormously complex set of analytic problems—of linking genotypes to proteins, and proteins to physiology, and physiology to basic levels of behavior, and basic behaviors to more complex ranges, constraints, and the resulting limitations on multiple behaviors. The prior methodological advances have carried us very far, but we now look forward to the future development of novel methods for these monumental tasks.

See also: Asymmetry of Body and Brain: Embryological and Twin Studies; Depression; Ethical Issues in the 'New' Genetics; Genes and Culture, Coevolution of; Genetics in Anthropology; Genetics of Complex Traits Through the Life Cycle; Human Evolutionary Genetics; Intelligence, Genetics of: Cognitive Abilities; Intelligence, Genetics of: Heritability and Causation; Schizophrenia

Bibliography

- Allison D B 1997 Transmission disequilibrium tests for quantitative traits. *American Journal of Human Genetics* **60**: 676–90
- Allison D B, Heo M, Kaplan N, Martin E R 1999 Development of sibling-based tests of linkage in the presence of association for quantitative traits that do not require parental information. *American Journal of Human Genetics* **64**: 1754–64
- Burmeister M 1999 Basic concepts in the study of diseases with complex genetics. *Biological Psychiatry* **45**: 522–32

- Dixon L K, Johnson R C 1980 *The Roots of Individuality: A Survey of Human Behavior Genetics*. Brooks/Cole, Monterey, CA
- Elston R C 1998 Linkage and association. *Genetic Epidemiology* **15**(6): 565–76
- Falconer D S, Mackay T F C 1996 *Introduction to Quantitative Genetics*, 4th edn. Longman, London
- Gambaro G, Anglani F, D'Angelo A 2000 Association studies of genetic polymorphisms and complex disease. *Lancet* **355** (9200): 308–11
- Gottlieb G 1998 Normally occurring environmental and behavioral influences on gene activity: From central dogma to probabilistic epigenesis. *Psychological Review* **105**: 792–802
- Johnson R C, McClearn G E, Yuen S, Nagoshi C T, Ahern F, Cole R R 1985 Galton's data a century later. *American Psychologist* **40**(8): 875–92
- Lander E S, Schork N J 1994 Genetic dissection of complex traits. *Science* **265**(5181): 2037–48
- Lynch M, Walsh B 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer Association, Sunderland, MA
- McArdle J J, Goldsmith H H 1990 Some alternative structural equation models for multivariate biometric analyses. *Behavior Genetics* **20**(5): 569–608
- McArdle J J, Prescott C A 1996 Contemporary models for the biometric genetic analysis of intellectual abilities. In: Flanagan D P, Genshaft J L, Harrison P L (eds.) *Contemporary Intellectual Assessment: Theories, Tests and Issues*. Guilford Press, New York, pp. 403–36
- McGill T E 1965 *Readings in Animal Behavior*. Holt Rinehart & Winston, New York
- Neale M C, Cardon L 1992 *Methodology for Genetic Studies of Twins and Families*. Kluwer, Dordrecht, The Netherlands
- Risch N, Merikangas K 1996 The future of genetic studies of complex human diseases. *Science* **273**(5281): 1516–17
- Schork N, Chakravarti A 1996 A nonmathematical overview of modern gene mapping techniques applied to human diseases. In: Mockrin S C (ed.) *Molecular Genetics and Gene Therapy of Cardiovascular Disease*. Marcel Dekker, New York, pp. 79–109
- Spielman R S, McGinnis R E, Ewens W J 1993 Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* **52**(3): 506–16
- Spector T D, Snider H, MacGregor A J (eds.) 2000 *Advances in Twin and Sib-pair Analysis*. Oxford University Press, London
- Strachan T, Read A P 1995 *Human Molecular Genetics*. Wiley-Liss, New York
- Wachs T D 1999 *Necessary But Not Sufficient: The Respective Roles of Single and Multiple Influences of Individual Development*. American Psychological Association, Washington, DC
- Waddington C H 1962 *New Patterns in Genetics and Development*. Columbia University Press, New York
- Wright S 1920 The relative importance of heredity and environment in determining the piebald pattern of guinea pigs. *Proceedings of the National Academy of Sciences* **6**: 320–32

J. J. McArdle and D. B. Allison

Genetic Studies of Personality

Genetic studies of personality focus on (a) evaluating the extent to which genetic influences are important

for individual differences in personality traits and (b) demonstrating whether or not specific genes are associated with personality traits. The thought that genetic effects may be important for personality is not new. Indeed, many cultures have sayings, for example, in Swedish and English, 'the apple does not fall far from the tree,' which suggest that genetic effects are important. Formal studies of genetic effects on personality started with Sir Francis Galton, who examined familial similarity for a variety of human characteristics. During the twentieth century, multitudes of behavioral genetic studies of personality have quantified the relative importance of genetic effects. Genetic differences among individuals account for on average 40 percent of the variance (individual differences) observed in most personality traits. Many specific genes are likely to be associated with personality, each with small effects.

1. Behavioral Genetic Studies of Personality

Twin, family, and adoption studies have been the primary source of information on the relative importance of genetic effects ('heritability') for personality. Behavioral geneticists study the similarity of individuals who share genes and environments to different degrees to quantify heritability. In the classical twin study, identical twins share all of their genes in common whereas fraternal twins share on average only half of their segregating genes in common. Thus, if genetic effects are important for a personality trait, identical twin pairs should be twice as similar as fraternal twin pairs. Similarity between parents and offspring or among siblings may also indicate the influence of genetic effects. The importance of the shared rearing environment can be evaluated by assessing the similarity of adoptees and their adoptive parents. Environmental influences can also be individual specific, or 'non-shared,' i.e., contributing to differences among family members. Behavioral geneticists typically apply structural equation modeling techniques to paper and pencil self-reports of personality when estimating heritability and the importance of shared and non-shared environmental influences (see *Genetic Studies of Behavior: Methodology; Genetics of Complex Traits Through the Life Cycle*).

1.1 Univariate Results

Most genetic studies of personality are based on very large twin and family samples. Personality has been assessed on over 50,000 pairs of twins in the national twin registries in Australia, Finland, Norway, Sweden, and The Netherlands, on volunteer twin registries in the USA, and on the 30,000 individuals in the Virginia study of twins and their families. Numerous smaller twin, family, and adoption studies have also been performed. Many of the studies (particularly those in

the national twin registries) have assessed personality, using versions of the Eysenck Personality Inventory. However, during the 1980s and 1990s increasing amounts of data have become available concerning other personality inventories, such as the Five-Factor Model of Costa and McCrae (McCrae and Costa 1987) and the Seven-Factor Model of Cloninger (Cloninger 1998).

Remarkably, there is considerable consensus across studies and cultures. Genetic effects account for on average 35–40 percent of the variation in personality traits such as neuroticism and extraversion. (In other words, the heritability of these traits is 0.35–0.40.) Evidence from studies of twins reared apart and adoption studies indicate that shared environmental influences account for another 5–10 percent of the variation in most personality traits, and the remaining variation is individual specific or ‘non-shared’ environmental variation. In other words, people differ for personality. Most of the individual differences we observe are due to individual specific experiences not shared by other family members and about 40 percent of the differences are due to genetic differences among us. The extent to which family members are similar to one another reflects primarily their genetic similarity, and only in part being raised in the same home.

There is some heterogeneity in genetic effects on personality. Traits related to extraversion, such as impulsivity, openness to experience, and sensation seeking tend to have slightly higher heritability estimates than those related to neuroticism and instability. Furthermore, the genetic influences for the extraversion-related traits may in part be operating in a nonadditive manner. That is, there may be interactions within or between genetic loci, such that the effect of one gene may depend on another gene. The latter effect, known as epistasis (see *Genetics of Complex Traits Through the Life Cycle*), may complicate efforts to identify specific genes that are of importance for personality traits.

Heritability estimates may be on average 40 percent for most personality traits, but there are traits for which genetic effects are of less importance. Examples of these are ‘agreeableness,’ ‘guilt,’ feelings of control over ‘luck,’ and ‘cynic hostility.’ For these measures, familial similarity is primarily due to environmental influences shared by family members rather than genetic similarity.

1.2 Genetic Influences on Stability and Change

There has been a great deal of controversy as to whether personality is stable throughout the lifespan. Some developers of personality inventories argued that finding significant heritability for personality would support hypotheses of stability in personality. Their argument rested on the misconception that if genetic effects are important they are immutable and

do not change. Despite the fact that humans are born with a full complement of genes, not all genes may be operating at all phases in life. Some genes turn on and others turn off at various transitional periods, while others may react dependent on external stressors. On the other hand, some genes may have a long-term effect that is stable across the lifespan (see *Genetics of Complex Traits Through the Life Cycle*). Thus, the issue of genetic contributions to stability for personality is empirically testable.

During infancy, there is less stability in temperament than is typically reported for adulthood. Nevertheless, results from longitudinal twin studies indicate that the moderate stability observed for infant temperament is influenced in part by genetic effects. During childhood, genetic effects are also important contributors to phenotypic (i.e., observable) stability. Longitudinal studies of adult twins confirm findings of substantial phenotypic stability in personality. Furthermore, the genetic effects themselves are very stable, i.e., the same genes appear to be influencing personality throughout adulthood. Nonetheless, genetic and environmental influences are equally important for phenotypic stability. Even though the genetic effects themselves are very stable, they are relatively less important than environmental effects, and thus the net influence on stability is reduced. Environmental effects are the most important influences on change. Thus, genetic influences on personality contribute to stability rather than to change. However not all personality stability can be explained by genetic influences alone.

1.3 Structure of Personality

Developments in theories of personality have been paralleled with developments in statistical analyses of genetically informative data. During the 1980s and 1990s, the three-factor approach to personality was gradually extended to the five-factor theory and the seven-factor theory (McCrae and Costa 1987, Cloninger 1998). The factor analytic approach generally applied at the phenotypic level is also applicable at the level of genetic and environmental influences. The question to be addressed is whether there is a genetic factor structure or an environmental factor structure that is similar to the phenotypic factor structure. Do genetic or environmental effects mediate the associations among various aspects of personality? This line of analysis has only recently been applied and promises to be an important area for future research concerning the structure of personality.

1.4 Genetic Mediation of Associations with Personality

Health psychologists have demonstrated that personality is important for health outcomes (such as

cardiovascular disease, cancer, and osteoarthritis) and response to treatment such as recovery from surgery. Current research efforts are focusing on the extent to which genetic and environmental effects mediate the associations between personality and health outcomes. By evaluating these associations, much can be learned about the mechanisms by which personality affects health. For example, neuroticism in young adulthood is predictive of joint pain and osteoarthritic symptoms some 25 years later in midlife. Genetic influences in common to neuroticism and joint pain and shared familial environmental influences mediate this association. There are no genetic influences unique to joint pain. Because neuroticism preceded the joint pain, it is likely that genetic effects are important for neuroticism which in turn influences joint pain and osteoarthritic symptoms through response to stress, immune, and inflammatory processes (Turk-Charles et al. 1999).

2. Molecular Genetic Associations with Personality Traits

The second major approach to the study of genetic influences on personality is to test whether specific genetic variants, or alleles, are associated with mean level differences in personality traits. Because personality traits are most often continuous measures and heritability analyses suggest that these traits are polygenic, the challenge is to find genes of medium effect size that influence personality levels. These genes are known as Quantitative Trait Loci or QTLs. Most efforts to identify which genes may be associated with personality traits are currently focusing on genes involved in neurotransmission, such as dopamine receptor genes and serotonin receptor and transporter genes. Many of the findings are only preliminary, and often with contrasting results or lack of replication in other studies or populations.

Finding genes that have relatively small effects will present great challenges to behavioral and genetic scientists, for several reasons. First, because many genes are likely to be involved, each with a small effect, very large sample sizes will be necessary to obtain sufficient statistical power. Second, many of the genes may be interacting with other genes, a phenomenon known as epistasis. For example, a particular allele for gene A may only influence personality negatively if a specific variant of gene B is present. Preliminary findings from large population-based studies suggest that some epistasis exists for personality and genes in the dopaminergic system. Third, there may well be substantial interactions between genes and environments that hamper the potential to find genotypic or even environmental associations. The impact of negative life events is greater in individuals with a genetic susceptibility to major depression, suggesting an interaction between genetic and environmental influ-

ences. Similarly, the effects of some allele may be enhanced or reduced depending on environmental circumstances. Again, large samples with genetic and detailed environmental characterization will be necessary to delineate the respective roles of genes and environments. In this fashion, scientists will come closer to understanding not only how much, but how genes influence personality.

3. Future Directions

Behavioral genetic analyses have provided a firm foundation for further studies on genetic influences on personality. Genetic influences are indeed important, and appear to be important for stability rather than change in personality. Further analyses of the genetic and environmental architecture of personality will expand our understanding of the associations within and among domains, such as the structure of personality and how personality influences health and other behavioral outcomes. Now that the Human Genome Project has provided a genetic map, the challenge to scientists will be to understand the function of the residents at the 30,000 addresses. Many genes, each of small effect, in interaction with each other and in interaction with the environment, are likely to influence variation in normal personality.

See also: Behavioral Genetics: Psychological Perspectives; Cultural Evolution: Overview; Evolution, History of; Gender Differences in Personality and Social Behavior; Genetic Studies of Behavior: Methodology; Genetics of Complex Traits Through the Life Cycle; Personality Development and Temperament; Personality Development in Adulthood; Personality Structure; Temperament: Familial Analysis and Genetic Aspects

Bibliography

- Cloninger C R 1998 The genetics and psychobiology of the seven-factor model of personality. In: Silk K R (ed.) *Biology of Personality Disorders*. American Psychiatric Press, Washington, DC
- Eaves L, Eysenck H, Martin N 1989 *Genes, Culture and Personality: An Empirical Approach*. Academic Press, London
- Loehlin J C 1992 *Genes and Environment in Personality Development*. Sage, Newbury Park, CA
- McCrae R R, Costa P T Jr 1987 Validation of the Five-Factor Model of personality across instruments and observers. *Journal of Personality and Social Psychology* **52**: 81–90
- Turk-Charles S, Gatz M, Pedersen N L, Dahlberg L 1999 Genetic and behavioral risk factors for self-reported joint pain among a population-based sample of Swedish twins. *Health Psychology* **18**: 1–11

N. L. Pedersen

Genetic Testing and Counseling: Educational and Professional Aspects

The sequencing of the human genome has enabled scientists to accelerate substantially the rate of discovery of gene mutations associated with disease. As these discoveries continue to stimulate the development of clinical genetic tests for disease susceptibility, there is a growing need to effectively educate and counsel individuals about the implications of this information. Genetic counseling is the professional practice established to provide this service. The unique familial aspects of genetic susceptibility and conditions have led to the development of medical genetics counseling subspecialties in medicine and nursing, as well as specialized master's degree programs in genetic counseling. While primarily a health communication process, genetic counseling also involves the discussion of relevant psychosocial issues. Thus, it may best be described as psychoeducational. Genetic counseling providers communicate with individuals and their relatives to improve their technical understanding of the cause of genetic conditions and genetic risk. In addition, they address feelings and coping strategies related to the condition and provide support in making preventive, reproductive, and other health-related decisions.

1. Genetic Counseling

1.1 The Role of Genetic Counselors

Members of families with inherited disease often ask not only why the condition occurred, but also why it happened to them. They seek an opportunity to find personal meaning in the senseless nature of it, and search for ways to live with the risk of future disease (Kessler 1979). In this sense, many of the psychological issues associated with inherited diseases are not unique. However, the shared, familial nature of genetic risk raises additional issues that are distinctive. Inherited risk implies that what is a health risk to one individual may also be a risk to their biological relatives. In addition, when a client seeks information about inherited patterns of disease risk, it may have unanticipated consequences. Information previously kept private (such as nonpaternity or adoption) may be revealed, producing enduring effects on family relationships. Yet there also may be potential harms from failing to disclose genetic information. Individuals who remain uninformed of their risk of a preventable disease, such as advanced-stage colorectal cancer, may suffer needless morbidity or mortality.

Because of their training and expertise, genetic counselors are uniquely positioned to assess and explain the genetic nature of a disease and to address its implications. The process aims to be therapeutic in

exploring with clients their feelings of guilt, shame, fatalism, or social isolation. Genetic counseling requires health care practitioners with expertise in medical genetics, knowledge of the disease for which the individual is at risk, effective communication skills, and short-term therapeutic counseling skills.

1.2 Genetic Counseling as a Profession

Until recently, genetic counseling focused primarily on risks for and consequences of rare conditions. Consequently, there are few genetic counseling providers worldwide. Also, because most of the evidence concerning the inherited causes of many diseases is relatively new, the practice of genetic counseling is only about 50 years old (Resta 1997). The short history of the field is reflected in the limited, but growing empirical evidence concerning the process of genetic counseling and its effectiveness in facilitating informed decision making and psychological adjustment.

The type of training received by genetic counseling providers is diverse, and thus it is likely that practices vary accordingly (Biesecker and Marteau 1999). In the United Kingdom and Australasia, genetic nurses often work alongside medical geneticists to provide genetic counseling services. In the USA, Canada, South Africa and increasingly, in the UK, genetic counselors receive specialized training at the master's level to provide genetic counseling. In the vast majority of European and South American countries, medical geneticists or social workers provide genetic counseling. Genetic counseling services are not universally available, as underdeveloped nations usually have no specialized genetics health care providers.

1.3 The Principle of Nondirectiveness

Nondirectiveness is often presented as a guiding ethical principle of genetic counseling due to the importance of personal autonomy in reproductive decision-making. Genetic information even outside the prenatal setting often has implications for future childbearing decisions. The principle of nondirectiveness is intended to suggest that within genetic counseling, patients are not persuaded or coerced into specific reproductive outcomes. However, there exists disagreement in the profession about the meaning of nondirectiveness and its importance in practice (Kessler 1997). An international survey of genetic providers revealed that nondirectiveness is a nearly universal guiding ethical principle in the majority of countries that offer prenatal genetic testing (Wertz et al. 1990). An exception is China, where providers are more inclined to advise their pregnant patients to terminate affected pregnancies in the interest of reducing the incidence of children affected with genetic conditions (Mao and Wertz 1997).

2. Genetic Testing

Genetic testing describes any test for inherited factors that may influence one's health or the health of one's relatives. The method of performing a genetic test ranges from DNA sequencing or direct mutation detection, to chromosomal analysis and protein function testing. In addition, the circumstances in which a genetic test may be done vary substantially. Therefore the genetics education and counseling that surround decisions about testing or implications of test results differ in content and emphasis.

2.1 Diagnostic Testing

Genetic tests have most often been used to diagnose genetic conditions. This has been done in three different ways. First, a genetic test may be used when a child or an adult is suspected to be affected with a certain condition based on the clinical features and pattern of inheritance. Second, newborn screening for a subset of rare largely treatable recessive genetic conditions has been conducted in the USA and most other developed countries for several decades. Third, prenatal genetic testing has been available since the early 1970s, used most often to diagnose Down syndrome and other chromosomal trisomies in the developing fetus.

The diagnostic tests used in a pediatric or adult medical genetics clinic are most often considered a part of good medical management when relatives (typically parents) or their health care providers are seeking to make a diagnosis. Genetics education and counseling typically addresses the genetic and medical implications of the diagnosis, as well as the psychological impact on the family of having to live with an affected family member, and the increased risk to other relatives.

Genetic counseling occurs infrequently in conjunction with newborn screening unless a diagnosis is made in an infant. When a genetic diagnosis is made, the parents usually require extensive education about medical management as well as recurrence risks. Although most of the conditions tested for are treatable (unlike the vast majority of single-gene conditions), they require exhaustive efforts by parents to keep their child on a specially restricted diet to ensure his/her health.

Prenatal genetic testing is most often preceded by genetic counseling due to the highly personal nature of reproductive decision making. In most countries where prenatal testing is available, it is offered as an option to pregnant women who are at increased risk of having an affected child. Most commonly, this concerns those over the age of 35 years. Prenatal genetic counseling addresses the personal values of the client or couple, as they are guided in the exploration of the potential consequences of adverse test results. This

includes envisioning both what their lives might be like if they had an affected child, or how they would feel subsequent to terminating a desired pregnancy.

2.2 Predictive Testing

Genetic testing is also used to predict genetic risk. Carrier testing, for example, is used to identify those who carry a gene mutation that does not affect them directly, but may place their children at risk. Relatives of individuals affected with certain genetic conditions and members of certain populations (e.g., Ashkenazi Jews) known to be at increased risk for certain conditions, are frequently offered carrier screening. Most carrier screening programs are accompanied by pretest genetics education, since it is important that those found to be carriers do not misinterpret the test result as indicating they have the disease.

Presymptomatic genetic testing is used to predict certain risk to an individual. Presymptomatic testing for Huntington's disease, a degenerative neurological disorder, is one example. It is now possible for those at increased risk for Huntington's disease to undergo presymptomatic testing prior to the onset of clinical symptoms. Inevitably, those found to carry a gene mutation develop the disease, for which there is no treatment. Due to the dire nature of these circumstances, the counseling that proceeds such testing is often extensive. Before presymptomatic testing was available, the majority of individuals at increased risk of developing Huntington's disease reported that they would choose to undergo testing if it was developed. However, once testing was available, only a small minority of those at increased risk chose to go forward with testing.

The newest type of predictive genetic testing, predisposition or susceptibility testing, results in a risk estimate about one's own future health. The first predisposition tests developed assess inherited susceptibility for cancer. An increased risk for at least two different types of colon cancer and breast and ovarian cancer can now be determined in members of hereditary cancer families. There are also populations of individuals more likely to carry susceptibility mutations due to shared heritage.

Susceptibility tests are being developed for many other common diseases and will be offered to an increasingly large portion of the population of developed nations. As demand for these tests increases, alternatives to in-person genetic counseling for pretest education will need to be developed and evaluated. Computer and Internet-based resources are likely to grow in their importance as sources of genetics information. These technologies allow information to be tailored to the characteristics of the individual user, allowing genetic counselors to focus on counseling and decision making rather than education and information delivery.

Initial studies of cancer susceptibility testing suggest that the majority of people express high levels of interest, but that many members of high-risk families do not utilize testing once it is offered (Croyle and Lerman 1999). The exploration of the psychological and behavioral costs and benefits of this type of testing have only just begun. Although individuals found to be noncarriers of cancer gene mutations report an alleviation of distress after testing, many mutation carriers do not engage in health behaviors that would allow the early detection of disease. These observations suggest that social and behavioral scientists will continue to play a key role in identifying ways to translate the achievements of genetics and biotechnology into improvements in public health.

See also: Bioethics: Examples from the Life Sciences; Cancer Screening; Ethical Issues in the 'New' Genetics; Ethics and Values; Eugenics as an International Movement; Eugenics, History of; Eugenics in Asia; Eugenics in Europe; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Engineering; Genetic Screening for Disease-related Characteristics; Mental Illness, Genetics of; 'New' Eugenics, The: Role in Society; Reproductive Medicine: Ethical Aspects; Risk Screening, Testing, and Diagnosis: Ethical Aspects

Bibliography

- Biesecker B B, Marteau T 1999 The future of genetic counselling: An international perspective. *Nature Genetics* **22**: 133–7
- Croyle R T, Lerman C 1999 Risk communication in genetic testing for cancer susceptibility. *Journal of the National Cancer Institute* **25**: 59–66
- Kessler S 1979 *Genetic Counselling: Psychological Dimensions*. Academic Press, New York
- Kessler S 1997 Psychological aspects of genetic counseling. XI. Non-directiveness revisited. *American Journal of Medical Genetics* **72**: 164–71
- Mao X, Wertz D C 1997 China's genetic services providers' attitudes towards several ethical issues: A cross-cultural survey. *Clinical Genetics* **52**: 100–9
- Resta R 1997 Historical perspective: Sheldon Reed and 50 years of genetic counseling. *Journal Genetic Counseling* **6**: 375–8
- Wertz D C, Fletcher J C, Mulvihill J J 1990 Medical geneticists confront ethical dilemmas: Cross cultural comparisons among 18 nations. *American Journal of Human Genetics* **46**: 1200–13

R. T. Croyle

Genetics and Development

This article describes the major historical misconceptions of the role of genetics in the development of the individual, whether human or nonhuman animal,

and also presents our current understanding of the necessarily co-operative roles of genes and environments in the developmental process.

1. Nature vs. Nurture

Many people believe that behavioral characteristics such as personality or intelligence are determined primarily, if not exclusively, either by heredity or environment. This belief, known as the nature-nurture controversy, has been with us for quite a long period of time and is the cause for much that is sorrowful and tragic in human history (e.g., the notion that some groups are behaviorally inferior or superior to other groups because of hereditary/genetic predispositions). The idea that genes are primarily responsible for some traits and that the environment is primarily responsible for other traits is now known to be scientifically invalid. This has been a hard-won scientific insight that has not yet percolated down to the mass of humanity. It is now known that both genes and environments are involved in *all* traits and that it is not possible to specify their respective weighting or quantitative influence on any trait. These crucial points are documented in the sections that follow.

2. Behavioral Genetics

In a remarkable experiment first conducted by E. C. Tolman (1924) and continued by R. C. Tryon (1942), it was found that it was possible to selectively breed laboratory rats for their ability to learn their way through a maze. Tryon selectively inbred the superior learners among each other and the poor learners among each other. After only seven generations of breeding the brightest rats within each of the brightest litters, and the dullest within each of the dullest litters, the performance of the two groups in learning the specific maze was virtually nonoverlapping, leading Tryon to label the superior line 'maze-bright' and the inferior line 'maze-dull.' It was as if Tryon had been able, by selective breeding, to create two lines of animals which differed greatly in their general learning ability. That turned out not to be the case, however. When Searle (1949) later compared the 'bright' and 'dull' lines on a variety of learning and performance tasks, the dull rats were sometimes superior and often equal to the rats from the bright line. Being bright or dull was not a property of the animals, but, rather, was restricted to the particular learning task that served as the basis for selectively breeding the animals. As it turned out, the specifics of the outcome of the breeding program were even further constrained by the manner in which the rats were reared, as will be described in Sect. 3.

Since the time of Tolman's and Tryon's breeding experiments, the field of behavioral genetics has been able to show that it is possible to selectively breed

Table 1

Examples supporting the notion that it is possible to selectively breed animals for differences in any behavior (more-less, better-poorer, higher-lower)

Species	Behavior
Rats	Maze learning
Mice	Speed in traversing runway
Rats	Spontaneous activity (wheel-running)
Rats	'Emotionality' in an open field
Mice	Sound-induced (audiogenic) seizures
Fruit flies	Orientation toward or away from light source (positive and negative phototaxis)
Rats	Saccharin preference
Chickens	Aggressiveness
Fruit flies	Upward or downward orientation with respect to gravity (negative and positive geotaxis)
Fruit flies	Mating speed
Mice	Alcohol preference
Fruit flies	General activity

The literature sources for these experiments can be found in Gottlieb 1992, p. 150.

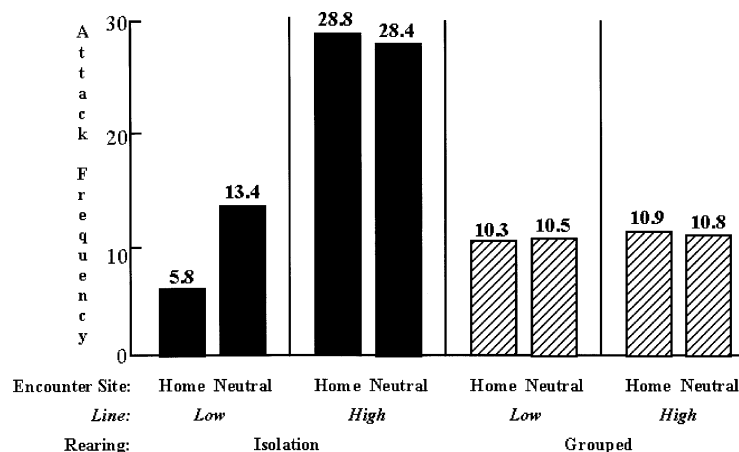
animals for differences on any trait, solidifying the idea mentioned earlier that genes are involved in *all* traits. As will be documented below, the results of behavioral genetics breeding experiments are also dependent on the rearing environment being kept constant from generation to generation (recall that the rearing environment is involved in all traits, too). As can be seen in Table 1, both favorable and unfavorable behaviors can be made the subject of selective breeding, not only maze learning but spontaneous activity

(wheel-running), 'emotionality,' sound-induced seizures, saccharin preferences, aggressiveness, mating speed, and so on.

As mentioned above, since all behavioral activities involve both genetic and environmental contributions, it is essential when doing selective breeding experiments to keep the rearing environment the same from generation to generation, otherwise the selective breeding experiments will be a failure. For example, in an experiment by Hood and Cairns (1989), mice were selectively bred for high and low aggressiveness. The standard rearing procedure was to keep the animals in isolation after they were weaned. After the high and low aggressive lines were established, to show the contribution of the isolation rearing procedure to the results of selective breeding Hood and Cairns reared half the animals from each line in social groups rather than in isolation. As can be seen in the right-hand side of Fig. 1, the high aggressive line did not differ from the low aggressive line when the animals were reared in social groups; the high aggressive line was only highly aggressive when reared in isolation (left-hand side of Fig. 1), the rearing condition under which the original selective breeding took place. These results lead into the topic of Sect. 3.

3. Reaction Range vs. Norm of Reaction

Until very recently some biologists and psychologists held the incorrect belief that genes strictly canalize behavioral development, creating a predictable reaction range when rearing environments are changed from the normal or usual (species-typical) to highly

**Figure 1**

Average attack frequency by male mice in two lines selectively bred for either high or low attack frequency. The original selection experiment involved males reared in isolation prior to testing. In this experiment, half of the males were reared in isolation and the others were group-reared, greatly reducing the attacks by the high line. During the test, the mice met in either the home cage, in which the mouse being tested was an intruder, or in a neutral cage (based on data from Hood and Cairns 1989. Figure modified from Gottlieb et al. 1998, p. 249)

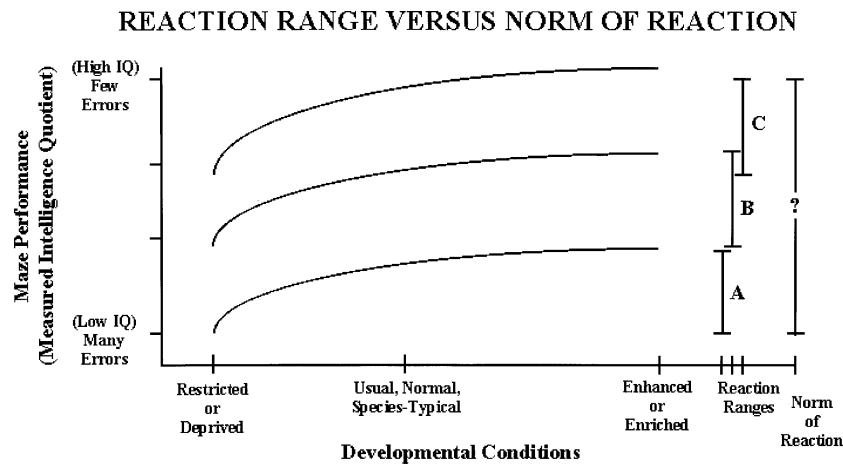


Figure 2

Contrasting predictions of reaction range and norm of reaction for maze performance or IQ of three genotypes or individuals, as a consequence of being reared under three developmental conditions. The reaction-range concept predicts a narrow range of variation for each individual over the three rearing conditions, whereas the norm of reaction predicts an indefinitely broader range of outcomes when the rearing conditions are changed from the original condition (usual, normal, species-typical) (modified from Gottlieb 1992, p. 178)

favorable (enriched) or unfavorable (deprived) conditions.

This point of view was expressed by one authority as follows:

Canalization is the genetic limitation of phenotypic development to a few possible phenotypes...The genetic restriction of possible phenotypes is a result of coadapted gene complexes (sets of genes that have evolved together) that buffer the developmental pattern against deviation outside of the normal, species range of variation...The limitation of possible phenotypes to a few rather than many is a fact of genotypes' ranges of reaction (Scarr-Salapatek 1976, p. 63).

The belief in the sheerly genetic canalization of behavioral development leads quite logically to the reaction-range idea. Thus, the basic idea of a reaction range is that genes determine (canalize) the upper and lower limits of psychological development so that, given a knowledge of the outcome of development under one rearing condition, one can make linear predictions of the upper and lower limits of outcomes under other rearing conditions (see as an example Fig. 2).

In contrast to the reaction range, the norm of reaction holds that the outcome of development is an emergent (epigenetic) property that is a nonlinear result of co-actions occurring at all levels of the developmental system (including but not restricted to genes and environments, as elaborated in Sect. 4). Thus, it is not possible to predict outcomes from one rearing circumstance to the next. Unfortunately, it is the reaction range, not the reaction norm, that has

had, until very recently, wide currency in psychology, as evidenced in many textbooks on developmental psychology. These texts reproduced a version of Fig. 2 depicting only the reaction-range concept for learning ability or mean IQ for three hypothesized genotypes reared in very different environments, and not the norm of reaction (far right).

The difference between the predictions of the reaction-range and reaction-norm concepts is testable. As indicated in Fig. 2, the reaction-range idea predicts that the ranking of the different genotypes on IQ or learning ability will remain the same when they are reared in different environments. In the enriched environment, each genotype will show an improvement but the 'brightest' genotype will still rank above the other genotypes. In the deprived rearing environment each genotype's performance will be poorer but the brightest genotype will still rank above the other genotypes. As shown in the far right of Fig. 2, the norm of reaction makes no prediction on relative ranking when the genotypes are reared in environments (enriched or deprived) that differ from the rearing environment in which the selective breeding took place.

There is only one empirical study in the psychological literature that explicitly addresses the reaction-norm concept, a study by Cooper and Zubek (1958). The results very clearly support the reaction-norm concept. It is interesting to note that the study was carried out with the idea of a reaction range in mind, and it is sometimes erroneously cited as supporting the reaction-range concept. Cooper and Zubek reared maze-bright and maze-dull rats in either an enriched

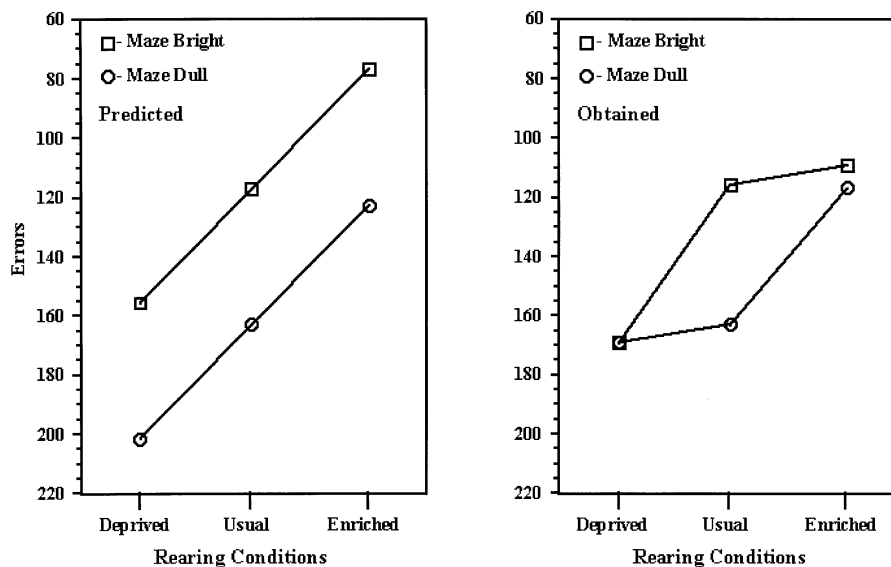


Figure 3

Behavioral reaction range (predicted) and norm of reaction (obtained) for maze-bright and maze-dull rats' performance in a Hebb-Williams maze after rearing in three different environments. Obtained deprived and enriched data points from Cooper and Zubek (1958); obtained usual data points from Hughes and Zubek (1956). Only the obtained usual data points are significantly different from each other (from Gottlieb 1995, p. 137)

BIDIRECTIONAL INFLUENCES

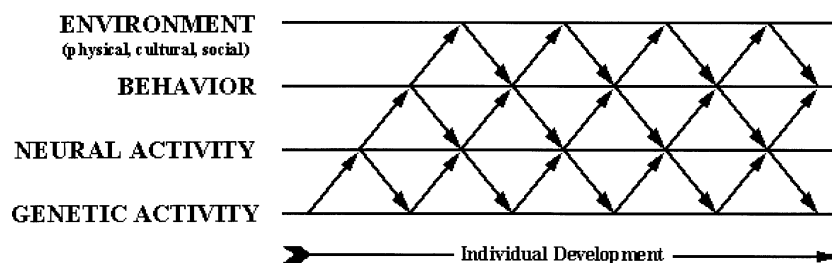


Figure 4

Depiction of the completely bidirectional nature of genetic, neural, behavioral, and environmental influences over the course of individual development (from Gottlieb 1992, p. 186)

or a restricted environment and then tested them in a Hebb-Williams learning maze. Since they had the reaction-range concept in mind in performing the experiment, Cooper and Zubek thought that the learning of both the bright and dull rats would improve relative to each other under the enriched rearing circumstances and would be poorer relative to each other when reared under the restricted (deprived) condition. This prediction is illustrated in the left-hand side of Fig. 3. Instead, as shown on the right-hand side of Fig. 3, they found equality of performance under both of the new rearing conditions. The dull rats made as few errors as the bright rats after enriched

rearing and the bright rats made as many errors as the dull rats after restricted rearing.

When the so-called bright and dull rats were tested in the Hebb-Williams maze after being reared in their usual way (neither enriched nor deprived), a significant difference between the lines appeared (middle points, right-hand side of Fig. 3). The reason is that this developmental situation repeats the rearing condition under which the original selective breeding for superior and inferior performance was carried out (Hughes and Zubek 1956). If the reaction-range idea were correct and the genes coded for upper and lower limits of learning ability, when these lines were reared under

Table 2

Normally occurring environmental and behavioral influences and gene activity

Species	Environmental signal or stimulus	Resulting alteration
Nematodes	Absence or presence of food	Diminished or enhanced neuronal <i>daf-7</i> gene mRNA expression, inhibiting or provoking larval development
Fruit flies	Transient elevated heat stress during larval development	Presence of proteins produced by heat shock and thermotolerance (enhanced thermal regulation)
Fruit flies	Light-dark cycle	Presence of PER and TIM protein expression and circadian rhythms
Various reptiles	Incubation temperature	Sex determination
Songbirds (canaries, zebra finches)	Conspecific song	Increased forebrain mRNA
Hamsters	Light-dark cycle	Increased pituitary hormone mRNA and reproductive behavior
Mice	Acoustic stimulation	Enhanced <i>c-fos</i> expression, neuronal activity, and organization of the auditory system
Mice	Light-dark cycle	<i>c-fos</i> -induced mRNA expression in hypothalamus, circadian locomotor activity
Rats	Tactile stimulation	Enhanced <i>c-fos</i> expression and increased number of somatosensory (sense of touch) cortical neurons
Rats	Learning task involving vestibular (balance) system	Change in nuclear RNA base ratios in vestibular nerve cells
Rats	Visual stimulation	Increased RNA and protein synthesis in visual cortex
Rats	Environmental complexity	Increased brain RNA diversity
Rats	Prenatal nutrition	Increase in cerebral DNA (increased number of brain cells)
Rats	Infantile handling; separation from mother	Increased hypothalamic mRNAs for corticotropin-releasing hormone throughout life
Cats	Visual stimulation	Increased visual cortex RNA complexity (diversity)
Humans	Academic examinations taken by medical students (psychological stress)	Reduced mRNA activity in interleukin 2 receptor (immune system response)

MRNA = messenger RNA; PER and TIM are proteins arising from activity of per (period) and tim (timeless) genes; activity of *c-fos* genes leads to production of C-FOS proteins. Sources documenting the findings can be found in Gottlieb 1998, p. 796

enriched or restricted conditions, the relative differences between them would be preserved. Instead, the experiment shows that the genes are part of a developmental system. The highly specific consequences of rearing under a certain developmental condition were realized by selective breeding under that condition. As called for by the norm-of-reaction concept, selective breeding under one developmental regimen (environment) does not predict outcomes under different rearing conditions. The results of selection depend on the entire developmental manifold, not merely on the genes that are involved.

4. Understanding Development from Genes to Behavior

To provide some insight into what would today constitute a truly developmental behavioral-genetic analysis, Fig. 4 shows a diagram of the four major levels that are involved in moving from genetic activity to environment and back again. This is a fully bidirectional co-actional system, with traditionally trained developmental psychologists working at the Behavior ↔ Environment level, developmental neuro-psychologists and neuroscientists at the Neural

Activity ↔ Behavior ↔ Environment levels, and biologists beginning now to work on the Genetic Activity ↔ Neural Activity levels. Viewed in this way, developmental understanding or explanation is a multilevel affair involving at least culture, society, immediate social and physical environments, anatomy, physiology, hormones, cytoplasm, and genes. Hierarchical, multilevel, or developmental systems' analysis is methodologically not theoretically reductionistic in the sense that psychological understanding or explanation does not come from lower levels (or reside in the lower levels), but that both higher and lower levels of analysis are necessary to explain developmental outcomes. The term co-action is used to emphasize the multilevel systems' nature of developmental analysis.

The notion that genetic activity is necessarily influenced by supragenetic factors (factors above the genetic level) may be unfamiliar, in part because of what has come to be known as the Central Dogma of Molecular Biology. The Central Dogma holds that DNA → RNA → Protein, in which DNA serves as a template for RNA which then serves as a template for the synthesis of a protein. In the Central Dogma there are master genes which contain a blueprint of the organism, so 'information' is said to flow outward from these master genes providing instructions for individual development. The actual story is rather different. Rather than being encapsulated (or 'safely sequestered in the nucleus of the cell,' as some biologists have written), the genes require signals from the internal and external environment to function appropriately. As mentioned above, genes are part of the developmental system of influences; they do not stand outside or apart from the developmental system as the Central Dogma would seem to indicate. Gene expression is influenced by cytoplasmic factors, which provide the immediate environment of the nucleus of the cell. (Structural DNA is found in the nucleus of each cell.) Cytoplasm can be influenced by the external environment of the organism (reviewed by Ho 1984) via the behavior of the organism, which gets the organism into and out of exposure to different environmental influences (physical, social, cultural). Likewise, hormones make their way into the nucleus of the cell and trigger gene expression, and hormones themselves are responsive to the organism's experiences in the external world. In fact, the action of some genes is dependent on sensory stimulation early in development (reviewed by Arenander and Herschman 1993). Finally, different proteins are formed depending on the particular factors influencing gene expression (Davidson 1986). Hence, genes do not stand outside the developmental system acting as independent causes. They are part of the developmental system (for a detailed discussion, see Wahlsten in press). One of the main reasons for the failure of gene therapy as currently conceptualized and practiced (Rosenberg and Schechter 2000) is that the genes do not function

as independent causes; they do not stand outside the developmental system.

Since there are in fact no master genes telling other genes what to do, in order for genes to function appropriately they require signals from the internal and external environment. Table 2 presents a sampling of the normally occurring external environmental and behavioral signals that activate genes during the normal course of development in a variety of species including humans. The findings summarized in Table 2 substantiate the main point of this article: genes and environments necessarily cooperate during the course of individual development.

See also: Behavioral Genetics: Psychological Perspectives; Genes and Behavior: Animal Models; Genetic Factors in Cognition/Intelligence; Genetics of Complex Traits Through the Life Cycle; Human Evolutionary Genetics; Intelligence, Genetics of: Cognitive Abilities; Intelligence, Genetics of: Heritability and Causation

Bibliography

- Arenander A T, Herschman H R 1993 Primary response gene expression in the nervous system. In: Loughlin S F, Fallon J H (eds.) *Neurotrophic Factors*. Academic Press, San Diego, CA
- Cooper R M, Zubek J P 1958 Effects of enriched and restricted early environments on the learning ability of bright and dull rats. *Canadian Journal of Psychology* **12**: 159–64
- Davidson E H 1986 *Gene Activity in Early Development*. Academic Press, Orlando, FL
- Gottlieb G 1992 *Individual Development and Evolution: The Genesis of Novel Behavior*. Oxford University Press, New York
- Gottlieb G 1998 Normally occurring environmental and behavioral influences on gene activity: From central dogma to probabilistic epigenesis. *Psychological Review* **105**: 792–802
- Gottlieb G, Wahlsten D, Lickliter R 1998 The significance of biology for human development: A developmental psychological systems perspective. In: Damon W, Lerner R M (eds.) *Handbook of Child Psychology: Theoretical Models of Human Development*, 5th edn. Wiley, New York, Vol. 1
- Ho M-W 1984 Environment and heredity in development and evolution. In: Ho M-W, Saunders P T (eds.) *Beyond Neodarwinism: An Introduction to the New Evolutionary Paradigm*. Academic Press, London
- Hood K E, Cairns R B 1989 A developmental-genetic analysis of aggressive behavior in mice: IV. Genotype-environment interaction. *Aggressive Behavior* **15**: 361–80
- Hughes K R, Zubek J P 1956 Effect of glutamic acid on the learning ability of bright and dull rats. I. Administration during infancy. *Canadian Journal of Psychology* **10**: 132–8
- Rosenberg L E, Schechter A N 2000 Gene therapist, heal thyself. *Science* **287**: 1751
- Searle L V 1949 The organization of hereditary maze-brightness and maze-dullness. *Genetic Social and General Psychology Monographs* **39**: 279–325

- Scarr-Salapatek S 1976 Genetic determinants of infant development: An overstated case. In: Lipsett L (ed.) *Developmental Psychobiology: The Significance of Infancy*. Erlbaum, Hillsdale, NJ
- Tolman E C 1924 The inheritance of maze-learning ability in rats. *Journal of Comparative Psychology* 4: 1–18
- Tryon R C 1942 Individual differences. In: Moss F A (ed.) *Comparative Psychology*. Prentice-Hall, New York
- Wahlsten D in press Genetics and the development of brain and behavior. In: Valsiner J, Connolly K (eds.) *Handbook of Developmental Psychology*. Sage, London

G. Gottlieb

Genetics and Mate Choice

In a sexually reproducing species, an individual must mate to reproduce. Mates are not chosen at random, but apparently according to sexual preferences. Such preferences may be genetically fixed at birth, or learned—acquired by sexual imprinting, mate-choice copying, etc. To the extent that they are reflected in actual matings, preferences may have immediate consequences for reproductive success and far-reaching effects on genetic evolution.

1. Avoidance of Inbreeding

Consanguinity is an important criterion of mate choice. Throughout the vertebrate, invertebrate, and plant kingdoms, the propensity to inbreed (mate with a blood relative) varies widely (Thornhill 1993). Some species of hermaphroditic plants mate predominantly with themselves—the seed is fertilized by pollen from the same individual. In others, mechanisms such as self-incompatibility have evolved to prevent self-fertilization. Hence, self-fertilization is not the unavoidable consequence of female and male functions occupying the same body. In species where the two sexes are separate, nuclear family incest—matings between father and daughter, mother and son, or full sibs (brother and sister with the same mother and father)—is the most extreme form of inbreeding. For some insects, mother–son and brother–sister matings are routine.

On the other hand, close inbreeding is unusual in birds and mammals. Of special interest are the non-human primates, where nuclear family incest, in particular, is rarely observed. Adult males of the chimpanzee, our closest relative, have seldom been seen mating with their mothers. Most, but not all, females refuse the sexual advances of males in their natal group old enough to be their fathers. The incidence of copulations between inferred maternal (possibly full) sibs is low (Pusey 1980). However, matings between paternal half sibs may not be uncommon in the gorilla.

The claim is often made that behavioral mechanisms have evolved in birds and mammals to prevent close inbreeding. However, avoidance of incest is not necessarily based on recognition of family members. Because of a sex difference in dispersal, sons and daughters may not encounter their opposite-sex sibs or parents as adults. In the black-tailed prairie dog, for example, daughters usually remain in their natal area whereas the sons emigrate. Fathers may also leave or be evicted before their daughters mature. Under these circumstances, incest may be prevented even when matings are indiscriminate. The reason for such a dispersal pattern is unknown, but avoidance of inbreeding is a possible explanation. It is perhaps surprising, then, that the incidence of close inbreeding, mostly between half sibs, does not differ from the number expected if matings were occurring at random in the study area (Hoogland 1992).

Nevertheless, kin recognition has been implicated in some field studies and demonstrated in the laboratory. A likely mechanism is ‘sexual imprinting,’ the phenomenon in which sexual preferences expressed later in life are acquired through exposure at a very young age to other individuals, usually parents or sibs. A prairie dog female, for example, is less likely to come into estrus when her father is still present. Similarly, an older female of the medium ground finch mates with a male who sings a song differing from her father’s. Since song is culturally transmitted from father to son in this species, close kin are apparently recognized and avoided (Grant and Grant 1996). Finally, Japanese quail reared with sibs choose the company of first cousins over that of either sibs, third cousins, or unrelated individuals. Bateson (1982) infers from this experiment that quails learn the characteristics of close kin and subsequently develop a preference for mates differing somewhat from familiar individuals.

It is interesting that cousin marriage, which is a comparatively mild form of inbreeding, is highly popular in some human societies. In particular, in south India and in many Muslim populations, the incidence of first-cousin marriages exceeds 30 percent (Bittles 1994). The reasons for this preference are likely different from those operative in quails.

In contrast to cousin marriage which exhibits much regional variation in incidence, nuclear family incest, marital and nonmarital, has been prohibited and usually avoided in almost all human societies past and present that we know about. A well-known counter example comes from Roman Egypt, where one-sixth of all marriages among commoners were between full sibs (Bagnall and Frier 1994). The evidence comprises over 300 census returns preserved on papyri, from which pedigrees have been reconstructed. Although this example is certainly the exception rather than the rule, it cautions against naive explanations for the existence of incest taboos.

Westermarck (1891) proposed that ‘there is an innate aversion to sexual intercourse between persons

living closely together from early youth.' In particular, full sibs reared together are expected to avoid mating with one another, but not full sibs reared apart. Avoidance is also predicted when unrelated boys and girls are reared together in the same household. Thus, studies on the Chinese custom of *sim-pua* and on marriage patterns among Israelis reared together as children in kibbutzim seem to support the 'Westermarck hypothesis.' For example, the *sim-pua* is a girl who is adopted, often as an infant, into a family with a young boy, to be his future bride ('little daughter-in-law'). After being brought up as sister and brother, they are forced to marry. Compared to other arranged marriages, the fertility of this type of marriage (minor marriage) is low, and the divorce rate is high (Wolf 1995).

This result is usually regarded as reflecting sexual dissatisfaction and supporting the Westermarck hypothesis. Nevertheless, marriages between foster sibs were in fact consummated. Moreover, in the case of institutionalized sib marriage in Roman Egypt, any aversion that may have existed was overridden. Hence, even if negative sexual imprinting occurs, it is apparently not sufficient to prevent an incestuous mate choice when other cultural pressures intervene.

If behavioral mechanisms such as sexual imprinting and sex-biased dispersal have evolved by natural selection to prevent close inbreeding in birds and mammals, there must be a selective disadvantage to mating with close kin. Sexual imprinting and sex-biased dispersal are assumed to be genetically determined capacities—hence innate. An inbred (born of related parents) offspring is known to have lower fitness than an outbred (born of unrelated parents) one. The decrease in fitness, normalized by the fitness of the outbred offspring, is called the inbreeding depression, and is estimated in practice by comparing survival to some arbitrary age. Inbreeding depression is believed to be caused by homozygosity for (having two copies of) deleterious recessive genes. The more closely inbred an individual is, the more severe the inbreeding depression is expected to be.

Estimates of the inbreeding depression associated with nuclear family incest average 33 percent for 38 species of mammals in captivity (Ralls et al. 1988). In humans, one study with good controls obtained an estimate of 29 percent, although this estimate is inflated by the inclusion of morbidity (Adams and Neel 1967).

The conditions under which avoidance of inbreeding will evolve are difficult to compute, but some theory, mostly heuristic, exists for full-sib matings (Maynard Smith 1980). If all males have the same number of mates, then sib mating is selected against, and avoidance will evolve. This would be true of a strictly monogamous species. However, if the inbreeding depression is less than one-third, and a female by being incestuous can give her brother one extra mating, then it will be in her interests to do so.

Another factor that may offset the natural selection against incestuous matings is 'reproductive compensation'—if offspring lost due to inbreeding depression are replaced. Similarly, if ecological or economic conditions are such that sibs can mate earlier than unrelated individuals, then lifetime reproductive success may be comparable—or even greater—for the incestuous pair. Keller and Arcese (1998) argue that this may be the reason why song sparrows do not purposefully avoid close inbreeding. In fact, the incidence of full-sib and parent-offspring matings was as expected if song sparrows were mating indiscriminately.

In modern human societies, age at marriage is lower when the spouses are cousins—and generally when related—than when they are unrelated. Perhaps because of this, or because of reproductive compensation, cousin marriages enjoy a higher fertility (live births) than nonconsanguineous unions (Bittles 1994). When on-going marriages in Roman Egypt are broken down by age, there is a relatively higher incidence of brother-sister and other close-kin marriages among the younger couples. Therefore, it is possible that sibs were marrying earlier than unrelated couples. Moreover, one of the most fertile couples are sibs, with eight surviving children (Bagnall and Frier 1994). If sib mating permitted an early start in reproduction during much of human evolutionary history, then sib mating rather than its avoidance may have been favored by natural selection.

2. Preference for a Mate Differing at the Major Histocompatibility Complex

In (positive) assortative mating, psychologically, behaviorally, or physically similar individuals pair up. Negative assortative, or disassortative, mating occurs as a result of attraction between dissimilar individuals. An example of positive marital assortment is the tendency of deaf persons to marry one another. Another example is the high correlation between the intelligence quotients of spouses. In the latter example, however, it is not known whether someone with an IQ comparable to one's own is actually preferred as a marriage partner. Rather, people who marry one another have often experienced similar environments as children, and since the environment is a major determinant of IQ, spouses may incidentally have similar IQs.

Recently, Wedekind et al. (1995) suggested that humans prefer a mate differing at the major histocompatibility complex (MHC). MHC is an essential part of the vertebrate immune system. It comprises many genes, each of which is highly variable and exists in many alternative forms called alleles. In the human, the genes are called HLA-A, HLA-B, HLA-C, etc. (HLA = human leucocyte antigen.) For illustrative purposes, let us focus on the HLA-A gene and assume

that four alleles are present, which will be called A1, A2, A3, and A4. The pair of alleles in an individual defines the genotype of that individual. The genotypes could be A1A1, A2A2, A3A3, or A4A4, which are called homozygotes, or A1A2, A1A3, A1A4, A2A3, A2A4, or A3A4, which are called heterozygotes.

Mates can share two alleles (as when the genotypes of husband and wife are A1A2 and A1A2), one allele (A1A2 and A2A3, for example), or no alleles (A1A3 and A2A4, say). The hypothesis of disassortative mating asserts that sexual attraction is negatively correlated with the number of alleles shared. Hence, an individual of genotype A1A1, say, given a choice of three different partners whose genotypes are A2A2, A1A2, and A1A1, say, is predicted—all other things being equal—to prefer the first (A1A1 and A2A2 share no alleles) over the second (A1A1 and A1A2 share one allele) over the third (A1A1 and A1A1 share two alleles).

Theoretically speaking, disassortative mating for MHC may make evolutionary sense. Yamazaki et al. (1976) argue as follows. Heterozygotes for MHC genes have a higher fitness than homozygotes, because the presence of two alleles rather than just one would permit an immunological response to a wider range of antigens. Although the genotype that an individual is born with cannot be altered, the genotype of offspring is under some personal control. Namely, by choosing an MHC-dissimilar partner, the chances are improved that offspring will be heterozygotes. A second reason why disassortative mating for MHC might be favored by natural selection is that it leads to avoidance of inbreeding.

A preference for MHC-dissimilar mates is fairly well established in house mice. There is good evidence, at least in mice, that this preference is mediated by body odor, which is influenced by the MHC genes. More precisely, mice imprint on the MHC identities of the individuals by, or with, whom they are raised, who under normal circumstances would be extended family. They subsequently choose a mate differing in MHC from these individuals. Since, close relatives are on average more similar for their MHC than unrelated individuals, the resulting preference would be for MHC-dissimilar mates.

The human evidence is of two kinds. First, in a provocative experiment, male subjects were asked to each wear a T-shirt, and female subjects to rate the odors of these T-shirts for pleasantness (sexiness). The subjects, all of whom were students at a Swiss university, were also typed for their HLA-A, HLA-B, and HLA-DR genes. It turned out that the females preferred the odors of T-shirts worn by MHC-dissimilar males to those worn by MHC-similar males (Wedekind et al. 1995).

Second, a direct test of whether humans mate disassortatively for MHC should be possible by typing married couples. To date, such an analysis has been done on the Hutterites (a North American repro-

ductive isolate of European ancestry), South American Indians from the lower Amazon basin, and the Japanese. Only among the Hutterites has disassortative mating been demonstrated (Ober et al. 1997). In the other two populations, mating is apparently random with regard to MHC.

It is not clear why such contradictory results have emerged. One possibility is that, since the preference for MHC-dissimilar mates is weak, it may sometimes be overwhelmed by other biological or cultural factors. In particular, a preference for cousin marriages will mask any tendency towards disassortative mating. However, there is no evidence of cousin marriages among the couples sampled in the above mentioned studies.

3. Genetically Fixed Preferences vs. Acquired Preferences

In all of the above examples, mating preferences are apparently acquired by sexual imprinting. Despite the title of this article, convincing demonstrations of genetically fixed preferences are hard to find. Even in insects, appropriate social behavior is often contingent on learning to distinguish different categories of individuals—nestmates from intruders, for example.

Nevertheless, a possible example of a genetically fixed preference occurs in ladybird beetles. Ladybirds are polymorphic for color and markings, and females show a preference for melanic males. Majerus et al. (1982) subjected this preference to artificial selection by permitting only those females who mated with melanic males to lay eggs. When the experiment was continued for four generations, a substantial increase in the level of preference was observed.

Cultural transmission of preferences is also difficult to prove. Surprisingly, the best evidence comes not from humans, but rather from small tropical fish—guppies, sailfin mollies, and three-spined sticklebacks—where the phenomenon is known as ‘mate-choice copying’ (Dugatkin 1992). Mate-choice copying is said to occur when a female is more likely to choose a specific male after observing another female doing so. The following three-phase experiment is conducted to demonstrate mate-choice copying.

In phase one, a female (focal female) is given a choice of two males. If the focal female shows a consistent preference for one of these males, then in phase two, a second female (model female) is placed near the nonpreferred male in view of the focal female. Finally, in phase three, the model female is removed, and the focal female is again given a choice of the two males. Mate-choice copying is implied, if the focal female’s preference is reversed as a result of seeing the model female and the nonpreferred male together. (It is not clear how long the reversal of preference is maintained.)

The experiment must control for confounding effects. For example, the focal female might be attracted by the nonpreferred male's activity during or after his interaction with the model female, rather than by the presence of the model female. Or the focal female may simply be exhibiting a tendency to join other females, as is apparently the case with fallow deer. White and Galef (1999), who investigated mate-choice copying in Japanese quail, point out other complications in the interpretation of such experiments.

Given this emphasis on socially learned sexual preferences in fish and birds, it is perhaps ironic that most research on humans—by the evolutionary psychologists—has attempted to identify preferences that are relatively impervious to social influence. For example, it is claimed that men are attracted to women with a low waist-to-hip ratio (WHR)—an hourglass figure. This appears to be true in several Westernized cultures. Moreover, since women with lower WHR are apparently more fecund (physiologically able to give birth) (Zaadstra et al. 1993), Singh (1993) argues that it is adaptive for a man to have such a preference, whence it has evolved to become a panhuman—culturally invariant—male characteristic. The logic is fine, but the prediction has yet to be verified. Recently, Yu and Shepard (1998) studied WHR preferences in an isolated Peruvian population unexposed to Western media. Contrary to prediction, an overweight woman with a high WHR was most attractive and desirable as a spouse for the men of this population.

See also: Divorce, Sociology of; Evolutionary Selection, Levels of: Group versus Individual; Family, Anthropology of; Family Theory: Economics of Marriage and Divorce; Incest, Inbreeding, and their Consequences; Incest Prohibition, Origin and Evolution of; Marriage; Natural Selection; Partner Selection across Culture, Psychology of; Personality and Marriage; Primate Behavior: Significance for Understanding Humans

Bibliography

- Adams M S, Neel J V 1967 Children of incest. *Pediatrics* **40**: 55–62
- Bagnall R S, Frier B W 1994 *Demography of Roman Egypt*. Cambridge University Press, Cambridge, UK
- Bateson P 1982 Preferences for cousins in Japanese quail. *Nature* **295**: 236–7
- Bittles A H 1994 The role and significance of consanguinity as a demographic variable. *Population Development Review* **20**: 561–84
- Dugatkin L A 1992 Sexual selection and imitation: Females copy the mate choice of others. *American Naturalist* **139**: 1384–9
- Grant B R, Grant P R 1996 Cultural inheritance of song and its role in the evolution of Darwin's finches. *Evolution* **50**: 2471–87
- Hoogland J L 1992 Levels of inbreeding among prairie dogs. *American Naturalist* **139**: 591–602

- Keller L F, Arcese P 1998 No evidence for inbreeding avoidance in a natural population of song sparrows (*Melospiza melodia*). *American Naturalist* **152**: 380–92
- Majerus M E N, O'Donald P, Weir J 1982 Female mating preference is genetic. *Nature* **300**: 521–3
- Maynard Smith J 1980 Models of the evolution of altruism. *Theoretical Population Biology* **18**: 151–9
- Ober C, Weitkamp L R, Cox N, Dytch H, Kostyu D, Elias S 1997 HLA and mate choice in humans. *American Journal of Human Genetics* **61**: 497–504
- Pusey A E 1980 Inbreeding avoidance in chimpanzees. *Animal Behaviour* **28**: 543–52
- Ralls K, Ballou J D, Templeton A 1988 Estimates of lethal equivalents and the cost of inbreeding in mammals. *Conservation Biology* **2**: 185–93
- Singh D 1993 Adaptive significance of female physical attractiveness: Role of waist-to-hip ratio. *Journal of Personality and Social Psychology* **65**: 293–307
- Thornhill N W (ed.) 1993 *Natural History of Inbreeding and Outbreeding*. University of Chicago Press, Chicago, IL
- Wedekind C, Seebeck T, Bettens F, Paepke A J 1995 MHC-dependent mate preferences in humans. *Proceedings of the Royal Society of London Series B* **260**: 245–9
- Westermarck E 1891 *History of Human Marriage*. Macmillan, London
- White D J, Galef B G 1999 Mate choice copying and conspecific cueing in Japanese quail, *Coturnix coturnix japonica*. *Animal Behaviour* **57**: 465–73
- Wolf A P 1995 *Sexual Attraction and Childhood Association*. Stanford University Press, Stanford, CA
- Yamazaki K, Boyse E A, Mike V, Thaler H T, Mathieson B J, Abbott J, Boyse J, Zayas Z A, Thomas L 1976 Control of mating preferences in mice by genes in the major histocompatibility complex. *Journal of Experimental Medicine* **144**: 1324–35
- Yu D W, Shepard G H 1998 Is beauty in the eye of the beholder? *Nature* **396**: 321–2
- Zaadstra B M, Seidell J C, Van Noord P A H, te Velde E R, Habbema J D K, Vrieswijk B, Karbaat J 1993 Fat and female fecundity: Prospective study of effect of body fat distribution on conception rates. *British Medical Journal* **306**: 484–7

K. Aoki

Genetics and the Media

Genetics is in the news as the scientific frontier, the cutting edge of current research. Every new discovery is reported in the media with optimistic expectations as a 'breakthrough'—even when the significance of a particular finding is in question. A survey of the coverage of genetics in the American press from 1970 to 1996 found that reports were overwhelmingly positive as journalists equated research with economic potential and medical progress (Lewenstein et al. 1998). Beyond the news, the gene has also become a prevalent media metaphor appearing in fictional narratives, popular magazines, visual images, and the cultural detritus of sound bites, slogans, and snippets: it's all in the genes. These media attribute extra-

ordinary powers to the gene, conveying a concept of 'genetic essentialism' (Nelkin and Lindee 1995). DNA in media narratives is an essentialist entity—the cause of disease, the definition of personhood, the location of the true self.

DNA is data without dimension, text without context, so that the equation of DNA with personhood requires a profound leap of faith. Yet, the gene as a solid source of identity holds media appeal. So too does the belief that the physical and behavioral traits of individuals, and the social and economic differences among groups are caused by inborn distinctions.

Genetic language has become a code in the media, readily applied and broadly understood. What is genetic is powerful, predictable, and permanent; 'hard-wired' in the human constitution. Underlying the notion of genetic destiny is the old idea of biological determinism: people are the way they are because they are born that way. This article reviews several themes conveyed by media gene talk:

- (a) Genetic technologies provide the means to predict health and disease.
- (b) Genes determine behavior and personality traits.
- (c) Genetic research holds the promises of therapeutic solutions.
- (d) But there are also potentials for abuse.

Following a brief description of these media themes, this article offers some reasons for the appeal of genetic stories. It suggests how the media coverage of genetics conforms to popular stereotypes and prevailing ideologies, and reflects the rhetorical skills of scientists who are eager to promote their work in the public arena. The article concludes by suggesting some of the problems with the media message that health, behavior, personhood—indeed, life itself—can be defined in genetic terms.

1. Media Themes

1.1 Predicting Genetic Destiny

Genetics is a science of prediction. Scientists emphasize the predictive possibilities of genetics as they seek to convey the importance of their science as a way to identify predispositions. Using media-appealing terms, geneticists call the gene a 'Delphic Oracle,' a 'time machine,' a 'trip into the future,' a 'medical crystal ball.' James Watson, the first director of the Human Genome Project, has announced in frequent media interviews that 'our fate is in our genes.'

Such metaphors convey several messages: a definition of the gene as the basis of human destiny, an expectation that genetic research will enable the prediction of future disease, and a promise of therapeutic solutions. These messages are disseminated by a receptive press.

The media, for example, cheered the discovery in the mid-1990s of the *BRCA1* and *BRCA2* genes that

helped to explain many cases of familial breast cancer particularly among Jewish women of Ashkenazi descent. News headlines announced that scientists had found the gene causing breast cancer, some referring to the 'breast cancer gene.' They suggested that heredity is the major cause of breast cancer and that those found to have the gene were certain to get the disease. In fact, genetic factors contribute to only about 5 percent of all breast cancer cases, and even among those with a *BRCA* gene, modifying factors, including diet and genetic variables, may limit the predictive power of the genetic diagnosis. But widespread publicity, according to former Director of the National Institutes of Health, Bernadine Healy, created an ambience of 'medical bookmaking and fortune telling' that was devastating for women who believe they are at risk (Healy 1997).

Joseph Alper describes the media coverage of genetics as OGD (One Gene One Disease) reporting (Alper 1996). When journalists report on the discovery of disease genes, they tend to minimize the complex, multicausal, and poorly understood sources of the genetic mutations that lead to disease. They are responding to the public desire for simple causal explanations. But Steven Rose (1995) describes genetic predictions as faulty reductionism based on 'reification, arbitrary agglomeration, improper quantification, misplaced causality, and dichotomous partition between genetic and environmental causes.' Similarly, Philip Kitcher (1997) describes how journalists tend to leap from an association of genes with particular conditions or traits to the attribution of genetic causation. Correlations turn into causes in the press.

Gene discoveries are newsworthy, but qualifications and disconfirmations are ignored. Studying the media reports about the causes of mental illness, Peter Conrad (1999) finds only limited coverage of inconsistent and contradictory information. The media message is rather that genes are destiny and the gene for mental illnesses will eventually be found. Science, however, is provisional and scientists routinely disconfirm their claims. These disconfirmations, like failed experiments, are important to the methods of science, but they are not regarded as news.

1.2 Genes Determine Behavior

Encouraged by the enthusiastic media response to research in medical genetics, behavioral psychologists in the 1990s began to extend the notion of genetic determinism to the inheritance not just of diseases, but of complex behavioral conditions and personality traits.

Such complex attributes as shyness, obesity, directional ability, aggressive personality, caring tendencies, exhibitionism, homosexuality, dyslexia, job success, arson, traditionalism, preferred styles of dressing, tendencies to tease, political leanings, re-

ligiosity, criminality, intelligence social potency, and zest for life have all appeared in media accounts as if they are simple Mendelian disorders, directly inherited like brown hair or blue eyes. Behavioral characteristics, once viewed in psychological or sociological terms, are now framed through genetic lenses (Alper and Beckwith 1993).

Seeking provocative and controversial copy, journalists have focused attention on claims attributing antisocial behavior to biological predisposition (Conrad 1997). There is violence, homelessness, alcoholism because people have genes that predispose them to aggression, depression, or alcoholism. Journalists refer to 'bad seeds,' criminal genes,' and 'alcohol genes.' To a *New York Times* writer, 'evil is embedded in the coils of chromosomes that our parents pass to us at conception' (Franklin 1989).

Journalists are also attracted to controversial theories about genetics and intelligence. In the 1970s, the dispute over Arthur Jensen's claims about the relationship between race and IQ attracted considerable media attention. Two decades later, the similar claims by Richard Herrnstein and Charles Murray in *The Bell Curve* again attracted an attentive press. Media interpretations of these studies, however, were more guarded than most reports on genetics.

Snyderman and Rothman (1988) saw the media as strongly biased against genetic explanations of intelligence, favoring arguments based on nurture or nature. Yet the coverage of the genetics and IQ debates has been extensive and the repeated articles about books like *The Bell Curve* disseminated their ideas very widely and ultimately served to amplify their importance.

In the 1990s, the media disseminated the claims about the genetic basis of behavior that were generated by studies of identical twins reared apart. These were controversial studies that often appeared in the media well before they were published in peer-reviewed journals. Offering a simple explanation of complex behavior and a reinforcement of prevailing stereotypes, this research attributed an extraordinary range of behaviors to 'the genes.' *US News and World Report* (April 13, 1987) published an authoritative-looking table providing precise percentages of how much personality traits were determined by heredity rather than culture: extroversion 61 percent, conformity 60 percent, worry 55 percent, creativity 55 percent, optimism 54 percent, and so on.

In the same issue, *US News* published an article called 'How Genes Shape Personality,' claiming that 'solid evidence demonstrates that our very character is molded by heredity.' It suggested that the future of Baby M, the child in a controversial surrogacy dispute, may not rest on which family got her, but in her genes. In 1992, *Time* was still offering the time-worn explanation of sex differences: 'nature is more important than nurture' and it is just a matter of time until scientists will prove it (January 20, 1992).

Media references to genetic influences draw on research in molecular and behavioral genetics, but they minimize the complexities of genetic and environmental interactions. They offer a distorting lens by ignoring the distance between the molecular level of genetic systems and actual behavior (Wilkins 1993). Beyond the curious notion that every quality has a gene, these media often convey a fatalistic and almost resigned attitude, for genetic predispositions appear to be immutable (Alper and Beckwith 1993).

One quantitative evaluation of magazine articles on genetics from 1919 to 1995 found a decreasing emphasis on genetic determinism over time (Condit et al. 1998). However, the way genetics is reported in the media of the 1990s suggests a retreat from the myth of 'can do' optimism that once led to the notion that every child could be president. The old story about those who make it despite difficult social circumstances is no longer a mystery. Genetic predisposition implies genetic destiny.

1.3 Biological Solutions

Good news, magic bullets, solutions, and cures appeal to public hopes and expectations. Journalists are strikingly optimistic as they cover the promise of 'medical breakthroughs' and 'wonder therapies.' Genetics claimed a journalist is 'the medical story of the century which will dramatically cure cancer, heart disease, aging and much more' (Rosenfeld 1992).

The media have welcomed every new claim about gene therapy with extravagant headlines and promotional hype as reporters convey—often uncritically—a wild array of futuristic scenarios presented by enthusiastic scientists. In the future, said a geneticist to *Discover*, 'present methods of treating depression will seem as crude as former pneumonia treatments seem now' (Wingerson 1982). In the future, said geneticist French Anderson to a *Time* reporter, 'Physicians will simply treat patients by injecting a snippet of DNA and send them home cured' (*Time* January 17, 1994). News headlines have promised: 'Genetic research leaves Doctors hopeful for Cures,' 'New Hope for Victims of Disease.' In a story called 'the Age of Genes,' *US News* reported that 'advances bring closer the day when parents can endow children not only with health but also with genes for height, good balance, or lofty intelligence' (Brownlee and Silberner 1991).

The development of pharmaceutical products and the proliferation of clinical trials on new therapeutic procedures have encouraged genetic optimism. The first official gene therapy experiment in 1990—the injection of cells containing ADA genes in a child with an immune system malfunction—became a major news event, touted as the arrival of 'the long awaited era of genetic therapy' (Culver 1991). The isolation of

the colon cancer gene in 1993 prompted an enthusiastic scientist to tell a *New York Times* reporter of its implications: 'deaths are entirely preventable' (*New York Times* December 3, 1993). The media reports implied that gene therapy would allow doctors to 'do something' about devastating disease. But in fact, diagnostic capabilities have far outreached the possibilities of solutions.

The gap between the promise and the reality of gene therapy has led to increased ambivalence in media reports of genetics. While few questions are raised about the ultimate benefit of research or the credibility of scientific claims, reporters are expressing concerns about the growing availability of genetic information and its potential abuse.

1.4 The Potential for Abuse

Stories extol the benefits of research, but then decry the risk of gathering genetic information. We are told of the dawn of a new genetic era but then cautioned about an impending eugenic nightmare.

Ambivalence is especially apparent in the coverage of genetic manipulation. Jon Turney (1998) notes the continuous use of Frankenstein metaphors in the media. An illustration for a *New York Times* article (September 16, 1990) on gene therapy and the potential of genetic engineering featured a drawing imitative of the famous Edvard Munch painting, 'The Scream.' A figure stands, horrified, mouth ajar, eyes wide open, its hair a mass of coiled DNA. Media reports have often associated experiments in gene therapy with genetic engineering or 'tampering' with genes. 'Lurking behind every genetic dream come true is a possible Brave New World nightmare' says a *Time* reporter. 'To unlock the secrets hidden in the chromosomes is to open up the question of who should play God with man's genes.' An accompanying image portrayed scientists balancing on a tightrope of coiled DNA (Dewitt 1989).

An important feature of these ambivalent media reports on genetics is their concern about the social and economic context in which this field is developing. The ties between the science of genetics and its commercial applications invite cynicism. Reporters call attention to the nonscientific interests—the investments and profits in this intensely competitive field. Science journalists, many of whom were once scientists themselves, have maintained an image of science as a pure and unsullied profession, a neutral source of authority, and an objective judge of truth. They are skeptical of corporate driven science. These days, according to some disillusioned journalists, biology has 'lost its virginity.' Scientists working in genetics are 'greedy entrepreneurs,' 'gene merchants,' or 'molecular millionaires' driven by economic interests that threaten their objectivity and override concerns about abuse (Nelkin 1995). Political cartoons portray genet-

icists in less than flattering terms as bumbling, naive, and unaware of the social implications of their discoveries. News reports and editorials repeatedly call attention to troubling aspects of the growing links between science and industry, the conflicts of interest that are inevitable when profits and ethics collide (Krimsky 1991). While the media have retained their history of genetic optimism, they are also expressing concern that expanding commercial interests will override important social considerations.

2. The Media Appeal of Genetic Explanations?

The Human Genome Initiative has replaced the space program as a new frontier, the cutting edge of high technology exploration, and, just as in the heyday of NASA, journalists are receptive to the enthusiasm of their scientific sources.

The stories of the 'gene of the week' meet the immediate pragmatic needs of the press for sound bites, newspegs, and events that will attract readers. The notions of genetic determinism offer relatively simple causal explanations of complex behavior—a welcome change from messy and uncertain social and environmental speculations. Genetic explanations conform to prominent cultural biases, intersecting with important American social values. Locating social problems within the individual, they support the ideology of individual responsibility. They are consistent with faith in scientific and technological progress. The status of the gene as a deterministic agent, a blueprint for life, offers certainty, predictability, and control. By promoting genetic explanations of behavior, however, the media are using the imprimatur of science to support a particular world view. They do so with little attention to the substance of science and the limits of theories purporting simple explanation of complex human conditions.

The appeal of genetics to the media largely reflects the promotional efforts of scientists themselves, for journalists rely on scientific sources for information. Scientists seek media coverage as a form of public relations, regarding public visibility as a means to attract funds for their research and support for their priorities (Shinn and Whitley 1985). They have become skilled in rhetorical strategies designed to attract media attention. Using media-ready metaphors, they have described the genes as 'master molecules': we are but 'readouts' of our genes. They describe the body in deterministic terms—as a set of instructions, a blueprint, a map, or a program that is transmitted from one generation to another. They suggest that by deciphering the text, classifying the markers on the map, and reading the instructions, they will unlock the key to human ailments and human nature, revealing the secrets of human life. Geneticist Walter Gilbert, introduces his public lectures on gene sequencing by pulling a compact disk from his pocket and announc-

ing to his audience: 'this is you.' Scientists, he claims, will 'provide ultimate answers to the commandment, know thyself' (Gilbert 1990).

The biotechnology industry has further encouraged media hype about gene therapy. The industry has made a major financial commitment to gene therapy, expecting this will be the basis of future medicine. Over 60 percent of gene therapy studies are directly financed by industry in anticipation of a profitable market in the near future. Corporate advertisements announce 'a great leap in the treatment of disease,' and promise 'a healthy future one gene at a time.' Corporate and university press releases are inclined to turn tentative experimental findings into magic bullets. An advertisement appearing in sports magazines said it all: 'Bad Genetics? Use Opti-genetics—The first genetic optimizer.'

When they report on complex scientific issues, journalists do not create science news; they mainly amplify and disseminate the messages provided by scientists themselves. They are seldom critical, responding with untempered enthusiasm to promotional hype especially when packaged in accessible media-friendly terms. However, the surfacing of hereditarian beliefs, their media appeal must be examined critically, for such beliefs have practical implications for both individual choices and social policies.

3. The Growing Power of the Gene

Tracing the history of popular images of genetics, Van Dijk (1998) has documented the changing metaphors that correspond to various stages in the development of this scientific field from the new biology of the 1970s to the rise of the biotechnology business and the Human Genome Project in the 1990s. At each stage, she says, 'performers'—scientists, journalists, and public relations people shape the media discourse (Van Dijk 1998).

Nelkin (1995) found considerable media skepticism in the 1970s when journalists wrote about the 'runaway science of genetic engineering.' But by the 1980s, gene splicing became a 'race toward better human health.' Today, genetics is the 'technological frontier.' Journalists describe geneticists as pioneers, 'unlocking the basic laws of nature,' discovering the 'secrets of life,' solving the problems of devastating disease. Genomic researchers are 'riding the DNA trail.' Geneticists are 'relentless hunters of genes,' involved in a 'race' to find the markers for disease. Media accounts in the 1990s have been reverent—almost religious: a 1994 cover of *Time* depicts a figure on a pedestal, his arms extended in a Christlike pose, his torso inscribed with a double helix. The caption reads: 'Genetics—The Future is Now.' The image, of course, is the Ascension. Though simply a biological entity, the gene in the media has assumed cultural significance extending far beyond its scientific meaning.

The media are mainly promotional, yet they are increasingly conveying mixed messages about the costs and benefits of biotechnology. Journalists seem to welcome the notion of biological determinism—simple, seemingly certain, and easy to convey—yet many writers remind their readers of the history of eugenics and place current research in this threatening historical context. While journalists report with enthusiasm and wonder the promises of gene therapy, they also warn about potential abuses of genetic manipulation.

Why should we be concerned about media images? Common and recurring media images provide insight into beliefs that are deeply embedded in society—beliefs that shape the questions of concern to scientists and the public appropriation of scientific findings. The media portrayal of the powers of the genes can have a powerful influence. For science has long served as a cultural resource appropriated and interpreted to support social beliefs and to legitimate social policies (Nelkin 1999). The language of determinism not only reveals but also creates social attitudes. The sound bites and headlines (gay genes, obesity genes, aggressive genes) attract readers, but they also endorse certain modes of thought and explanations.

Media coverage of science and technology provides a useful window on public attitudes, but also has an important influence on public perceptions. The way people perceive research in human genetics—the way they interpret their costs and benefits—may be influenced less by the details of scientific evidence than by the repeated messages conveyed in the popular press. These media messages help to create the beliefs and assumptions that underlie personal decisions, social policies, and institutional practices.

See also: Science and the Media

Bibliography

- Alper J S 1996 Genetic complexity in single gene diseases. *British Medical Journal* **312**: 196–7
- Alper J S, Beckwith J 1993 Genetic fatalism and social policy. *Yale Journal of Biology and Medicine* **66**: 511–24
- Brownlee S, Silberman J 1991 The age of genes. *US News and World Report* **November 4**: 64
- Condit C M, Ofulue N, Sheedy K M 1998 Determinism and mass media portrayals of genetics. *American Journal of Human Genetics* **62**: 979–84
- Conrad P 1997 Public eyes and private genes: historical frames, news constructions and social problems. *Social Problems* **44**: 139–54
- Conrad P 1999 Media images, genetics and culture. In: Hoffmaster B (ed.) *Bio-Ethics in Context*. Cambridge University Press, Cambridge, UK
- Culver K W 1991 Splice of life. *The Sciences* **33**(1): 18–24
- Dewitt P E 1989 The perils of treading on heredity. *Time* **March 20**: 70
- Franklin D 1989 What a child is given. *New York Times Magazine* **September 3**: 36

- Gilbert W 1990 *Current State of the HGI*. Dibner Institute, Cambridge, MA
- Healy B 1997 BRCA genes—bookmaking, fortunetelling, and medical care. *New England Journal of Medicine* 336: 1448–9
- Kitcher P 1997 *The Lives to Come*. Simon & Schuster, New York
- Krimsky S 1991 *Biotechnics and Society*. Praeger, New York
- Lewenstein B U, Allaman T, Parthasarathy S 1998 *Historical Survey of Media Coverage of Biotechnology in the United States*. Presented to the Association for Education in Journalism and Mass Communication, Baltimore, MD
- Nelkin D 1995 *Selling Science: How the Press Covers Science and Technology*, rev. edn. Freeman, New York
- Nelkin D 1999 Behavioral genetics and dismantling the welfare state. In: Rothstein M (ed.) *Biology and Culture*. Johns Hopkins University Press, Baltimore, MD
- Nelkin D, Linde M S 1995 *The DNA Mystique: The Gene as a Cultural Icon*. Freeman, New York
- Rose S 1995 The rise of neurogenetic determinism. *Nature* 373: 380–2
- Rosenfeld A 1992 The medical story of the century. *Longevity* May: 42–53
- Shinn T, Whitley R (eds.) 1985 *Expository Science*. Reidel, Dordrecht, The Netherlands
- Snyderman M, Rothman S 1988 *The IQ Controversy*. Transaction Books, New Brunswick, NJ
- Turney J 1998 *Frankenstein's Footsteps: Science, Genetics, and Popular Culture*. Yale University Press, New Haven, CT
- Van Dijk J 1998 *Imagination: Popular Images of Genetics*. MacMillan, London
- Wilkins A S 1993 Jurassic Park and the 'Gay Gene': the new genetics seen through the distorting lens of the media. *The FASEB Journal* 7(13): 1203–4
- Wingerson L 1982 Searching for depression genes. *Discover* February: 60–4

D. Nelkin

Genetics in Anthropology

Anthropologists of biological orientation began to undertake the study of genetics as soon as it became apparent about 1900 that biological inheritance was indeed particulate, and that the supposed essence of inheritance, genotype, could be inferred from phenotype. Within 'the science of man' theories were soon developed which applied these insights to the solution of problems that had long been central to anthropology, many of which were deeply entangled with political and religious controversies concerning 'race' and evolution. These problems especially included: (a) the extent and significance of human variation, (b) the phylogenetic relationships among humans and other primate species, (c) the migration and adaptation of human populations to different continents and geographical regions since the earliest times, and (d) the relationships between biological variation and other kinds of human variation, especially in language and culture. Despite some episodes of resistance to this new source of information, by the end of the twentieth

century it was clear that genetics had become central to the science of anthropology, not only for helping to illuminate these problems, but for discovering new problems and creating new fields of study within the discipline (Crawford 2000).

1. Human Variation

When laboratory techniques for studying Mendelian traits first became available, international anthropology had just emerged from a serious controversy over racial typology during which Franz Boas, recognized by many as the founding father of American anthropology, had repudiated to the satisfaction of most anthropologists the racial theories of Gobineau and his intellectual heirs. Originally trained as a physicist, Boas presented data and arguments between 1910 and 1913 showing that 'race,' 'language,' and 'culture' were three independent phenomena which, he asserted, should all be studied by anthropologists, but using different kinds of theories and techniques. On the heels of this observation, the Hirtzfelds, building on the work of Landsteiner and others, showed in 1919 that while the ABO blood system was universal among human populations, the phenotypes and hence the alleles, the alternative forms of a gene, were in different proportions, thereby supporting Boas' contention that the differences among 'racial' populations were not qualitative but quantitative (Weiss 1998). The ability to taste PTC (phenylthiocarbamide), a simple Mendelian trait, also seemed to occur in many different human populations, likewise in different frequencies. But other apparently Mendelian traits examined early in the century—eye color, skin color, and ear lobe shape—turned out to embody genetic relationships with a much more complex genetic architecture (McKusick's current list of over 30,000 Mendelian traits can be found at World Wide Web URL: <http://www3.ncbi.nlm.nih.gov/Omim/>).

Data on human phenotypes assembled for military purposes in World War I continued to encourage anthropological geneticists in the decades following the war. In addition to complex and composite anthropometric traits such as height and weight, the military was especially interested in those factors that affected the compatibility of blood and blood components used in transfusion. The collection of such data was continued through World War II, when most European and American soldiers and sailors were typed for ABO and Rh factors during initial processing, and this body of data became generally available to scientists after 1945 (Harrison et al. 1964).

For anthropologists, the major theoretical model which emerged to describe human genetic variation in this period was the notion of cline, the gradual geographic change in the frequency of an allele, which also served to undermine further the idea of racial typology. As regularities began to emerge, especially

in the European data, it was found that the alleles of a gene, and also phenotypic traits such as red hair, seemed to vary across the landscape in a regular manner. These were mapped using arrows to show the direction of increasing frequency of one of the component alleles of a gene, or by isographs which identified geographical areas where allele frequencies were within a particular range. Traits mapped in this manner included blood system alleles, phenotypic traits, and also diseases which were suspected of having a simple genetic basis, such as albinism, color blindness, and some types of cancer (Harrison et al. 1964, pp. 234–76).

Some patterns which emerged from the analysis of clines conformed to more general biological principles pertaining to other animals. It was found, for example, that Gloger's, Bergmann's and Allen's Rules for variation within a mammalian species (pertaining to pigmentation, mass, and limb length, respectively) were also valid for the human species, reflecting adaptation to the different latitudes, climates, and environments inhabited by local human populations. Since these traits were so clearly adaptive, it was suspected that other patterned traits, with definite clines, likewise reflected genetic adaptation to local conditions.

2. Adaptation

The most forceful new example of human genetic adaptation was presented by Livingstone in 1958 with his study of sickle cell disease, especially as it existed in Africa. In addition to presenting convincing evidence of several types to prove that sickled blood cells were a response to the presence of malaria, Livingstone's work illustrated several other lessons in anthropological genetics. First of all, it was the heterozygote of hemoglobin S (HbS) which had the adaptive advantage, not the HbS homozygote, who often sickened and died. The second lesson was that an apparent 'disease,' sickle cell anemia, was on closer inspection actually a successful adaptive mechanism, a proposition which is still better understood by anthropologists and population geneticists than by their colleagues in clinical medicine (Weiss 1993).

Another highly visible connection between genetic variability, culture, and environment concerns the ability of human adults to digest milk and milk products from dairy animals, an ability that seems to be under genetic control. When mapped geographically, the alleles responsible for adult lactose tolerance in human populations seem to be correlated, in the Old World, with the locations of pastoral peoples. Significantly, populations where lactose tolerance is high include pastoral societies all the way from Scandinavia to southern Africa, producing east–west clines which cut across the north–south clines describing the distributions of all sorts of other phenotypic

characteristics, including those related to latitude and climate (Durham 1991).

There are indications that the frequencies of many other diseases, such as cancer and diabetes, correlate with blood types and also with alleles involved in the digestion of food. People of A phenotype, for example (in the ABO system), may carry an increased risk of death from cancers of the esophagus, pancreas, and the stomach, but whether these diseases are in any sense 'caused' by type A blood, as the sickling trait can cause anemia, or whether both disease and blood type are derived from an independently variable genetic or environmental condition, is very problematic to clarify. Northern Scandinavia, for example, has a high frequency of type A individuals and a high incidence of stomach cancer. But their diet is different from other Europeans, and their climatic environment is different as well, so it is difficult to separate the variables which might cause stomach cancer (Molnar 1998, pp. 100–7).

Perhaps the most useful way of looking at the various genetic and environmental factors correlating with disease frequencies is to say that each 'contributes' to the likelihood that an individual will exhibit the disease (Terwilliger and Göring 2000). Many anthropologists argue that from the standpoint of adaptation, no gene of itself is a 'bad' gene. Like HbS, any allele manifested in the phenotype might be beneficial in certain circumstances. Among anthropologists, it is increasingly recognized that the relationship of gene to environment, like the relation of gene to gene, is very complex. Also, the distinction between 'genetic disease' and 'infectious disease' begins to break down when we consider that smallpox, for example, takes a more virulent course in people of blood type A or AB, as opposed to O or B (Stein and Rowe 2000, pp. 427–30).

The study of adaptation to cold climate and high altitude has provided an opportunity to differentiate between the effects of genetic differences among human populations and phenotypical differences engendered by growth and development—more simply, the difference between nature and nurture. Most recently, Beall has shown that Tibetan and Peruvian populations have a different genetic basis for their respective adaptations to thin air (hypoxia), but many other aspects of human adaptation to climatic extremes remain to be examined (Jones et al. 1992, pp. 46–51). These studies require the comparison of highland to lowland Peruvian Indians, and the examination of such characteristics as body type, especially lung capacity, and oxygen saturation rates among migrants and colonists in both locations.

3. Population Genetics

With increased funding for science in the United States after World War II, and increased access to field sites,

anthropologists began to investigate isolated, small-scale human populations with renewed vigor, often including a biological anthropologist in the research team. The most visible of these joint enterprises was probably the comprehensive study of the Yanomamö, who live in the upper drainage of the Orinoco River in Venezuela and Brazil. In the 1960s, James Neel of the University of Michigan organized a comprehensive multidisciplinary series of field expeditions which investigated the environment, ecology, biology, culture, and social organization of these then little-known people. After a series of books and articles was published over the next two decades, the Yanomamö became not the least known but one of the best known tribal societies in the world.

For the genetics part of the research, Neel and Ward examined six blood systems relatively well known at the time (MNS, Rh, Kidd, Duffy, Diego, and haptoglobin) among seven villages of Yanomamö, seven villages of nearby Makiritare, and 12 other tribes of Central and South America (Neel and Ward 1970). After comparing the different groups, the authors concluded that human evolution might have been proceeding at a pace '100 times more rapid' than previously thought, if these small-scale societies could be taken as analogues of earlier Pleistocene or Paleolithic human peoples.

In the same period, Friedlaender and Harpending were proceeding with similar investigations in Oceania and Africa, respectively. Working in the Solomon Islands from 1966 to 1973, sometimes alone and sometimes as part of a multidisciplinary team, Friedlaender simultaneously undertook studies of demography, anthropometry, genetics, and social structure, seeking to integrate his results into some general conclusions that cut across these traditional fields of research. Like Neel and Ward, he emphasized the genetics of blood systems, in this case ABO, MNS, Rh, Kell, and Duffy. In these years, blood system alleles were not sequenced directly, but determined from immunological reactions to blood proteins, usually through the intermediary use of rabbit antibodies (Jones et al. 1992, p. 304). From this kind of data, Friedlaender calculated phenotype frequency and then allele frequency by techniques described in his major ethnographic work, *Patterns of Human Variation* (Friedlaender 1975). His most general conclusion, combining results from anthropometry, dermatoglyphics, genetics, and demography, was simply that Bougainville communities tended to exhibit high biological variability within the group, but low variability among the groups, and that some traits, such as skin color, eye color, and hair form, were 'monotonous' on the island, while dermatoglyphic characteristics and some blood system alleles were discontinuous from village to village. Those expecting that these early studies would provide fundamental insights into the mechanisms of human evolution were soon disappointed. Summariz-

ing a widely shared consensus of opinion in 1974, Harpending concluded (1974, p. 229):

... studies of the genetic structure of small populations have made particular and incidental contributions to formal genetics, to regional history and prehistory, to epidemiology, and to several other fields to which they are peripheral, but ... they have not advanced our understanding of human evolution in a global sense. The sample sizes available have been too small to allow reliable inferences about natural selection; the extensive occurrence of what is presumably local random genetic drift has little or no consequence for evolution over long time periods over large areas; and the presumed selective agents in the various environments of these peoples differ greatly so that few of the generalizations which have been put forward hold for many groups.

4. *Evolution of Primates*

Although anthropologists, even before the publication of Darwin's *Descent of Man* had recognized that humans and apes were closely related, the significance of the differences between them had been both exaggerated and misunderstood. Based on the fossil record, the split of humans from the ape line had been put far back in time, about nine million years, and it was assumed that all the ape species were closely related to each other and quite distinct from humans. Therefore the allegation of Sarich and Wilson, published in 1967 and based primarily on immunological data, that humans and apes had diverged as recently as five million years ago (mya), created a serious confrontation between 'the bone people' (paleontologists) and 'the lab people,' human geneticists such as Sarich and Wilson. The former accused the latter of inaccuracies in their assumptions about the speed of 'evolution,' in this case the evolution of blood albumin, while the latter accused the former of manipulating primate taxa to preserve the supposed profound differences between ape and human. In fact the fossil *Ramapithecus*, with an antiquity of about 14mya, soon fell out of the human lineage on osteological grounds, so that even the fossil evidence seemed to point to a more recent divergence between human and apes.

Beginning in the 1970s, Morris Goodman and his associates have become prominent in the reconciliation of fossil and biological evidence which, taken together, provide the best description of the evolution of apes and humans. In 1974 Goodman proposed that humans, chimpanzees, and gorillas should join humans as hominines, a subfamily of the family Hominidae, perhaps leaving the more distantly related gibbons and orangutans as separate families or subfamilies. Goodman's conclusions, like those of Sarich and Wilson, were based on molecular data, in this case amino acid sequences from blood proteins.

Goodman's current synthesis (1999) includes data from many kinds of direct sequencing of DNA from

humans and other primates, especially the β -globin gene cluster, and from the analysis of morphological features, especially certain diagnostic skeletal features on living and fossil primates. His current picture of relationships among living humans and apes shows a split between the Tribe Hominini and the lesser apes about 18mya, between orangutans and the Subtribe Hominina about 14mya, and among gorillas, humans, and chimpanzees within the last seven million years. The timing and significance of the chimpanzee–bonobo separation has not yet been firmly established, since they have only recently been recognized as distinct species, and basic comparative genetic studies of both groups are now underway.

5. Methods

Although thousands of laboratory procedures have been developed over the last half-century, especially by medical researchers, to elucidate the structure of the human genome and show how it reproduces itself, certain techniques and certain targets have been more interesting to anthropologists because of their particular research agenda. Ideally, anthropologists interested in adaptation would like to know the locations of all human genes on the chromosomes, the number of alleles for each gene, and the locations of all loci involved in constructing a protein, but in practice researchers have had to settle for much less. Early in the twentieth century, the identity of alleles had to be teased out of complex immunological procedures in the laboratory. Later, proteins could be examined directly, but without much knowledge of the DNA structures, the precise sequence of nucleotides which had produced the protein. But with the development of slow though reliable DNA sequencing procedures in the 1970s, anthropologists became much more selective about which portions of the chromosomes they wanted to study, and for what purposes.

Surprisingly, with the inception of direct sequencing, it was discovered that perhaps 93–97 percent of the mass and length of human chromosomes consisted of apparently inactive regions variously called introns, satellites, tandem repeats, noncoding regions, or even ‘junk.’ Almost immediately these regions began to attract the attention of anthropologists even though they did not contain genes. It was the presumed inactivity of these regions which was interesting, since accumulated mutations on these regions were apparently selectively neutral, that is neither favored nor disfavored by natural selection. Therefore, it was reasoned, they might in a more accurate, clock-like manner reflect not only the relationships among primate species, but also the migrations and gene flow of ancient human populations. That is, satellite DNA was considered to constitute a ‘passive’ record of evolutionary events, rather than comprising active agents under selective pressure.

5.1 Stains

As soon as chromosomes could be observed directly through microscopes, attempts were made to stain them chemically so that bands would appear which might show differences in structure. Initially, the purpose was to discriminate among the different chromosomes of the human karyotype, or suite of chromosomes (22 autosomes, 1X and 1Y), so they could be counted and numbered. During the course of this research, it also became apparent that patterns of stains on a particular chromosome might vary from one individual to another within a population. In addition, when it was noted that certain patterns of staining were characteristic of particular species, yet another method became available for evaluating the phylogenetic relationships among apes and humans. At first the functional significance of the patterns was unknown, but this is changing rapidly as new techniques allow the identification of gene-rich regions identified by stains.

5.2 Restriction Enzymes

In 1973 it was discovered that certain enzymes, mostly recovered from bacteria, could sever a strand of DNA selectively at locations defined by a known base pair sequence. For example, the enzyme *AluI* recognizes the sequence AGCT and cleaves DNA at that location to produce fragments which are on average 300 base pairs in length in humans. The resulting distribution of segments of different length serves to define the structure of a chromosome, even without knowing the entire succession of base pair sequences on each segment. An appropriate enzyme, or several enzymes applied in series, can be selected to produce a small or a large number of segments, long or short, depending on the purposes of the research. By comparing the phenotypes of individuals contributing DNA samples and examining their pedigrees, ‘linkages’ could then be discovered between particular alleles included in the same segment.

Electrophoresis, which had previously been used to separate blood proteins obtained from serum, can be used to separate DNA segments according to size and net electric charge, with the segments sometimes enhanced or differentiated visually by using a large number of new stains developed in the last two decades of the twentieth century, some of which are fluorescent. In the 1980s, the PCR (polymerase chain reaction) ‘revolution’ occurred, which allowed researchers to clone or amplify small pieces of DNA indefinitely. PCR procedures make it possible to amplify a tiny bit of DNA clipped from an electrophoresis gel, or even a single sequence from a single cell, to obtain sufficient material for analysis (Hillis et al. 1996). All of these became useful tools for describing genetic variation in a population within a species, or to compare and contrast variations between and among species.

5.3 Chromosome Mapping

A major objective of human genetics research is to locate all genes on their respective chromosomes as well as to identify all the other functional regions of DNA, including some formerly dismissed as 'junk.' One technique for gene mapping involves the use of genetic markers, like the restriction enzyme sites described earlier, or other DNA sequence variants discovered by various means at hundreds of other chromosome locations. Because many of these are relevant to human disease, high levels of funding were made available to the Human Genome Project to identify the location of these variable sequences or 'polymorphisms.' This support has led to the development of very fast computer and molecular technology and machinery, and much of the resultant data appears in easily searched form on the worldwide web.

5.4 Mitochondrial DNA (mtDNA)

Most human genes reside in the nucleus of the cell and occur in two copies, one each inherited from the mother and the father. However, the body or cytoplasm of each cell also contains structures known as mitochondria, which are involved in the production of energy for the cell and contain their own DNA. This DNA is special in many ways. At fertilization, the mitochondria of the ovum are transmitted to the offspring, but the mitochondria of the sperm are not. Thus mtDNA is inherited in a strictly maternal manner, a biological phenomenon which parallels several important sociological phenomena of special interest to anthropologists, such as matrilineal clan membership.

Since mtDNA is haploid rather than diploid in form (a single rather than double strand), it does not recombine during replication, and over time, as sequence variations accumulate through mutation, it produces distinct, diverging, lineages of mtDNA sequences. This variation can be used to reconstruct the mtDNA structure of the common, apical ancestor (the coalescent sequence) from a sample of mtDNAs collected from individuals alive today. Because the mitochondria are so numerous in the cell, mtDNA can be recovered much more easily than DNA from the nucleus.

The mtDNA chromosome takes the form of a relatively small ring of 16,500 base pairs (16.5kB), which was entirely sequenced by 1981. By 1987, 147 human individuals had contributed additional sequences and the structure of the converged mtDNA lineage was announced, which 'coalesced' on a woman who was promptly dubbed 'mitochondrial Eve,' alleged to have lived in Africa about 200,000 years ago, and whose 'modern human' progeny had emigrated from Africa and repopulated the world, which was at

that time allegedly occupied entirely by earlier forms of human, such as Neanderthals and *Homo erectus* (Cann et al. 1987).

5.5 Y Chromosomes

There is one DNA molecule that is uniquely inherited in the paternal line, a kind of mirror image of mtDNA. When it was discovered that certain regions of the Y chromosome were very stable and apparently did not recombine during meiosis, it was hoped that the Y chromosome would become the analytic reciprocal of mtDNA. That is, Y chromosome lineages should parallel patrilineal social practices in the same way that mtDNA lineages paralleled matrilineal social practices. A dramatic confirmation of the utility of Y chromosome analysis has been provided by Thomas et al. (1998) who showed how Y chromosome lineages confirmed the patrilineal inheritance of the Jewish priesthood among widely dispersed Jewish populations. Seielstad and others have opened another promising line of research by comparing the geographical distributions of mtDNA and Y chromosome lineages in attempts to reconstruct the marriage practices of prehistoric peoples (see *Y-chromosomes and Evolution*; Stoneking 1998).

5.6 Satellites

Although the picture is constantly changing, researchers at this point recognize three general categories of tandemly repeated sequences, which are classified by length. They are satellites, the longest sequences, minisatellites, defined as sequences which are 500–40,000 base pairs in length, usually consisting of five or fewer repeated nucleotides, and microsatellites, which tend to be short and simple in motif, even consisting of repeats of a single base pair, or a triad or tetrad of nucleotides (for example, CACACA or CAGCAG) (Jackson et al. 1996, pp. 171–210).

These short, apparently neutral sequences are useful to anthropologists because their rate of mutational change is much higher than the single nucleotide mutation of DNA, so that minisatellites and microsatellites, like mtDNA, provide a finer control of historical relationships among individuals and the populations they represent. While the mutation rate for single nucleotides is about 10^{-7} , the rate for microsatellites is about 10^{-3} per generation.

6. Grand Synthesis

Before the entry of genetics into anthropology, the recent history of our species, over the last 6,000 years, was determined largely from written accounts or, in

the case of preliterate societies, from oral narratives. To reconstruct events farther back, in the late prehistoric period, archaeologists examined architecture and technology, with architecture becoming less important farther back in time and technology, especially stone tools, becoming more important, especially before the Neolithic period began about 10,000 years ago. In all periods, compared to cultural remains, biological evidence in the form of graves and fossil humans has been rare, and the interpretations have been controversial.

In some ways, the present evolutionary debate among anthropological geneticists can be seen as a continuation of an historic dispute in paleontology between 'lumpers' and 'splitters.' Lumpers have historically emphasized the general evolution of the whole species, while splitters have tended to group human fossils into named regional and temporal categories, which were considered to evolve somewhat separately and independent of other regional populations. Splitters tended to assign new names to new fossil finds, while lumpers were reluctant to inaugurate new taxa, emphasizing instead the ongoing wide diversity within the human species.

The nature of genetic evidence lends itself to the continuation of this kind of debate. To a considerable extent, all modern human populations share similar alleles, and can be distinguished in this respect only by their different frequencies. An array of allele frequencies from a population can in principle be analyzed as a combination of two or more previous arrays, identified by name and allegedly specific to some region. This is a splitter's approach. On the other hand, genetic 'lumpers' can maintain that the human species has always been diverse, and is no more heterogeneous now than it was hundreds of thousands of years ago.

The debate can also be characterized as between cladistic theorists who see local or regional human populations as largely maintaining their social and biological integrity, increasing in size, dividing into daughter groups, and then migrating to replace less successful populations. By contrast, ethnogenetic theorists postulate that gene flow by reciprocal marriages or matings among small adjacent populations might be as significant as long distance migration in changing regional allele frequencies, and that while language and culture might change rapidly within a region, biological characteristics change more slowly.

Luca Cavalli-Sforza and his colleagues have been prominent in this debate, and viewed by many as leaders of the cladistic school of thought. In the final decades of the twentieth century, they have produced several important theoretical books, followed by a huge compendium of maps, tables, and texts showing known global genetic variation (Cavalli-Sforza et al. 1994). In this school of thought, current human genetic variation is explained as the result of historic and prehistoric migrations of named ethnic groups from

region to region and continent to continent, especially including the 'African Replacement' by Eve and her progeny.

If these works have not entirely convinced all of their academic readers, they at least have served as a convenient target for 'lumpers' and advocates of ethnogenetic theory. Prominent among these critics has been Milford Wolpoff, who argues, primarily on the basis of osteological evidence, that global human evolution has been mostly *in situ*, since modern populations, in the Far East for example, tend to reflect unique or unusual osteological features exhibited by pre-Eve residents of the region. The other main line of criticism has been from linguists and social anthropologists, who argue that Cavalli-Sforza and his co-workers have got the languages and ethnography wrong, and that their hypothesized migrations and other social and linguistic processes cannot be confirmed independently (MacEachern 2000, Terrell 2000). Such critics argue that human groups, especially the small-scale hunting-collecting societies which existed in Paleolithic times, were more fluid and less bounded than required by the migration hypotheses.

Accumulating genetic evidence seems to support the migration hypotheses. Concerning Eve, other genetic lineages besides mtDNA also seem to coalesce at about 120–200,000 years ago, supporting the Out of Africa hypothesis. However, rates of evolution of sequences on the autosomal chromosomes are difficult to judge, and the particular histories of gene families and other segments of chromosomes, especially their evolutionary odysseys of travel among the chromosomes over the last millions of years, are just beginning to unfold.

The analysis of ancient DNA from human fossils has the potential for resolving many problems of human prehistory and evolution, provided that appropriate and usable DNA can be retrieved from scarce and damaged human bones, of great antiquity. Already the examination of mtDNA sequences from a few Neanderthal specimens has shown deep rather than shallow genealogical relationships to modern human sequences, as if indicating that modern Europeans replaced them, rather than being descended from them. A human of lesser age, the so-called and well-publicized Tyrolean Ice Man, has an estimated antiquity of about 5,000 years, according to carbon-14 dating, and could have been ancestral to modern Central Europeans (Herrmann and Hummel 1994, Kolman and Tuross 2000).

The most formidable but most satisfying task facing anthropological geneticists in the near future is to continue the confirmation and criticism of the fit, or lack thereof, among genetic, linguistic, and archaeological evidence for human migrations. No technique has yet been found for differentiating allele distributions resulting from gene flow from those resulting from a sudden migration after several generations

have passed. Inter-marriage and recombination tends to blur the distinction, so that anthropologists for the next several decades will no doubt continue to debate the accumulating evidence for global, continental, and regional migration and gene flow.

7. Problems

The greatest barrier to using variation in DNA sequences—polymorphisms—to trace the antiquity of relationships among human individuals and populations is the unreliability of the molecular clock. While mtDNA sequences, DNA lineages constituting gene families, and specific microsatellite sequences can be arranged into evolutionary trees, no one knows for sure the time between each generation or branch of the tree. It is clear that the frequencies of base pair deletions and other forms of mutation are not the same in different chromosomal regions, and similarly the amount and dynamics of variation seem to vary greatly among regions on the same chromosome, and among the autosomes, the Y chromosome, and mtDNA. There also seem to be ‘hot spots’ where mutations or recombinations are more frequent, near telomeres and centromeres, for example. An additional concern in calibrating the molecular clock is that the ‘proofreading,’ repair, and editing functions of genetic reproduction seem to be uneven for different genes and loci (especially, mtDNA seems to have weaker repair systems), so that some sequences are actually much older than they appear, because they have been systematically repaired (Kumar and Hedges 1998). What all of this means is that each gene has its own history, and only by aggregating information from many such histories do some aspects of the picture of human history begin to emerge.

Bibliography

- Cann R L, Stoneking M, Wilson A C 1987 Mitochondrial DNA and human evolution. *Nature* **325**: 31–6
- Cavalli-Sforza L L, Menozzi P, Piazza A 1994 *The History and Geography of Human Genes*. Princeton University Press, Princeton NJ
- Crawford M 2000 Anthropological genetics in the 21st century: Introduction. *Human Biology* **72**(1): 3
- Durham W H 1991 *Coevolution*. Stanford University Press, Stanford, CA
- Friedlaender J S 1975 *Patterns of Human Variation: The Demography, Genetics, and Phenetics of Bougainville Islanders*. Harvard University Press, Cambridge, MA
- Goodman M 1999 Molecular evolution '99: The genomic record of humankind's evolutionary roots. *American Journal of Human Genetics* **64**: 31–9
- Harpending H 1974 Genetic structure of small populations. *Annual Review of Anthropology* **3**: 229–43
- Harrison G A, Weiner J S, Tanner J M, Barnicot N A 1964 *Human Biology*. Oxford University Press, Oxford, UK
- Herrmann B, Hummel S 1994 *Ancient DNA*. Springer, New York
- Hillis D M, Moritz C, Mable B K 1996 *Molecular Systematics*. Sinauer Associates, Sunderland, MA
- Jackson M S, Strachan T, Dover G (eds.) 1996 *Human Genome Evolution*. BIOS, Oxford, UK
- Jones S, Martin R, Pilbeam D 1992 *The Cambridge Encyclopedia of Human Evolution*. Cambridge University Press, Cambridge, UK
- Jorde L B, Bamshad M, Rogers A R 1998 Using mitochondrial and nuclear DNA markers to reconstruct human evolution. *BioEssays* **20**: 126–36
- Kolman C J, Tuross N 2000 Ancient DNA analysis of human populations. *American Journal of Physical Anthropology* **111**: 5–11
- Kumar S, Hedges S B 1998 A molecular timescale for vertebrate evolution. *Nature* **392**: 917–20
- Long J C 1993 Human molecular phylogenetics. *Annual Review of Anthropology* **22**: 251–72
- MacEachern S 2000 Genes, tribes, and African history. *Current Anthropology* **41**(3): 357–84
- Molnar S 1998 *Human Variation*. Prentice-Hall, Saddle River, NJ
- Neel J V, Ward R H 1970 Village and tribal genetic distances among American Indians, and the possible implications for human evolution. *Proceedings of the National Academy of Sciences* **65**(2): 323–30
- Sarich V M, Wilson A C 1967 Immunological time scale for Hominid evolution. *Science* **158**: 1200–3
- Stein P L, Rowe B M 2000 *Physical Anthropology*. McGraw-Hill, New York
- Stoneking M 1998 Women on the move. *Nature Genetics* **20**: 219–20
- Terrell J E (ed.) 2000 *Archaeology, Language, and History*. Bergin and Garvey, Westport, CT
- Terwilliger J D, Göring H H 2000 Gene mapping in the 20th and 21st centuries: Statistical methods, data analysis, and experimental design. *Human Biology* **72**(1): 63–132
- Thomas M G, Skorecki K, Ben-Ami H, Parfitt T, Biadman N, Goldstein D 1998 Origins of Old Testament priests. *Nature* **394**: 138–9
- Weiss K M 1993 *Genetic Variation and Human Disease*. Cambridge University Press, Cambridge, UK
- Weiss K M 1998 Coming to terms with human variation. *Annual Review of Anthropology* **27**: 273–300

J. H. Moore

Genetics of Complex Traits Through the Life Cycle

The purpose of this article is to present a schema for conceptualizing the sources of individual differences in lifespan developmental processes. The two broad domains of genetic and environmental factors are presented as co-acting in their influence on complex traits, with the relative influence differing from trait to trait. Furthermore, intricate interactions are possible among genes and between genes and environments. The system is dynamic, with the possibility that the relative contributions of genetic and environmental sources can change over time. Newly available

methods assure finer-grained descriptions of the inter-relationships of genes and environment in the 'architecture' of complex systems in the near future.

1. Mendelian Genetics

Prior to the research of Mendel, the general conceptualization of the hereditary material was of an amorphous, miscible substance, in which the contributions of the parents were blended and influences from the environment altered the mix. Mendel's ingenious contribution was to postulate the existence of discrete elements (later termed 'genes'). Further postulations, in abridged form, were that an organism possesses many such genes, and each occurs as a pair, with an organism receiving one element from each parent. There are different forms of the genes (later to be named 'alleles'). Just as an individual received one allele for each gene from each parent, it transmits only one of each pair to its offspring. The alleles do not blend and are not modified during the association with their partner alleles or with the alleles of other genes, but are transmitted from the organism in the same functional condition in which they were contributed to the organism. Likewise, environmental factors do not alter the nature of the elements.

A critical distinction in Mendelian genetics is made between the phenotype, the measured attribute, and the relevant genes, or genotype, and the crux of the theory was the set of rules determining the relationships that could obtain between genotype and phenotype.

Mendelian genetics deals with categories. In the ideal case, all individuals are assignable unequivocally to one of two or three categories. A particular reason that the genotype/phenotype distinction is important is that the relationship between the two is not always additive. Knowing the phenotype is not always unambiguously informative about genotypes. For example, in one classical Mendelian situation, those individuals possessing two copies of a particular allele of the gene belong to category A, those possessing two copies of another allele of the same gene belong to category B, and those possessing one copy of each also belong to category B. That is to say, one copy of the allele 'for' category B is as effective as two. This relationship is described as dominance; the allele for which one copy is as effective as two is the dominant allele, and the other is termed the recessive allele. The two genotypes with identical alleles are called homozygotes, and the genotype with one each is called a heterozygote. Other relationships are possible: In some situations, the relationship between allelic 'dose' and phenotype is strictly additive, and intermediate degrees of dominance are found for other phenotypes.

Even when dominance is present, there is, of course, an additive component to the relationship. Thus, the phenotypic variability can be regarded as due to an

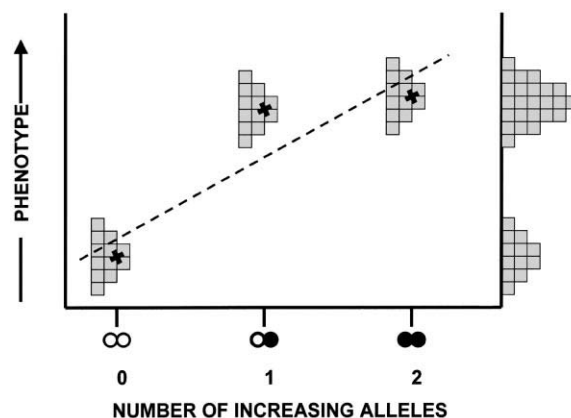


Figure 1

Depiction of a Mendelian partial dominant situation with environmental influences responsible for within-genotype phenotypic variation. Phenotypic values of all genotypes are projected onto a single distribution on the right Y axis

additive component and to dominance deviations. Another major reason that the distinction between genotype and phenotype is critical is that genes may not be the sole determinants of the phenotypic outcome. The basic principle of multiple influences underlying quantitative genetics is illustrated in Fig. 1. Phenotypic values are shown for a number of individuals of each genotype for a single gene with a partially dominant functional relationship to the phenotype. The within-genotype variability is assumed to be due to environmental circumstances that have uniquely affected each individual. A projection of these phenotypic values onto the right Y-axis reveals a distribution, the variability of which is attributable both to genetic differences (with some of this genetic influence due to the additive effect of allelic 'substitution' (represented by the dashed regression line), and some due to dominance (the deviations of the X's marking the means of each group from the regression line) and also with a portion of the variance due to environmental differences among the individuals. The consequence when more than one gene can affect the phenotype will be discussed below.

The source of genotypic diversity is, of course, the existence of different alleles. These alternative forms of the gene arise through mutational events, and their ultimate fate in a population depends upon their influence on the reproductive 'fitness' of their possessors, among other matters (see Falconer and Mackay 1996, Hartl and Clark 1997). In any population, the relative frequencies of the alternative alleles (there may be many more than two, but this discussion will be restricted to the simple case of two) will ordinarily be far from equal. Thus, the different genotypes will also not be equally frequent. If p represents the

frequency of the most abundant allele (say A1) and q represents the least frequent one (A2), and if mating is at random, then the relative frequencies of the genotypes $A1A1$, $A1A2$, and $A2A2$ will be p^2 , $2pq$, and q^2 , respectively. Different values of p and q will thus obviously lead to different genotype frequencies, and thus to different means and to different variances.

2. Polygenic Inheritance

For many phenotypes, it is not possible to assign individuals to categories; instead, they are arrayed along a continuous distribution. Although for a time there was a conjecture that different genetic systems were engaged in the categorical and continuous, a model (Fisher 1918) that postulated the existence of many loci, each performing according to Mendelian rules, but each having only a small influence on the same phenotype, resolved the issue. The collective effect of these multiple genes (polygenes) yields statistical expectations of resemblance of relatives of different degrees of biological relationship. For example, our understanding of the mechanism of gene transmission informs us that a parent shares half of his or her alleles with a child; grandparents share one quarter of their alleles with their grandchildren; siblings share half of their alleles on average; fraternal twins are just like siblings in this regard, and identical twins share all of their alleles. Comparison of the degree of similarity of phenotypes to the fraction of shared alleles permits the estimation of the proportions of phenotypic variance that can be attributed to genetic sources and to environmental sources. Each of these broad categories can be further divided. The overall genetic component can be described in terms of an additive component and a non-additive one. The non-additive component itself can be subdivided into effects arising from dominance, which can be regarded as the interaction of alleles at a locus, and from epistasis, which is the interaction of genotypes at different loci. This latter effect will be discussed in more detail later.

In the quantitative genetic model, environment is usually defined by exclusion—it is all of the influences that are non-genetic; i.e., not arising in a causal sense from the coded material of the DNA. Thus the scope and range of environment in this conceptualization is enormous, including in its compass factors from cellular biochemistry to peer group dynamics. This environmental contribution to phenotypic variance can also be partitioned in a number of ways, depending on the purpose of the research and on the data available. Long-term environmental effects may be distinguished from short term ones, for example, for developmental studies. One common distinction is that between environmental effects that are shared by family members and non-shared environmental influences that are experienced uniquely by individuals.

A simple, broad-stroke conceptual model represents the phenotype as the additive outcome of genetic and of environmental influences. Algebraically, this can be stated as

$$P = G + E$$

In a system such as this, the variances can be described as

$$V_P = V_G + V_E + 2\text{Cov}_{GE}$$

As noted previously, these terms are further divisible. Thus, as noted, V_G includes both an additive component (V_A) and a nonadditive (V_{NA}) component, and V_E is similarly subdivisible. These components yield an estimate of the relative influence of the two general domains. (There are also formulations that include a component of stochastic or of nonlinear dynamic influences [see Molenaar et al. 1993, Finch and Kirkwood 2000].) Measurement error is usually found in the environmental term, and as will be discussed below, there are also interaction terms that can be of great importance. The most commonly encountered index is that of heritability, which is the proportion of total phenotypic variance that is due to genetic differences among individuals. In the narrow sense, heritability is V_A/V_P , and in the broad sense, it is V_G/V_P . These two values are differentially useful for different descriptive and predictive purposes (see, for example, Falconer and Mackay 1996). The same information can be captured by an index of environmentality ($1 - h^2$), of course, but this term is infrequently encountered. It is very important to recognize that heritability and environmentality are descriptive statistics describing the state of affairs of a specific population under specific environmental circumstances. Different gene pools, or different environmental forces, might yield quite different estimates. Note also that high heritability does not mean that there is no conceivable environmental influence that can affect the phenotype. It means that in the population that was studied, there was either no sufficiently powerful environmental influence, or that if one existed, it impinged on so few of the population that it had little effect on the overall variability. Thus high heritability for a condition such as Attention Deficit Hyperactivity Disorder should not inhibit the search for environmental interventions. The classical single locus exemplar of this point is that of phenylketonuria, in which the homozygous recessive condition for a particular allele at the phenylalanine hydroxylase locus results in devastating mental retardation in ordinary environmental circumstances. However, the understanding of the metabolic process in which phenylalanine hydroxylase is involved led to the development of a rational intervention—a diet without phenylalanine, which cannot be metabolized normally by the homozygous recessive individuals.

Note that a gene is an unobserved hypothetical element both in a Mendelian system and in a polygenic system. The basic sequence involved in identifying a Mendelian inheritance pattern is to observe a categorical distribution of some phenotype, then to determine if the transmission pattern within families accords to Mendelian rules. In a polygenic, multifactor system, the issue is whether the pattern of resemblance among relatives conforms to expectations of the quantitative model. It is not required to know anything concrete about the gene or genes involved in either of these situations. But in fact, historically, while the range of application of Mendelian rules was being explored and the elaboration of the quantitative genetic model was proceeding, much was also being learned about the physical reality of genes. The key observation was that small bodies in the nucleus of cells behave according to rules similar to those of Mendel's elements. The hypothesis that these small bodies, the chromosomes, were the vehicles of the genes was quickly confirmed, and the field of cytogenetics emerged. Examination of the fine structure of chromosomes revealed distinctive banding patterns that permitted specification of chromosomal location, and subsequent research showed that the Mendelian elements resided in fixed positions on the chromosomes. The position of a gene came to be called its locus, for obvious reasons. The term, locus, was a terminological advance, making it possible to distinguish easily between the concept of gene, which everyone has, and the alleles, which different individuals can have in different combinations.

Throughout the first half of the twentieth century, a major biological mystery concerned the mechanism through which genes exert their influence. In mid-century occurred the revolution of molecular biology that discovered the chemical nature of genes and revealed the processes through which the genetic information in DNA is transcribed into RNA and translated into protein. This work connected the genetics of hypothetical genes to anatomy, physiology, immunology, endocrinology, biochemistry, neurobiology—to all of the disciplines that have explored the biological functioning of the organism. The question of the mechanism of gene action is thus answered in broad principle (although the details of specific cases may be unknown). Genes have no special route of influence to the phenotype; they operate through the complex systems comprising the familiar structures and functions of the body.

Complex systems such as this have some properties of great importance for the developmental sciences. A fundamental feature is the existence of feedback relationships of the sort involved in regulated processes, such as the familiar negative feedback in homeostatic systems. Another related feature is the typical presence of redundant pathways leading to the phenotype. With such properties, deficiencies in a particular element of a pathway might be compensated by up-

regulation of other elements in that pathway, or by enhanced activity in the parallel pathway. Kacser and Burns (1979) have emphasized that in biochemical pathways with feedback loops, it becomes impossible to specify the impact of any element without specifying the other elements in the system. In general, the simple notions of linear, unidirectional causation must yield to circular or network causal concepts (Sattler 1986). Specific to interpretation of genetic influences, we are cautioned that the effect of genotypic differences at a particular locus may be greatly influenced by other loci in the system, or by the environment in which the organism exists. The magnitude and complexity of these interactions will be illustrated later.

3. New Tools

Recent developments in molecular genetics have provided some remarkable new methods with which to approach the issues identified in this review. Some of the most relevant are the following.

3.1 Quantitative Trait Loci

Although the original quantitative genetic model postulated an indefinitely large number of loci with equal small effects, it has long been suspected that within a polygenic set there may exist a distribution of effect sizes. Prior to the genome projects that have provided chromosomal markers throughout the genomes of human beings and several model species, attempts to locate such loci have been fraught with difficulties. With the availability of the new markers, it has become possible to identify segments of the chromosomes in which are located genes that have a detectable influence on the variability of a phenotype. This is obviously a critical step on the way to further characterization of the locus at the molecular level. These individuated polygenes are termed quantitative trait loci (QTLs).

3.2 Transgenic Procedures

Among other enticing possibilities, there are now procedures that bring about changes in the DNA molecule constituting a gene so as to prevent its expression (knockout preparations), to alter its regulatory system in order to cause overexpression, or to insert a new gene. These preparations, permitting the examination of the downstream results of altering single loci, provide new avenues of approach to the causal nexus between gene and phenotype.

3.3 Arrays

Wedding of the technological advances of microchips with molecular biology has provided chips on which

are arrayed hundreds or thousands of gene segments. When exposed to preparations containing DNA from particular organs or tissues, the chips can identify, through the hybridization of the tissue DNA to the segments on the chips, those genes that are being expressed in the tissue being investigated. Both the bright promise and the methodological issues of this new technology have been well presented by Claverie (1999) and by Lockhart and Barlow (2001).

Although these procedures and techniques are new on the scene, they are already providing powerful complements to the accumulated data on interactions and developmental changes in the architecture of complex phenotypes.

4. Epistasis

The phenomenon of interaction among genes has been known since the early days of the last century. An epistatic variance component can be incorporated in the quantitative genetic model, as well, but the detection of interaction effects in polygenic systems is challenging. There are numerous examples of the dependence of single locus effects on a polygenic background, however. A now-classic result can illustrate. Coleman and Hummel (1975) investigated the phenotypic consequences of homozygosity for a particular allele (*db*) at the diabetes locus in inbred mice. Inbred strains approximate the condition of being homozygous at all loci (see McClearn and Hofer 1999); animals within a strain are identical, and different inbred strains differ from each other at some unknown number of loci. It was possible to arrange the *db/db* genotype on the background of two different (but related) inbred strains—C57BL/6 and C57BL/Ks. The phenotypic outcome differed dramatically, in respect to body weight, glucose level, pathophysiology (atrophy of the Islet cells in one strain, and hypertrophy in the other), and in effects on lifespan.

The advent of QTL methods has greatly facilitated the study of epistasis. It has become possible to examine the interaction of these as yet unidentified (but approximately located) loci. In mice, to illustrate the point, QTL interactions have been described affecting blood pressure (Rapp et al. 1998), diabetes (Galli et al. 1999), epileptic seizure frequency (Frankel et al. 1995), and lung cancer (Fijneman et al. 1996).

In human research, an epistatic interaction concerning risk for dementia has been reported. The $\epsilon 4$ allele of the *ApoE* locus has been implicated by many studies as a risk factor for the development of Alzheimer's disease. Overall, an allelic dose-response relationship has been observed, with a significant risk in heterozygous condition and an even greater risk in homozygous condition. Kamboh et al. (1995) describe an interaction of this locus and the *ACT* locus, alleles of which are labeled *A* and *T*. If an individual is homozygous *TT*, the risk associated with two $\epsilon 4$ alleles is not greater than with one. The $\epsilon 4$ -associated risk is

elevated if individuals are also *AA*, and individuals homozygous both for $\epsilon 4$ and for *A* are especially vulnerable.

These few examples represent the scope for the influence of genetic context on the effect of allelic differences at specific loci (i.e., the effect of allelic substitution, or the allelic dose-response effect). Characterization of a gene as 'the' gene for some phenotype must be understood to involve acknowledging the potential of this contextual influence.

5. Gene-Environment Interaction

Under the rubric of gene-environment interaction are those phenomena in which the relation of genotype to phenotype is altered as a function of environment, or, conversely, in which the impact of an environmental agency or condition is dependent upon the genotype of the individual exposed to the environment.

This theme can be illustrated by the results of Fosmire et al. (1993), who, in the context of risk factors in Alzheimer's disease, explored the effects of exposure to high levels of dietary aluminum on brain aluminum levels in five strains of mice. Three strains showed no elevation as a consequence of the exposure, one showed a moderate effect, and one showed a threefold increase.

Logically parallel to evidence such as this on genetic influence on risks of deleterious outcomes engendered by environmental factors is evidence of genetic influence on responsiveness to intended therapeutic interventions. Yaffe et al. (2000), for example, investigated the effects of estrogen used by women 65 years of age and older on risk of cognitive impairment, and found an interaction with *ApoE* genotype. Risk was not affected in women who possessed at least one $\epsilon 4$ allele, but was reduced in those who were $\epsilon 4$ -negative.

A final example illustrating the potential scope and complexities of gene-environment interactions may be taken from the work of Vieira and colleagues (2000). Seventeen QTLs were identified affecting lifespan in *Drosophila* reared in one of five different environments: three levels of rearing temperature, one involving a brief high temperature exposure, and one involving starvation stress. The results revealed an astounding level of interactions of the QTLs with sex and with environment. Some QTLs were effective only in one or a few environments and not in others; some were effective only in one sex; some were effective in both sexes, but only in one environment; and so on. In a summary statement, the authors reported that *all* of the genetic variance in this system was involved in some type of interaction!

6. Temporal Considerations

Dynamic change in the network of genetic and environmental factors over time is a key concept for developmental genetics. A fundamentally important

principle to emerge from investigations of the mechanisms of gene action is that genes may be turned on and off, or may have their level of expression up-regulated or down-regulated. There were temporal aspects to many of the previously cited studies; in a number of researches, the temporal dimension is the major focus. A particularly apposite example is the work of Rogina and Helfand (1995), who measured the expression of a gene in the antennae of *Drosophila melanogaster*. At hatching, there is little or no activity of this gene, but expression rises systematically to a peak in mid-life, with a subsequent gradual decline to near-zero levels at the end of life. When lifespan was shortened under conditions of an elevated rearing temperature or through the introduction of a different gene that shortens lifespan, the expression of the gene followed the same pattern, but in an accelerated manner, scaling to the proportion of the lifespan rather than to chronological time.

The existence of a program that 'knows' what proportion of a lifespan has already passed (and what remains) is an intriguing prospect. But for our present purposes, the important point is that this unambiguous demonstration of lifespan-scaled change in expression of this single gene informs us of the sort of changes that can be taking place in a polygenic system.

An example of dramatic developmental change in a mammalian system is provided by Cheverud et al. (1996) who studied mouse growth. Two sets of QTLs were identified: one affecting early growth but with diminishing influence by 6 weeks of age, and the other coming into play in the interval from 3 to 6 weeks of age.

It remains unclear how many or what proportion of loci can be expected to undergo systematic changes during the lifespan. The new chip technology is poised to generate a flood of data that will illuminate this issue, among many others. As a promissory illustration, Lee and colleagues (1999) have described differences between 5-month and 30-month old mice in respect to gene expression in skeletal muscle. Of 6,347 genes examined, 58 showed a more-than-twofold increase in expression levels and 55 showed a decrease of that magnitude. Although these numbers indicate that only a small percentage of genes are changing expression of this magnitude in this tissue over this time span, the impact of these changes in complex aging systems could be substantial. Information from other tissues is urgently needed, and will undoubtedly be reported soon.

Life stage has also been shown to be critical to the genetics of development in human beings. In a single locus example, Ferrari et al. (1998) showed that the major locus for the vitamin D receptor has an appreciable effect on the bone mass of pre-pubertal girls, but does not have any such influence during post-pubertal adolescence.

Concerning late life, there is a widely shared expectation that with advancing age the cumulative

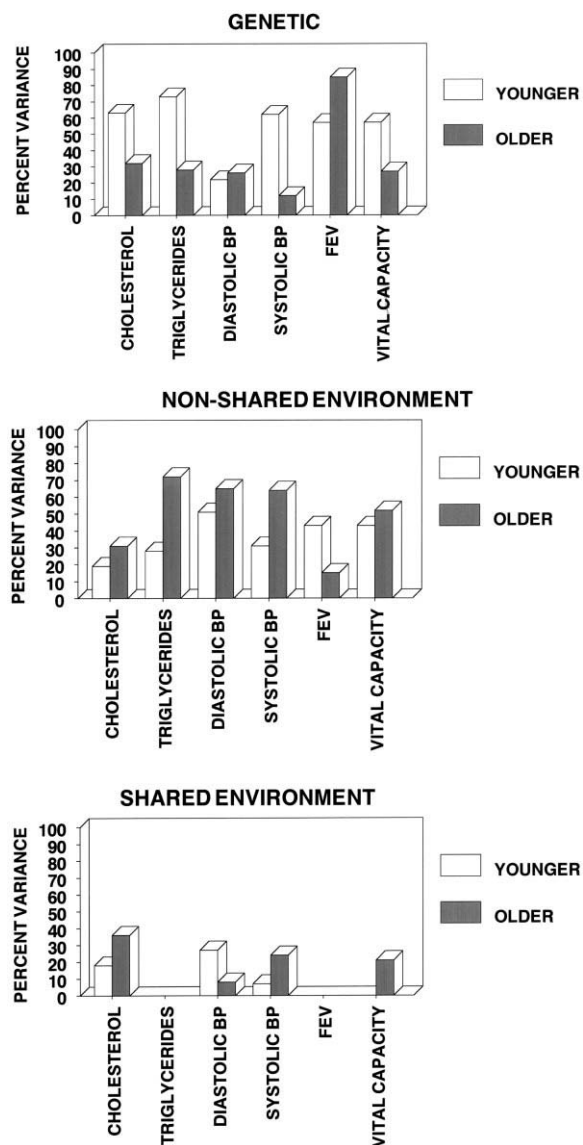


Figure 2

(a) Percent of variance of various biomedical phenotypes attributable to genetic sources. Data from the Swedish Adoption/Twin Study of Aging. Estimates are shown for a younger group of twin pairs (aged less than 65 years) and an older group (65 years of age and older). (b) Percent of variance of various biomedical phenotypes attributable to non-shared environmental sources. Data from the Swedish Adoption/Twin Study of Aging. (c) Percent of variance of various biomedical phenotypes attributable to shared environmental sources. Data from the Swedish Adoption/Twin Study of Aging.

effects of uniquely experienced environmental factors will lead to reduced proportional contribution of genetic factors to that variance. Evidence on shifting variance components in aging has been provided by twin studies on the elderly; the classic expectation is fulfilled in the case of some phenotypes, but not all. Fig. 2 gives results from some biomedical variables included in the Swedish Adoption/Twin Study of Aging (SATSA) (Heller et al. 1993, Hong et al. 1994, McClearn et al. 1994). Estimates were derived from structural equation modeling of genetic, shared environmental, and non-shared environmental components of variance.

Participants in the study were Swedish identical and fraternal twins, some of whom were reared together and some of whom were separated in early life and reared separately. Estimates were derived separately for a younger group (less than 65 years of age) and an older group (65 years of age and older). Figure 2a gives the estimates of heritability. The difference between the older and younger cohort clearly conforms to the classical expectation of reduced heritability for measures of cholesterol, triglycerides, systolic blood pressure, and vital capacity. Heritability of diastolic blood pressure is, however, modest and essentially the same in both cohorts, and, with respect to forced expiratory volume (FEV), heritability is very substantial in the younger cohort, and even higher in the older cohort.

Fig. 2b gives the estimates for the influence of non-shared environment for these phenotypes. Most of these phenotypes (FEV being the exception) do display the classically expected result of greater non-shared environmental influence in the older group. Shared environmental influences, shown in Fig. 2c, are generally smaller than the nonshared environmental components. An interesting result is the hint that the effects of early shared rearing environment on cholesterol, systolic blood pressure, and vital capacity may be more manifest in the older cohort than in the younger one.

In addition to the fluctuation of genetic influence over the lifespan, there is growing evidence of shorter-term temporal variability. In rats, for example, the liver mRNA concentration of aldolase B, an enzyme of the glycolytic system, declines during fasting, and increases four- to eightfold upon re-feeding of a carbohydrate-rich diet (Munnich et al. 1985). Investigating the mechanisms involved revealed that regulation of the expression of the relevant gene is accomplished very differently in liver, kidney, and small intestine.

Another example of environmentally mediated alteration of gene expression is provided by the work of McMahon and colleagues (1992), who studied the effects of immobilization stress in rats. Among other results, it was found that a single stress resulted in elevation of levels of tyrosine hydroxylase mRNAs in the adrenal glands.

Thus, the responsiveness of the genetic system to the environmental stimulus can occur quickly. From the point of view of behavioral science, an especially pertinent observation is that the expression of a gene in a particular brain region can be classically conditioned (Smith et al. 1992).

7. Summary

The studies reviewed here are but a few, selected, examples from a vast literature pertaining to an enormously broad array of phenotypes observed on species representing the phyletic range from prokaryotes to eukaryotes. Obviously, a brief review such as this can provide only the sketchiest outline. But, hopefully, it provides a hint as to the possibilities of dynamic interaction of genes with genes and of genes with environment and of the changes of this whole complex system over time. It may also serve as a preventive against simplistic interpretations of the old nature vs. nurture variety in issues of lifespan development. The current data are impressive; the anticipated flood of new data from new methodologies will provide us both with sharpened reductionist understanding, and, in the context of the quantitative model, an integrated view appropriate to the complex systems with which the developmental sciences are concerned.

See also: Behavioral Genetics: Psychological Perspectives; Familial Studies: Genetic Inferences; Genes and Culture, Coevolution of; Genes and Languages; Genetic Factors in Cognition/Intelligence; Genetic Studies of Behavior: Methodology; Genetic Studies of Personality; Genetics and Development; Human Evolutionary Genetics; Longevity, Genetics of; Mental Illness, Genetics of; Schizophrenia and Bipolar Disorder: Genetic Aspects; Senescence: Genetic Theories

Bibliography

- Cheverud J M, Routman E J, Duarte F A M, van Swinderen B, Cothran K, Perel C 1996 Quantitative trait loci for murine growth. *Genetics* **142**: 1305–19
- Claverie J-M 1999 Computational methods for the identification of differential and coordinated gene expression. *Human Molecular Genetics* **8**: 1821–32
- Coleman D L, Hummel K P 1975 Influence of genetic background on the expression of mutations at the diabetes locus in the mouse. II. Studies on background modifiers. *Israeli Journal of Medical Science* **11**: 708–13
- Falconer D S, Mackay T F C 1996 *Introduction to Quantitative Genetics*, 4th edn. Longman, London
- Ferrari S L, Rizzoli R, Slosman D O, Bonjour J P 1998 Do dietary calcium and age explain the controversy surrounding

- the relationship between bone mineral density and vitamin D receptor gene polymorphisms? *Journal of Bone and Mineral Research* **13**: 3363–70
- Fijneman R J, de Vries S S, Jansen R C, Demant P 1996 Complex interactions of new quantitative trait loci, *Sluc1*, *Sluc2*, *Sluc3*, and *Sluc4*, that influence the susceptibility to lung cancer in the mouse. *Nature Genetics* **14**: 465–7
- Finch C E, Kirkwood T B 2000 *Chance, Development, and Aging*. Oxford University Press, New York
- Fisher R A 1918 The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**: 399–433
- Fosmire G J, Focht S J, McClearn G E 1993 Genetic influences on tissue deposition of aluminum in mice. *Biological Trace Element Research* **37**: 115–21
- Frankel W N, Valenzuela A, Lutz C M, Johnson E W, Dietrich W F, Coffin J M 1995 New seizure frequency QTL and the complex genetics of epilepsy in EL mice. *Mammalian Genome* **6**: 830–8
- Galli J, Fakhrai-Rad H, Kamel A, Marcus C, Norgren S, Luthman H 1999 Pathophysiological and genetic characterization of the major diabetes locus in GK rats. *Diabetes* **48**: 2463–70
- Hartl D L, Clark A G 1997 *Principles of Population Genetics*, 3rd edn. Sinauer, Sunderland, MA
- Heller D A, de Faire U, Pedersen N L, Dahlén G, McClearn G E 1993 Reduced importance of genetic influences for serum lipids in the elderly: A study of twins reared apart. *New England Journal of Medicine* **328**: 1150–6
- Hong Y, de Faire U, Heller D A, McClearn G E, Pedersen N 1994 Genetic and environmental influences on blood pressure in elderly twins. *Hypertension* **24**: 663–70
- Kacser H, Burns J A 1979 Molecular democracy: Who shares the controls? *Biochemical Society Transactions* **7**: 1149–60
- Kamboh M I, Sanghera D K, Ferrell R E, DeKosky S T 1995 *APOE**4-associated Alzheimer's disease risk is modified by α 1-antichymotrypsin polymorphism. *Nature Genetics* **10**: 486–8
- Lee C-K, Klopp R G, Weindruch R, Prolla T A 1999 Gene expression profile of aging and its retardation by caloric restriction. *Science* **285**: 1390–3
- Lockhart D J, Barlow C 2001 Expressing what's on your mind: DNA arrays and the brain. *Nature Reviews Neuroscience* **2**: 63–8
- McClearn G E, Hofer S M 1999 Genes as gerontological variables: Uses of genetically heterogeneous stocks. *Neurobiology of Aging* **20**: 147–56
- McClearn G E, Svartengren M, Pedersen N L, Heller D A, Plomin R 1994 Genetic and environmental influences on pulmonary function in aging Swedish twins. *Journal of Gerontology: Medical Sciences* **49**: M264–8
- McMahon A, Kvetnansky R, Fukuhara K, Weise V K, Kopin I J, Sabban E L 1992 Regulation of tyrosine hydroxylase and dopamine β -hydroxylase mRNA levels in rat adrenals by a single and repeated immobilization stress. *Journal of Neurochemistry* **58**: 2124–30
- Molenaar P C M, Boomsma D I, Dolan C V 1993 A third source of developmental differences. *Behavior Genetics* **23**: 519–24
- Munnich A, Besmond C, Darquy S, Reach G, Vaulont S, Dreyfus J O C, Kahn A 1985 Dietary and hormonal regulation of aldolase B gene expression. *Journal of Clinical Investigation* **74**: 1045–52
- Rapp J P, Garrett M R, Deng A Y 1998 Construction of a double congenic strain to prove an epistatic interaction on blood pressure between rat chromosomes 2 and 10. *Journal of Clinical Investigation* **101**: 1591–5
- Rogina B, Helfand S L 1995 Regulation of gene expression is linked to life span in adult *Drosophila*. *Genetics* **14**: 1043–8
- Sattler R 1986 *Biophilosophy: Analytic and Holistic Perspectives*. Springer-Verlag, Berlin
- Smith M A, Banerjee S, Gold P W, Glowa J 1992 Induction of *c-fos* mRNA in rat brain by conditioned and unconditioned stressors. *Brain Research* **578**: 135–41
- Vieira C, Pasyukova E G, Zeng Z-B, Hackett J B, Lyman R F, Mackay T F C 2000 Genotype-environment interaction for quantitative trait loci affecting life span in *Drosophila melanogaster*. *Genetics* **154**: 213–27
- Yaffe K, Haan M, Byers A, Tangen C, Kuller L 2000 Estrogen use, APOE, and cognitive decline: evidence of gene-environment interaction. *Neurology* **54**: 1949–54

G. E. McClearn

Genocide: Anthropological Aspects

Genocide is a crime in international law. It was established as such by the United Nations Genocide Convention of December 9, 1948 which committed 'contracting' states to 'prevent and to punish ... acts committed with the intent to destroy, in whole or in part, a national, ethnical, racial or religious group.' This legal definition was reached following political compromises between the major powers at the UN, and a more precise, academic definition is adopted here, according to which genocide is 'a series of purposeful actions by a perpetrator(s) to destroy a collectivity through mass or selective murders of group members and suppressing the biological and social reproduction of the collectivity' (Fein 1990, pp. 28–9). After many years of academic avoidance of the topic, there has been a recent surge of interest across disciplines from psychology through political science to anthropology, the best of which work has been interdisciplinary in perspective. Sociological and anthropological investigation of genocide covers four main areas: the definition of genocide and the characterization of particular instances of mass slaughter; the sociology of genocides themselves; the sociology of representations of genocide (from 'survivor testimony' through memorialization and forms of material culture such as architecture to studies of war crimes tribunals); and the sociology of the after effects of genocide. These will be treated in turn.

1. Problems of Definition

The term genocide is a neologism, coined in 1944 by Raphael Lemkin, a legal scholar and Polish Jewish

émigré to the USA. Lemkin took over ancient Greek roots for 'tribe' and 'killing' to create a word that would adequately convey the enormity of Nazi and Axis crimes in Central and Eastern Europe during World War II. This definition inspired a UN General Assembly resolution of 1946 but when, two years later, the same body established a binding Genocide Convention, the range of reference was narrowed in two important ways. First, cases where a culture is destroyed without the extermination of people—now commonly referred to as *ethnocide*—were excluded. Secondly, and much more critically, in order to accommodate concerns of the USSR, which might have faced indictment for mass killing of 'class' enemies (as in the deliberately created famine in the Ukraine in 1932–3, or the then still undiscovered 'gulag'), the 1948 UN definition excluded the extermination of 'political and other groups' such as social classes.

These questions of legal definition are important because the Convention commits states which subscribe to it 'to prevent and to punish' the crime of genocide by pursuing perpetrators 'whether they are constitutionally responsible rulers, public officials or private individuals.' Moreover, since the Convention explicitly denies that genocide can be a 'political' crime, perpetrators are subject to the laws of extradition.

Perhaps precisely because of its ambitious promise, the Convention has been invoked remarkably little since 1948. Both the desire of member states to preserve inviolate their claims to sovereignty and of others not to have to commit resources to humanitarian interventions have played a role. Leaving aside UN silence during the slaughter of half a million Indonesian communists in 1965 and the death of 1.5 million Cambodians in 1975–9 (since it could be argued that the 1948 Convention excluded such cases) it is a matter of record that US officials at the UN were instructed in 1992 to avoid using the term 'genocide' with respect to Bosnia so as not to trigger obligations under the Convention. In fact, the only major development since 1948 has been the establishment, in 1994, of the *ad hoc* International Criminal Tribunals for former Yugoslavia and Rwanda, though under the peace enforcement provisions of the UN Charter (Chapter VII), not the Genocide Convention, and refused permanent status due to Chinese objections.

While considerable academic energy has been expended in attempting to classify particular mass murders as genocide or otherwise (the destruction of the Albigensian Cathar heretics in France in the thirteenth century, the removal of up to 50 million persons from Africa during the slave trade, the slaughter of the Hereros of South West Africa by German forces in 1904–7, the massacres of the Armenians in 1915, the persecution of political enemies under Stalin in the USSR, the elimination of various indigenous peoples such as the Ache of

Paraguay, to name but a few) perhaps the most broadly attended and resonant debates have concerned the uniqueness or otherwise of the Nazi genocide of the Jews. On the one hand, some scholars argue that the Jewish Holocaust is the only case in history of the attempted destruction of an entire people because they were that people. Other genocides were committed as a by-product of the course of war (Hereros) or political struggles (Democratic Kampuchea), or aimed to destroy a culture or way of life by eliminating some of its bearers (Gypsies in World War II), or occurred indirectly as a consequence of actions not *per se* intended to destroy a people (the slave trade). The Nazi persecution of the Jews differed from all these in intent, extent and 'success' (Bauer 1984).

Other scholars, however, resist the implication of the 'exceptionalist' position that the Holocaust occurred in some sense 'outside of normal history' and thus had neither precedent in the past nor lessons for the future. Unsurprisingly, this debate leads on to conflict over the interpretative framework for understanding the causes and the motivations of the Holocaust and its perpetrators. For their part, most anthropologists would favor retaining a comparative, sociological framework and resist the suggestion that any sociological phenomenon can be unique in the strong sense argued by exceptionalist scholars, or require recourse to abstract notions of 'evil' (Conte and Essner 1995, Clendinnen 1999).

This debate has implications beyond German historiography. An anthropologist working in former Yugoslavia has recently argued that the public understanding of what constitutes genocide, which derives from images of the Nazi Holocaust, prevented observers from recognizing an actually existing genocide as it developed before the world's eyes (Sorabji 1994). Indeed, one of the puzzles of the sociology of genocide is why outside observers tend in general to be so unwilling to face the facts until it is too late, and why even after the event such effort is put into obfuscation and evasion of the brute historical record.

2. *The Sociology of Genocidal Mass Murder*

Though anthropologists aim to study the sources of human social and cultural variation, surprisingly little effort has been contributed by this discipline to the study of genocidal mass murder. This gap in the literature is all the more surprising in that many of the acknowledged genocides of the last century and a half have taken place in parts of the world where anthropologists work. Naturally, part of the explanation lies in the way anthropologists carry out their research. Participant observation in the life of a community is not a methodology that is easily adapted to the study of murderous gangs manning roadblocks or units of

the local military on active service. Beyond methodological considerations, there may also be a sense, shared with those who work for Amnesty International, that to explain is in some sense to forgive, and that the task of observers is merely to record, lest we forget.

Recently, however, anthropological avoidance of this topic—an evasion that was characteristic of most social science, with some noteworthy exceptions like Leo Kuper (1982)—has been coming to an end. Inspired in part by a wave of ethnographic writing about political violence, especially in South Asia (Spencer 1990), a new generation of anthropologists have written on particular genocides (Bosnia, Rwanda, and Cambodia most notably, but also Nazi Germany). In this work, the anthropologists attempt, with more or less success, to bring to bear a sense that violence is rarely simply pragmatic and is often patterned in ways specific to a particular time and place and always interpreted in local, cultural terms.

As in the other social sciences and humanities which deal with cases of mass murder, anthropological analyses aim above all to answer the question, 'how could this happen?' Under what conditions did one group of humans attempt to wipe out another? Responses can be divided along a familiar line into two camps: those who stress cultural factors and those who stress social, political, and economic factors. Taking the former first, for authors in this school the goal is to provide a specifically 'anthropological contribution,' to prove that an ethnographer's interest in 'cultural variation' adds an otherwise missing ingredient to analyses of genocide.

Much of this 'culturalist' work (notably writing on Rwanda and Cambodia) takes a cue, wittingly or not, from an essay by the historian Natalie Zemon Davis (1973) on differential patterns of violence in the religious wars in France in the seventeenth century. Davis showed that the forms of Protestant on Catholic violence were not the same as those perpetrated by Catholics on Protestants. Each religious tradition had its own construction of the nature of demonic power in the other and adherents of the two faiths acted to cleanse the world accordingly. Because of their interest in demonstrating a 'cultural logic,' the anthropological writing on Rwanda and Cambodia does not advance theoretically on Davis's original insight (Taylor 1999, Hinton 1998). Evidence is presented that shows how in some instances cultural patterns seem to have given form to and shaped people's violent behavior, but little or no attempt is made to explain how particular segments of the population or groups of individuals came to participate in the genocide at a particular time and place.

Curiously, this kind of explanation of 'how it happened' has several features in common with the approach adopted by Daniel Goldhagen (1996) in *Hitler's Willing Executioners*, a book which enjoyed runaway popular success across the USA, Europe,

and particularly Germany in the mid-1990s (while being almost universally condemned by the historians with period expertise). Goldhagen's answer to the question 'how could it happen' was that it was Germans who killed Jews in World War II (it was not just they, in fact) and they did so because of the activation of a tradition which he dubbed 'eliminationist antisemitism' which had long characterized German culture (while in reality antisemitism had actually been a vote-loser for the Nazis prior to 1933).

For authors working in this tradition, genocide is derived or emerges out of a 'cultural model.' What tends to be missing from such explanations is any attempt to detail political manipulation by those who planned the genocides, to analyze the institutional apparatus which prepared and carried through the genocides, and to establish the sociological and even biographical profile of identified killers. In the case of Rwanda one has to turn away from the anthropologists, to the work of the political scientist Rene Lemarchand to discover, for instance, that the genocide was planned by a core of one extended family and advisors, promulgated through 300–500 middle level, rural cadres, set off by a presidential guard of around 6,000 persons who massacred all opponents of the genocide (Hutu and Tutsi alike) and finally continued by militias who may have made up 1–2 percent of the population, each member of which may have killed 200–300 people (Lemarchand 1997).

Work that searches for the cultural models which underpin and apparently feed directly in to genocide bears a close affinity with the first efforts by anthropologists to deal with issues of genocide during and shortly after World War II. Then adherents of the culture and personality school provided textbook/manual type explanations of the personality of the enemy in an effort to explain the success of totalitarian regimes in Japan and Germany and as a prelude to reconstruction and democratization. At the beginning of the twenty-first century, with decades of detailed analysis of the paradigmatic Nazi case behind us, there really is little excuse for allowing this kind of cultural reductionism to pass for 'anthropological explanation.'

An alternative approach within anthropology is to be found in work on Bosnia. Here the research was carried out by a British-trained *social* rather than US-trained *cultural* anthropologist (Sorabji 1994). This work owes much to the tradition of historical and sociological investigation of the Holocaust represented by such scholars as those around Herbert (2000), as well as the kind of political science of Bosnia pursued by Bougarel (1997). It puts stress less on 'culture' than 'local knowledge,' and tries to provide a fuller sense of 'what happened' before answering the more difficult question, 'why?'. Anthropological expertise is combined with a journalist's sense of political manipulation, a sociological interest in institutions, and a historian's concern with events of the recent and

remembered past. Not only is the Bosnian genocide not represented as being rooted in 'cultural traditions,' but it is shown how this very representation is part of the armory of the perpetrators, used, with great success in this case, to keep outsiders from interfering. As long as US Secretary of State Warren Christopher, or British Prime Minister John Major was advised that 'ethnic cleansing' was not 'genocide' and that deep-rooted cultural patterns on all sides were to blame, they felt justified in evading calls for intervention. Indeed, the very 'disorganized' quality of the genocide (compared with the Nazi model case) came to be taken as evidence of its nonexistence. In reality, however, institutional patterns were at work: Violent criminals were being released from Belgrade jails and sent to join militias. These operated formally independently but under the cover of the official military. Intellectuals were among their first targets, often identified on pre-prepared lists; local schools and other sites of intercommunal, neighborly interaction were chosen as locales for internment and torture; neighbors were deliberately sent to kill neighbors. The 'sites of memory' of Bosnian social life were being systematically and consciously destroyed by the instigators of a genocidal violence that was presented as medieval but was, in reality, wholly modern.

This kind of work points to a need to think more broadly about the psychological and mental underpinnings of the violent behavior of individuals and collectivities and to relate these findings to work elsewhere within anthropology. It is striking that the language of genocide in Bosnia, as in some other cases, was one of sacrifice, an invitation, one might think, to introduce anthropological work on the relation between religious representations of transcendence and political violence (e.g., Bloch 1992).

3. *The Sociology of the Representation of Genocide and Memorialization*

It is possible to argue that much of what has recently passed for analysis of the cultural soil of genocide in anthropology is in reality merely the re-presentation of local, folk explanations of genocidal behavior. This data, which is often part of the discourse around the genocide and ought to be part of the material analyzed and explained by a sophisticated social science, is instead relied upon, more or less uncritically, as if descriptively neutral.

In analyzing perpetrator discourse anthropologists have pointed out how genocide, like political violence more generally, tends to be presented by the perpetrators as a just defense against others who are about to attack, or are already attacking, the perpetrators. But a less sceptical attitude has been adopted, for instance, over accounts which focus on the rural, resentful character of supposed killers (Hinton 1998). Going

further into the psychological roots of genocidal violence, the fundamental challenge, as everyone knows who has ever investigated these issues, is that no one ever admits to having personally been a genocidal killer. No researcher has yet overcome this obstacle.

A more fruitful domain has been analysis of survivor testimony and of the reconstructions of generations who have come after (a field which overlaps with broader interests in memory, narrative, and representations of the past). Discussion of the relationship between trauma, personal memory, and reconstruction of the 'self' is still at an early stage in all the disciplines which touch on this area. One feature that is commonly pointed to in victim testimony is a flat, matter-of-fact tone adopted, a technique of distancing and dissociation. In innovative work in transcultural psychiatry, Laurence Kirmayer has developed a notion of *landscapes of memory* to characterize some of the differences in narratives of trauma between child abuse and holocaust victims: 'while holocaust stories involve bearing witness to what is widely ... recognized as a human catastrophe, personal stories of abuse are revelatory, shameful and damaging' (Kirmayer 1996, p.188). In this context memorialization may be a key feature of legitimating the experience of suffering. Work is, however, only just beginning on situations where memorialization does not take place (as for instance with most Gypsies in Eastern Europe).

One important new area of investigation is the social and political understandings and implications of particular ways of representing genocides. Early work in material culture on the activity of memory in holocaust memorials has shown how the meanings of memorials are contingent on social and political realities (Young 1993). At a literary level, work by Inga Clendinnen (1999) has opened up one path for a generation who no longer have access to direct testimony from survivors; more challengingly, Peter Novick's (1999) iconoclastic study of the Holocaust in US public life argues that some representations of genocide may act to inoculate a society against its own history and current world duties.

Apart from academic analysis, anthropologists have, at times, become centrally involved in acts of memorialization. In 1989, Polish Prime Minister Mazowiecki convened a commission to reconstruct the Auschwitz-Birkenau museum and monument for a postcommunist world. At the behest of the Polish Ministry of Culture, Jonathon Webber, an anthropologist with field research in Poland, and head of Oxford University's Centre for Hebrew Studies, invited an international body of experts to discuss 'the future of Auschwitz.' Alongside determining major issues of factual history—that 1.6 million had died there, and not four million as previously claimed, that 90 percent of those killed were Jews and not Polish and so on—the commission also addressed such tricky questions as to whether and how one should preserve

the ruins: reconstruction of the gas chambers was out of the question, but what sort of renovation should be permitted?

4. Postgenocide Societies?

Work on the representation of genocide is intimately connected with an interest in the afterlife of such massive social trauma. Many have suggested an association between German and Japanese urban terror in the 1960s with the phenomena of the 'second generation' but the link has never been established systematically. Now, however, a body of clinical work on individual responses to massive trauma is emerging. Imaginative work here often comes at the interface of disciplines, as in Lynne Jones's (2000) ethnographic-psychological examinations of how subjective understandings and political context mediate the impact of political violence on adolescent mental health in Bosnia. Jones, who is based at Cambridge University's Centre for Family Research, has, among other things, investigated children who had watched others being cleansed from their own town. In this case the 'therapeutic remedy' of 'the search for meaning' exposed children to uncomfortable truths about their community and appeared to lead to psychological discomfort, depression, and anxiety. Those children who avoided examining such events had better psychological health, but at the cost of leaving their elders' discourse unchallenged for a second generation.

For many victims of direct attack, the search for meaning, to name the unnameable is unavoidable. Here too, it is to be hoped, ethnographic work will emerge examining the impact of the International Criminal Tribunals. The president of the Yugoslav tribunal has sought to bring as many witnesses as possible to the Hague, so they can see their stories heard and entered in the register. But what effect has this had, and how will the expanding work of the Tribunal change the 'landscapes of memory' in Bosnia? It has been argued that significant use was made in 1990–2 of representations of experiences of World War II in cultivating an environment in which violence came to be seen as a serious option (Hayden 1994). Investigation of how societies deal with these kind of events is thus not merely of academic interest.

At an even broader level of analysis, anthropology would have much to contribute to the study of the experience of societies like Israel—for long unwilling to incorporate the genocide into its collective memory and now, having made it central to the nation, finding it hard to create a unified identity with Arab Jews who did not suffer. In the case of the perpetrator states the issue is almost invariably one of the social consequences of denial. Racist attitudes to Gypsies in Eastern Europe are likely related to denial and ignorance of Gypsy persecution in 1940–5. In this regard the German case is unique—a perpetrator state

that has accepted historic responsibility. German foreign policy will be constrained by the Nazi past for the foreseeable future and perhaps as long as there is a German political entity, but it is also clear that for the time being this is a society which cannot shrug off its past. Daniel Goldhagen's work struck a profound chord in Germany perhaps because it offered a new generation a fresh chance to stare at an unadulterated image of the heart of some of their ancestors' 'darkness.' At that very moment there was also anguish over the Berlin Holocaust memorial. One particularly radical proposal was to construct a bus station in central Berlin with timetabled buses taking visitors and pilgrims to and from the death and concentration camps—a reminder of just how close so many of these camps lay to the capital of the Reich and lie to the new capital of a reunified Germany. It was rejected for 'lack of taste' and necessary 'monumentality.' Such phenomena can provide sociologists and anthropologists with rich seams of cultural and social significance.

See also: Anti-Semitism; Collective Memory, Anthropology of; Ethnic Cleansing, History of; Ethnic Conflicts; Ethnocentrism; Ethnonationalism; Cultural Concerns; Genocide: Historical Aspects; Holocaust, The; National Socialism and Fascism; Public Commemoration; Racism, History of; Racism, Sociology of; Social Trauma; Xenophobia

Bibliography

- Bauer Y 1984 The place of the Holocaust in contemporary history. In: Frankel J (ed.) *Studies in Contemporary Jewry*. Institute of Contemporary Jewry, Hebrew University of Jerusalem; Indiana University Press, Bloomington, IN, Vol. 1, pp. 201–24
- Bloch M 1992 *Prey into Hunter: The Politics of Religious Experience*. Cambridge University Press, Cambridge, UK
- Bougarel X 1997 Bosnia and Hercegovina: State and communitarianism. In: Dyker D, Vejvoda I (eds.) *Yugoslavia and After: A Study in Fragmentation, Despair and Rebirth*. Addison-Wesley Longman, New York, pp. 87–112
- Clendinnen I 1999 *Reading the Holocaust*. Cambridge University Press, Cambridge, UK
- Conte E, Essner C 1995 *La Quete de La Race: Une Anthropologie du Nazisme*. Hachette, Paris
- Davis N Z 1973 The rites of violence: Religious riots in sixteenth century France. *Past and Present* 59: 51–91
- Fein H 1990 Genocide: A sociological perspective. *Current Sociology* 38(1): 1–126
- Goldhagen D J 1996 *Hitler's Willing Executioners: Ordinary Germans and the Holocaust*. Knopf, distributed by Random House, New York, London
- Hayden R M 1994 Recounting the dead: The rediscovery and redefinition of war-time massacres in late and post-communist Yugoslavia. In: Watson R (ed.) *Memory, History and Opposition under State Socialism*. School of American Research Press, Santa Fe, NM, pp. 167–84
- Herbert U (ed.) 2000 *National Socialist Extermination Policies: Contemporary German Perspectives and Controversies*. Berg-hahn, New York

- Hinton A L 1998 A head for an eye: Revenge in the Cambodian genocide. *American Ethnologist* 25(3): 352–77
- Jones L 2000 Adolescent understandings of political violence and their relationship to mental health: A qualitative study from Bosnia Herzegovina.
<http://www.waraaffectedchildren.gc.ca/socscimedpaper-e.asp>
- Kirmayer L J 1996 Landscapes of memory: Trauma, narrative and dissociation. In: Antze P, Lambek M (eds.) *Tense Past: Cultural Essays in Trauma and Memory*. Routledge, London, pp. 173–98
- Kuper L 1982 *Genocide: Its Political Use in the Twentieth Century*. Yale University Press, New Haven, CT
- Lemarchand R 1997 The Rwanda genocide. In: Totten S, Parsons W S, Charny I W (eds.) *Century of Genocide: Eyewitness Accounts and Critical Views*. Garland, New York, pp. 408–23
- Novick P 1999 *The Holocaust and Collective Memory: The American Experience*. Bloomsbury, London
- Sorabji C 1994 A very modern war. In: Watson H, Hinde R (eds.) *War: A Cruel Necessity: The Bases of Institutionalised Violence*. I.B. Tauris, London, pp. 80–95
- Spencer J 1990 Collective violence and everyday practice in Sri Lanka. *Modern Asian Studies* 24(3): 602–23
- Taylor C C 1999 *Sacrifice as Terror: The Rwandan Genocide of 1994*. Berg, Oxford, UK
- Young J E 1993 *The Texture of Memory: Holocaust Memorials and Meaning*. Yale University Press, New Haven, CT

M. Stewart

Genocide: Historical Aspects

Genocide is any intentional act to physically eliminate a non-combatant group in whole or in part by direct assault, or by inflicting deadly life conditions including measures to prevent biological, economic and cultural reproduction.

This article will outline the development of genocide research as well as the tasks of this young academic discipline (Sect. 1). It will, in Sect. 2, deal with the problem of applying the twentieth-century term genocide to a wide variety of mega-killings which have happened throughout history. The term genocide has generated a tendency not only to rewrite, but also to morally judge anew, historical events to settle recent political accounts. Groups whose ancestors have suffered massacres at a period in which neither they themselves nor their victorious opponents had drawn up or adhered to international legal standards outlawing such atrocities may yet label their past fate as genocide to attract attention to contemporary plights. The entry will try to circumvent this extra-scholarly or ideological use of the term genocide by implicitly or explicitly using a caveat of the sort that an event called genocide 'today' was not seen as such then. It has to be stressed that all figures used in this entry are estimates. There is no such thing as an undisputed figure for even the best researched genocides.

1. The Term Genocide and the Development of Genocide Research

Genocide is not only a historical concept but also an international crime defined by the 1948 United Nations *Convention on the Prevention and Punishment of the Crime of Genocide*. In its Article II

genocide means any of the following acts committed with intent to destroy, in whole or in part, a national, ethnical, racial or religious group, as such: (a) killing members of the group; (b) causing serious bodily or mental harm to members of the group; (c) deliberately inflicting on the group conditions of life calculated to bring about its physical destruction in whole or in part; (d) imposing measures intended to prevent births within the group; (e) forcibly transferring children of the group to another group.

The crimes mentioned in paragraphs (a) to (e) of Article II of the Convention are generalizations of practices employed by Nazi Germany after September 1, 1939 (beginning of World War II) against Slavs and/or Jews, and gypsies. The UN Convention was mostly written by a Polish lawyer and Holocaust refugee, Raphael Lemkin (1900–59), who had first in 1933 proposed a similar measure to the 'League of Nations.' In 1943, he had termed the crime in Polish as *ludobójstwo* (*lud* = people; *zabójstwo* = murder). A year later, he translated it into English as *genocide* (from Greek *genos* = people, and Latin *caedere* = kill, see Lemkin 1944, p. 79).

Since the UN Convention—today signed by some 140 nations—excluded 'politicide'—the murder of any person or people because of their politics or for political purposes—as well as 'ideological genocide'—the murder of any person or people because of their ideology, theory or knowledge—scholars have broadened the concept. Rudy Rummel's term 'democide' may be considered as the most comprehensive one. It is the murder of any person or people including genocide in the international legal sense, politicide as well as ideological mass murder (Rummel 1996, p. 39).

Besides the narrower term genocide and the more inclusive term democide the term 'crimes against humanity' is still in wide use. It was coined on May 24, 1915 by France, Great Britain, Russia, and the USA on behalf of the slaughtered Ottoman Armenians as 'new crimes of Turkey against humanity and civilization.' The term 'crimes against mankind' created, on October 12, 1942, in London (*Declaration of St. James*) to denounce the Nazi Holocaust of European Jewry is also still in use. Less known though quite appropriate is Gracchus Babeuf's French term 'populicide,' a shorthand for his 'système de dépopulation' executed—during the French Revolution—in the Vendée (1793/94) by killing its entire royalist peasantry of some 120,000 (Babeuf 1795).

The discipline of comparative genocide research draws on the entire corpus of historical research of the

last three millennia. Mega-killings have always attracted the attention of laymen and scholars alike. Yet, a first world history approach to genocide was tried by Lemkin who had drafted three volumes on the *History of Genocide* by 1959 (Jacobs 1999). Several publishing houses had turned him down. The manuscripts are only now being prepared for publication by Steven L. Jacobs.

The 1967–1970 massacres of half a million Ibos in Biafra/Nigeria triggered a new interest in the subject. In 1970, Yvan van Garse published *A Bibliography of Genocide Crimes against Humanity and War Crimes* (Garse 1970). The genocide, in 1971, of some 1.5 million citizens of East-Pakistan (today's Bangladesh) also has drawn a steady albeit not widespread attention to the subject. Both killings had world wide television coverage. Only a year later, in 1972, *The Twentieth Century Book of the Dead* by Gil Elliot became the first major published monograph devoted to comparative genocide research (Eliot 1972). In 1975, sociology entered the field with Dadrian (1975). Irving Louis Horowitz's *Genocide: State Power and Mass Murder* followed in 1976 (Horowitz 1997). In 1981, Leo Kuper reached a major academic publisher, Yale University Press, with *Genocide: Its Political Use in the Twentieth Century* (Kuper 1982). More systematically and historically minded works followed in 1990 with Fein's (1990) *Genocide: A Sociological Perspective* as well as Chalk's and Jonassohn (1990) *The History and Sociology of Genocide*.

The first—one person—institute devoted to the subject: 'Institute on the Holocaust and Genocide,' was founded in 1979 by Israel W. Charny in Jerusalem. Helen Fein followed in 1982 with the 'Institute for the Study of Genocide,' in New York. Next were Frank Chalk and Kurt Jonassohn with the Canadian 'Institute for Genocide and Human Rights Studies' (Montreal, Concordia University). Australia with the 'Centre for Comparative Genocide Studies' (Macquarie University), and Europe with the 'Raphael-Lemkin-Institut für Xenophobie- und Genozidforschung' (University of Bremen, Germany) followed in 1993. An international 'Association of Genocide Scholars' was established in 1995 with some 150 members by mid-2000. The first encyclopedia of genocide was published in Germany (Heinsohn 1999). A similar endeavor in English followed shortly afterwards (Charny 1999). Both works, with a combined volume of some 1,200 pages, cover most of the contents of this entry.

Historical, sociological, and sociopsychological studies of genocide aim at the assessment of genocidal potential as well as the formulation of means to prevent the crime. Interventions against ongoing genocides are rather dealt with in studies of warfare than in comparative genocide research. The prevention of genocide requires an early warning system. Historically, roughly 25 percent of high risk nations have transformed their perpetrator potential into

actual genocide (the rate is higher for democide). Since every genocide needs certain steps of preparation which are difficult to conceal, like, e.g., registering the victims and training the killers, all nations with perpetrator potential (today close to 100) may one day be scanned for these red alert signals indicating the imminence of a genocide.

2. History of Genocide

Because of its modern provenance it is difficult to employ the term genocide to the common periodization of history. Homer's *Iliad* (VI: 55–62), e.g., paints, as perfectly legitimate, an intention that today would be considered as genocidal in its full legal sense:

"My good Menelaos" [king of Sparta], said he [Agamemnon, king of Mycene], "this is no time for giving quarter. Has, then, your house fared so well at the hands of the Trojans? Let us not spare a single one of them—not even the child unborn in its mother's womb; let not a man of them be left alive, but let all in Ilios [Troy] perish, unheeded and forgotten." Thus he did speak and his brother was persuaded by him, for his words were just.

Moreover, a practice of killing considered the expectable as well as acceptable outcome of conflict in an early period of history may continue to this very day. As an example in question one may look at the Yanomami Indians of Southern Venezuela and Northern Brazil. They must—since 1980—be considered a possible victim of genocide by modern settlers and—at the very same time—were successful in nearly wiping out a neighboring tribe (Keegan 1993, p. 97).

2.1 Tribal Societies

Genocidal acts in the modern sense were hardly distinguishable from acts of war in tribal societies—covering the longest stretch of human history—because there was no separation of noncombatant civilians and warriors. Warfare was "ecological" in motivation' and led to the redistribution of hunting and fishing grounds or arable land 'from the weak to the strong' (Keegan 1993, p.101). Means of these wars were 'battles of annihilation' or, if space allowed, 'displacement of the weaker' (Keegan 1993, pp. 29, 387). Genocidal tribal raids continued well into the second half of the twentieth century in territories of the Amazon, New Guinea, the Philippines and elsewhere: 'Typically the goal in such raids is to destroy an entire village or settlement, the common intention being to kill everyone, perhaps excepting a few of the younger women and children who are carried off for incorporation into the home group' (Glick 1994, p. 45).

2.2 Early Antiquity

Early stages of civilization (Early and Middle Bronze Age with priest-kingship feudalism) saw tribal practices of annihilation on a larger scale: 'I slew them without sparing them, they sprawled before my horses, and lay slain in heaps in their blood,' Egypt's Pharaoh Ramses II (conventionally dated 1279 to 1212) boasted after a campaign against the Hittites whose leader, then, begged him for mercy: 'Look, you spent yesterday killing a hundred thousand and today you came back and left no heirs. Be not hard in your dealings victorious king' (Lichtheim 1976, pp. 70 f.). Other powers of the Ancient Near and Far East behaved very much in the same way, as did Bronze Age, i.e., Mycenaean Greece.

2.3 Antiquity/City States

The annihilation practices of time immemorial were cast into a principle by Athenians. When, in the Peloponnesian War (431 to 404 BC), Athens had decided to wipe out Sparta's ally Melos, the admirals in charge tried to formulate an eternal law underlying their action:

Our knowledge of men leads us to conclude that it is a general and necessary law of nature to rule wherever one can. This is not a law that we made ourselves, nor were we the first to act upon it when it was made. We found it already in existence, and we shall leave it to exist for ever among those who come after us. [...] We know that you or anybody else with the same power as ours would be acting in precisely the same way. [...] The Melians [in 415 BC] surrendered unconditionally to the Athenians, who put to death all the men of military age whom they took, and sold the women and children as slaves (Thucydides, *The Peloponnesian War*, (1954 trans) V: 85 ff.).

As a first step away from mutually slaughtering their males, the Greeks developed the practice of exchanging prisoners of war.

Within the Roman Empire—a private ownership based society on a much larger scale—the practice of destroying enemies who had caused problems in the past was continued. The elimination of Corinth (metropolis of mainland Greeks) and Carthage (metropolis of Phoenicians) in 146 BC, of Numantia (capital of Spanish Celtiberians) in 134 BC, and the elimination of the Etruscans (92 BC) and the Samnites (82 BC) as major opponents in Italy proper provide typical examples with 50,000 to 150,000 dead in each case. Of course, the opponents acted in the same manner as, e.g., in the slaughter of the inhabitants of the Greek cities Selinus and Himera (Sicily) in 409 BC by the Phoenicians of Carthage. In 88 BC all Romans and Italians, including women and children, living in Greek Asia Minor were put to the sword by orders of Mithradates VI. The Romans of Britannia met a similar fate in 61 BC. Within the Roman Empire a *ius gentium* was developed for strangers who did not hold

citizenship with, sometimes, more protection for these minorities than for full citizens.

Parallel to practices of the Greek City States and the Roman Empire, movements against blood sacrifice emerged around 600 BC. Well known are Buddhism in India and the Pythagoreans of Classical Greece. It was, however, monotheistic Judaism whose moral code—albeit slowly—came to influence international ethics most deeply. To overcome expectations of salvation tied to the slaughter rituals of blood sacrifice, Hebrew prophets found a new moral vision: 'I desired mercy, not sacrifice; and the knowledge of God more than burnt offerings' (*Hosea* 6:6). The sanctity of life became the highest principle: 'Thou shalt not kill' (*Exodus* 20:13 and *Deuteronomy* 5:17) including—in its universal character—the strict outlawing of infanticide.

The Torah—rather a moral compilation than a historical record—emphasizes the sanctity of life as the core of the law and as a value identical with goodness ('the good law'): 'See, I have set before thee this day life and good. [...] I call heaven and earth to record this day against you, that I have set before you life and death, blessing and cursing: therefore choose life, that both thou and thy seed may live' (*Deuteronomy* 30:15, 19). From the principle of life's sacredness derived other commandments like 'Thou shalt love thy neighbor as thyself' (*Leviticus* 19:18). Anti-xenophobic teaching followed suit: 'And if a stranger sojourn with thee in your land, you shall not vex him. But the stranger that dwelleth with you shall be unto you as one born among you, and thou shalt love him as thyself' (*Leviticus* 19:33–34). 'Are ye not as children of the Ethiopians unto me, O children of Israel? Saith the Lord. Have not I brought up Israel out of the land of Egypt, and the Philistines from Caphtor, and the Syrians from Kir' (*Amos* 9:7).

Against Judaism, the first recorded order (of Haman) for an ideological, though abortive, genocide (also called religiocide) was issued in the Akhaemenid Empire of fourth century BC: 'And the letters were sent by posts into all the king's provinces, to destroy, to kill, and to cause to perish all Jews, both young and old, little children and women, in one day' (*Esther* 3:8/13). Christians who had adopted the Jewish rejection of blood sacrifice to planetary deities actually suffered the first well recorded religiocides. The edict of Emperor Decius of AD 250 forced them to eat sacrificial meat or suffer death. Four more edicts by Diocletian from AD 303/304 repeated the order upon pain of death. The massacres ended in AD 312 when Constantine the Great converted to the new faith.

2.4 Middle Ages

The Middle Ages saw attempts at a *Pax Romana*, directed by the Papal See. Its concept of a *Pax Dei* (peace of God)—developed since AD 975 and extended over all Christian lands at the Synod of Clermont

(AD 1095)—outlawed war from Wednesday night to Monday morning as well as on Christian holidays. Excommunications and interdicts were the only means of enforcement. It was the beginning of international law based on the Judeo-Christian code of ethics. Yet, the transgressions were severe and, in modern terminology, definitely genocidal.

The crusades—starting in AD 1096—not only brought massacres of thousands of Jews in Germany and tens of thousands of Jews and Muslims in Jerusalem (AD 1099), but also of fellow Christians. Against Pope Innocent III's interdict, Dalmatia's port Zadar, arch rival of Venice, was bathed in the blood of all its inhabitants in November 1202 during the fourth Crusade by French and Italian troops transported on Venetian ships. The same men butchered 2,000 inhabitants of Greek Constantinople on April 13, 1205 before they went on a three-day-rampage of indiscriminate slaughter, not sparing women and children.

Since 1167, the Cathars (from Greek *katharoi* = pure), called Albigensians after the French city of Albi, suffered a series of religiocides by Catholic forces, because they rejected any spilling of blood and did not believe in the resurrection of Christ, and the Holy Communion as the residue of blood sacrifice within Christianity. By 1234 the Cathars were wiped out with the massacre of Béziers as the largest slaughter of this crusade in which all 15,000 inhabitants were killed.

A few years later, in AD 1211, Gengis Khan's Mongols (Morgan 1986)—with no authority to arouse their consciousness by threats of excommunication or invoking the image of all men as protected children of the one God—began their terrifying campaigns from the Japanese to the Baltic Sea. To turn their fertile areas into horse pastures the Mongols have, up to AD 1290, eliminated some 25 million Chinese. Another 5 million perished in Persian, Arab, and Christian territories where the Steppe warriors revived the memory of Attila's fifth century Huns. Behind their lines, the Mongols used to kill everyone perceived as a future threat, i.e., all males not suitable to join their ranks. Defeated armies were allowed a brief lease to life by serving as shields or trail blazers for advancing Mongol troops. Tamerlane, a pious Muslim, supposedly had, in a single day of 1398, some 100,000 Hindus strangled in sight of the walls of Delhi which had rejected surrender. In 1402, he dealt a mortal blow to Arab civilization by killing 90,000 inhabitants of Baghdad.

2.5 Modern Period

A sovereign's right to war (*ius ad bellum*) dominated the period of sixteenth to nineteenth century in which Europe came to dominate the world. What happened within a war was more or less up to the sovereign. Laws of warfare or the definition of crimes of war (*ius in bello*) had to wait for the turn of the nine-

teenth/twentieth century. In Europe as well as overseas, the results were genocidal. To give only two examples, Tenochtitlan, the Aztec capital in the valley of Mexico with its 100,000 inhabitants, was totally annihilated by Cortes' Spaniards in 1521. All 30,000 inhabitants of German Magdeburg, the richest city of Protestantism, were slaughtered on May 10, 1631 by Catholic troops under General Tilly in the Thirty Years' War.

The European Population Explosion—beginning at the end of the fifteenth century, and still not fully understood—bestowed this small continent with 25 percent of world population by 1900. During the same period, some 60 million young Europeans, with the superior military technology of their ownership-based economies but no attractive positions in an overcrowded continent, conquered their settlement space in the Americas, Australia, Africa, and Northern Asia. In the wake of this 'emigration,' a population of 55 to 85 million native Americans estimated for 1492 (Spanish conquest) was reduced by 85 to 95 percent around 1650. Between 250,000 and 500,000 native Australians estimated for 1788 (British conquest) were reduced to 60,000 by 1921. All means of killing were employed from outright genocide to infections—including intentional ones. Africa, between 1500 and 1870, lost up to 20 million inhabitants into American slavery with 2 million dying on transport. The slave trade within Africa may have cost, for the same period, the lives of 4 million. On the slave routes to—mostly Islamic—Asia another 11 million may have perished in these gigantic democides (Rummel 1996, p. 48).

Though not an invention of modern times, politicide became one of the most heinous activities of this period. It does not necessarily include the killing of the families of the victims. However, in a so-called 'Vespers' in which politicide is turned against a foreign elite, it usually grows into a full blown genocide. It took its name from the Easter massacre in Palermo on March 29, 1282, at the hour of evening prayer (Vespers) when the French in control of the Kingdom of Sicily were killed with their families ('Sicilia Vespers'). It had been preceded, in 1268, by the butchering of some 5000 members of the Hohenstaufen elite of Sicily by the hands of the French conquerors. The best remembered example of politicide was the annihilation, after eight inconclusive civil wars, of the Protestant leadership of France (ca. 100,000 victims), starting on Saint Bartholomew's Day (August 24, 1572) when the Huguenots—easily recognizable by their black clothing—were guests at the royal marriage of Henri of Navarra and Marguerite of Valois. The politicide, in February 1579, against Novgorod, Russia's richest city, by Ivan IV (the Terrible) was peculiar insofar as the Tsar had put together special killing units—the 'Oprichniki'—because the army was not trusted to act ruthlessly enough. In the 'Ulster, Massacre' (October 23, 1641), as the last 'Vespers' of

Table 1

Rough estimates of numbers of victims by genocides, politicides and ideological mass murder of a wide variety of regimes usually lumped together under the term Communism (Rummel 1996).

Soviet Union, 1917–87	40–60,000,000
Red China plus Maoists before 1949	40–60,000,000
Khmer Rouge in Cambodia, 1975–79	2,000,000
North Korea, 1945 onwards	1,700,000
Yugoslavia of Tito, 1941–87	1,100,000
North Vietnam, 1945–87	1,700,000
Poland (deportations of Germans), 1945ff.	1,400,000
Czechoslovakia (deportations of Germans), 1945ff.	240,000
Ethiopia, 1974–91	4,000,000
Albania, 1944–90	100,000
Communist Germany 1945–52	96,000

Western Europe, Irish rebels killed most of the Protestant British settlers and administrators on their island. The Creoles of Haiti delivered their French masters a ‘Vespers’ in 1791/1805 as did, in 1857/58, albeit with no lasting success, the sepoy to the British masters of India.

2.6 Twentieth Century

The Hague Conventions of 1899 and 1907 with *The Regulations Respecting the Law and Customs of War on Land* prohibited, in Articles 22–8, the murder, ill-treatment, or deportation of enemy civilians and prisoners of war. The use of poison and special ammunitions was banned. However, a nation’s harm to its own civilians was not taken into consideration. This had to wait for the ‘Istanbul Trials,’ installed in January 1919 to inflict punishment on those responsible for the Armenian massacres. On April 10, 1919, Mehmet Kemal Bey, procurator of the province Bogazliyan, became the first person executed for genocide, then called crime against humanity. In the Nuremberg Trials of Nazi-Germany’s leadership (1945/46) as well as in the Tokyo Trials of Imperial Japan’s leaders (1946–48), ‘International Military Tribunals’ further developed the legal tools to punish perpetrators of crimes against humanity. Special United Nations courts explicitly devoted to the crime of genocide started, in 1993, to prosecute the perpetrators in Yugoslavia (‘ethnic cleansings,’ since 1991) and, a year later, in Rwanda (genocide of Tutsis in spring 1994). The statute for a permanent ‘International Criminal Court’ was passed in Rome on July 17, 1998.

The entire twentieth century is on record for close to 180 million victims of democide according to Rummel’s statistics. The extermination of the Ottoman Armenians (1894–96, 1909, 1915/16, 1919, 1923)

combined strategic, racial, and ideological intentions (Dadrian 1995). A shrinking empire was to be consolidated by a homogenous pan-Turkic population within an undivided settlement space under Islam as the only religion. That’s why not only more than 1 million Armenians suffered genocide but also other Christian minorities like Maronites (1860), Assyro-Chaldaeans (1915), and Greeks (1915–23) with a combined death toll of some 440,000.

Since 1917, single party totalitarian regimes with agendas may have killed up to 110 million people (Rummel 1996, Courtois 1997). After World War II, ‘one-party communist states are 4.5 times more likely to have used genocide than are authoritarian states’ (Fein 1933b, p. 79)

The figures given in Table 1 are all estimates. Totalitarian regimes are not only known for most of the killings of the twentieth century, but also present the most difficult problems for research because archives were destroyed and many deeds never recorded. Thus, totalitarian regimes have to live with possibly exaggerated figures of their atrocities.

Extreme anti-capitalism slowed economic development though superior prosperity was promised. The shortcomings were frequently blamed on sabotage and/or party comrades still infested with the reactionary spirit of ownership. This view was used to justify massive party purges like the ‘Great Terror’ in the Soviet Union (1934, 1936–38) with more than 7 million or the ‘Cultural Revolution’ in China (1964–75) with nearly 8 million dead.

Annihilation through forced labor in camps was initiated by Trotzki on June 26, 1918. By 1922, 23 Soviet camps were at work, forming the nucleus of Stalin’s Gulag fully emerging in 1928, and eventually running some 8000 camps. Working people to death became, in the long run, the most devastating killing tool of history. It allowed the elimination of opponents by simultaneously using them for the creation of big projects—like canals, railroads, subways, etc.—to prove by ‘miracles’ the promised productive superiority of communism that otherwise was wanting.

The traditional siege tool of man-made famines had to wait for the twentieth century to be welded into the most efficient fast-killing techniques ever. To crush national ambitions of the Ukraine, for example, a famine was turned into the mass death of some 7 million in the single winter (4 months) of 1932/33. Ethiopia’s Derg managed to kill up to 4 million peasants between 1982 and 1986 by the same means, often intentionally aggravating natural disasters.

The mega-killings by left totalitarian regimes came as a combination of politicides (elimination of like minded rivals for power), genocides (elimination of minority nations, mostly in the Soviet Union and the Peoples Republic of China), and ideological genocides (foremost the elimination of proprietors or property-infested citizens). The latter category probably became the largest victim group in history with some 50

million dead (55 percent in China, 40 percent in the Soviet Union, 5 percent in the other communist regimes).

Of all twentieth-century totalitarian regimes, Nazi Germany was the most genocidal. Following September 1, 1939, deformed newborn and mentally-handicapped Germans—as well as some German soldiers severely wounded in the military campaign against Poland—were killed by injections or in gas chambers. Close to 130,000 handicapped eventually ended in this so-called euthanasia operation. Later exterminations struck Gypsies, stigmatized as socially inferior, as well as non-reproductive homosexuals etc. A bigger target, however (outlined in the ‘*Generalplan Ost*’ (Masterplan East)) were 100 million Slavs (Czechs, Poles, White-Russians, Ukrainians, and Russians) living between Germany’s Eastern border and the Ural Mountains, a territory marked to become Germany’s ‘*Lebensraum*.’ Of the Polish elite, 75,000 members, most of them holders of academic degrees, were killed in autumn 1939. The second Eastern campaign, against the Soviet Union (June 22, 1941), was aimed at the immediate killing of 30 million Slavs. Another 30 million were destined to be worked to death to lay the infrastructure for German settlements. The remaining 40 million were to be driven into Siberia with a calculated death toll of another 10 million. With 13 million Slavic civilians and prisoners of war killed by 1945 (Kumanev 1990, Vitvitsky 1990), Nazi Germany had fallen short of its 70 million goal but still had committed the largest twentieth-century genocide explicitly directed against a particular ethnic group. Nazi Germany’s Eastern campaigns, thus, were not only massive wars but also militarily enforced genocides (Hillgruber 1993).

In connection with the Eastern wars, 5 to 6 million Jews were killed. Though a ‘theory of the Holocaust is lacking’ (Herbert 1998, p. 66), Hitler’s practice of ‘strengthening’ Germany by killing its handicapped and of gaining settlement space through genocide of its Slavic inhabitants put him inevitably against the Torah. To him the sanctity of life was a Jewish ‘invention,’ a weakening ideology that could be cleansed from the German sphere of influence by exterminating its ‘carriers.’ The Holocaust, in this perspective, was a ‘genocide for the purpose of reinstalling the right to genocide’ (Heinstohn 2000, p. 424).

Since 1944/45, some 100 acts, fulfilling the definition of the UN Genocide Convention have taken place. In the two decades from 1980 to 2000, some 80 distinct ethnic and religious groups—mostly in Africa, Asia, and Latin America—were victimized with 12 of them going through full blown genocide: Bosnian Muslims, Kurds, Tibetans, Dinka (Sudan), Nuba (Sudan), Shilluk (Sudan), Isaaq (Somalia), Tutsi (Rwanda), Mayans (Guatemala), Albanians (Serbian Kosovo), and Timorese of Indonesia (Harff and Gur 1996, updated by the author).

2.7 The Future

Ecologically and ideologically motivated genocides most probably will have a future in the twenty-first century. Politicides may also occur wherever solidly established democratic structures are missing. Democracies usually do not go to war against each other. The same cannot be said for democide like nuclear and carpet bombing against totalitarian powers with, e.g., a death toll of 1.5 million in Germany (600,000) and Japan (900,000) during World War II.

The prevention of genocide by spreading democracy remains the preeminent goal for the future. Meanwhile, some experience with intervention has been made, in 1999, in Kosovo and Timor. To preempt the cruelties unavoidable in an intervention by force, genocide early-warning stations are now seriously considered. They may be in place within a few years.

See also: Communism; Ethnic Cleansing, History of; Ethnic Conflicts and Ancient Hatreds; Cultural Concerns; Eugenics, History of; Genocide: Anthropological Aspects; Holocaust, The; National Socialism and Fascism; Religion: Nationalism and Identity; Religion: Peace, War, and Violence; Totalitarianism; Violence, History of; War Crimes Tribunals; Xenophobia

Bibliography

- Babeuf F N 1795 *Du système de la dépopulation ... avec la guerre de la Vendée*. Franklin, Paris
- Chalk F, Jonassohn K 1990 *The History and Sociology of Genocide. Analysis and Case Studies*. Yale University Press, New Haven, CT
- Charny I W (ed.) 1999 *Encyclopedia of Genocide*. ABC-CLIO, Santa Barbara, CA
- Courtois S (ed.) 1997 *Le Livre Noir du Communisme: Crimes, terreur, répression*. Laffont, Paris
- Dadrian V N 1975 A typology of genocide. *International Review of Modern Sociology* 5: 201–12
- Dadrian V N 1995 *The History of the Armenian Genocide: Ethnic Conflict from the Balkans to Anatolia to the Caucasus* (1995), Berghahn. Providence, RI
- Elliot G 1972 *Twentieth Century Book of the Dead*. Allan Lane/The Penguin Press, London
- Fein H 1993a *Genocide: A Sociological Perspective*. Sage Publications, London
- Fein H 1993b Accounting for genocide after 1945: Theories and some findings. *International Journal on Group Rights* 1: 79–106
- Garsse Y v 1970 *A Bibliography of Genocide, Crimes Against Humanity and War Crimes*. Studiecentrum voor Kriminologie en Gerechtelijke Geneeskunde, Sint Niklaas Waas, Belgium
- Glick L B 1994 Religion and genocide. In: Charny I W (ed.) *The Widening Circle of Genocide. Genocide: A Critical Bibliographic Review*. Transaction, New Brunswick, NJ, Vol. 3, pp. 43–74
- Harff B, Gurr T 1996 Victims of the state: Genocides, politicides and group repression from 1945 to 1995. In: Jongman A J (ed.) *Contemporary Genocides: Causes, Cases, Consequences*. PLOOM. Leiden, The Netherlands, pp. 33–58

- Heinsohn G 1999 *Lexikon der Völkermorde*, 1st edn. 1998, Rowohlt, Taschenbuch Verlag, Reinbek bei Homburg, Germany
- Heinsohn G 2000 What makes the Holocaust a uniquely unique genocide. *Journal of Genocide Research* 2: 411–30
- Herbert U (ed.) 1998 *Nationalsozialistische Vernichtungspolitik 1939–1945: Neue Forschungen und Kontroversen*, Fischer, Frankfurt, Germany
- Hillgruber A 1993 *Hitlers Strategie: Politik und Kriegsführung 1940–1941* [1963]. Bernard & Graefe, Bonn, Germany
- Horowitz I L 1997 *Taking Lives: Genocide and State Power*. Transaction, New Brunswick, NJ
- Jacobs S L 1999 The Papers of Raphael Lemkin: A First Look. *Journal of Genocide Research* 1: 105–14
- Jonassohn K, Björnson K S 1998 *Genocide and Gross Human Rights Violations in Comparative Perspective*. Transaction, New Brunswick, NJ
- Keegan J 1993 *A History of Warfare*. Knopf, New York
- Kumanev G A 1990 The German occupation regime in occupied territory in the USSR (1941–1944). In: Berenbaum M (ed.) *A Mosaic of Victims: Non-Jews Persecuted and Murdered by the Nazis*. New York University Press, New York, pp. 128–41
- Kuper L 1982 *Genocide: Its Political Use in the Twentieth Century*. Yale University Press, New Haven, CT
- Lemkin R 1944 *Axis Rule in Occupied Europe*. Carnegie Endowment for International Peace, Washington, DC
- Lichtheim M 1976 *Ancient Egyptian Literature. Volume II: The New Kingdom*. University of California Press, Berkeley, CA
- Morgan D 1986 *The Mongols*. Blackwell, Oxford
- Rummel R J 1995 *Statistics of Democide: Genocide and Mass Murder Since 1900*. Center for National Security Law, Charlottesville, VA
- Rummel R J 1996 *Death by Government 1994*. Transaction, New Brunswick, NJ
- Thucydides 1954 *The History of the Peloponnesian War*, trans. R Warner. Penguin Books, Harmondsworth, UK
- Vitvitsky B 1990 Slavs and Jews: Consistent and inconsistent perspectives on the holocaust. In: Berenbaum M (ed.) *A Mosaic of Victims: Non-Jews Persecuted and Murdered by the Nazis*. New York University Press, New York, pp. 101–8

G. Heinsohn

Genotype and Phenotype

The fundamental distinction between genotype and phenotype was first introduced into biology by Wilhelm Johannsen in 1911 to make explicit a distinction that was implicit in Mendel's explanation of heredity, and which marked it off from previous theories of inheritance. Mendel found that when he crossed two pure-breeding varieties of garden pea, the first generation of hybrid offspring were indistinguishable from one of the two parental strains, but if these hybrids were crossed with each other, both of the original types appeared in the offspring of these hybrids. So, for example, a pure bred red-flowered crossed with a pure bred white-flowered form produced all red-flowered plants, but these, when

interbred with each other produced both red- and white-flowered plants.

In order to explain this extraordinary result, Mendel postulated the existence of internal, unseen 'factors' contributed by each parent to the offspring, factors that were causal determinants of the outward appearance of the plants, but whose nature was unaffected by their passage from one generation to the next. So, pure bred white-flowered plants had two white 'factors' one contributed by the pollen and one by the egg of their male and female parents, while pure bred red-flowered plants had received two red 'factors' from their two parents. The hybrids of red-flowered and white flowered forms would then contain one red and one white 'factor,' yet these plants were indistinguishable in appearance from a parent that had two red 'factors.' There was then a fundamental distinction between the state of the internal causal 'factors' and the outward state of the plant that carried them. Moreover, to explain the reappearance in the next generation of some white-flowered plants, Mendel had to suppose that the temporary coexistence of red and white 'factors' in the hybrids had no effect on their nature, but that they could be passed on uncontaminated to the offspring of the hybrids. An offspring of the hybrids that happened to contain two white 'factors' was then white like its white grandparent.

This scheme of explanation makes a radical separation between the internal factors that behave according to one set of physical relations in their passage between generations, the laws of *heredity*, and the external appearance of the organism that results from a different set of physiological relations which determine the consequences to the organism of carrying those factors, the laws of *development*. This distinction between the phenomenon of heredity and the phenomenon of development which we usually associate with August Weissman's separation of the germplasm from the somatoplasm had already been made by Mendel twenty years earlier. Earlier theories of heredity had not made such a separation of inner factors from outward manifestation and had been posed as the problem of the inheritance of the organism's characters themselves. It was Mendel's realization that characters are not inherited, they develop. In Johannsen's terminology the description of an organism's internal factors that were passed on to it in the gametes from its parents is a specification of the organism's *genotype*, while the description of the total physical and behavioral state of the organism is a specification of its *phenotype*.

1. Description of Genotypes and Phenotypes

As a result of the discoveries in molecular and cellular biology since Johannsen's time genotypic and phenotypic descriptions have become more concrete, more detailed and greatly amplified in their scope. Originally genotypes had no independent physical descriptors

but were distinguished from each other entirely on the basis of the external phenotypic manifestations and the patterns of similarity in phenotypes of related individuals. 'Genes' were simply abstract placeholders of unknown physical composition, inferred entities that could be characterized only by their phenotypic effects.

The modern identification of DNA as the physical material that behaves in heredity in the manner inferred for Mendel's mysterious 'factors' has allowed a precise physical description of the genotype of an organism. It is the complete spatially-ordered sequence of the four alternative nucleotide bases, adenine, cytosine, guanine and thymine contained in the DNA of the organism's cells. In its most complete form this genotypic description includes not only the DNA in the nucleus of the cell but in certain cellular organelles such as mitochondria or chloroplasts that carry the DNA that is related to their physiological function and that is passed across generational lines. Phenotypic description has also been greatly amplified beyond the external characteristics of individuals. It now includes the gross structure of the various cells of the organism, the physico-chemical description of the various molecules within cells, including their shape and intra-cellular localization, and the network of biochemical reactions in which these molecules participate.

The concepts of phenotype and genotype require a distinction between *types* and *tokens*. Phenotype and genotype are, as the words indicate, types, sets of which individual organisms are members, sets delineated by the complete descriptors of those organisms. Any individual member of these type sets is a token of the type. In addition it is necessary to distinguish the descriptors of organisms from the physical object or behavior that is being described. The concepts phenotype and genotype refer to the descriptors. The physical objects that are being described are the *phenome* and *genome* of an organism, the actual physical material or behavior of some particular individual and its actual DNA molecules. If we take the definitions literally, then no two organisms belong to the same phenotype because there is always some physical difference in morphology, physiology and behavior between any two organisms, even between identical twins, cloned individuals, or cells that result from asexual reproduction. Moreover, except for the results of twinning, cloning or vegetative reproduction, no two individuals belong to the same genotype, as the process of mutation guarantees some variation in their DNA between any two individuals.

The rate of mutation per nucleotide per cell division is of the order of one in a billion, so for mammals, with over a billion nucleotides of DNA making up their genes, even identical twins are likely to have some small difference in total genotype. There is, in addition, an ambiguity in the assignment of an individual to a genotype because mutations occur during the millions

of cell divisions that occur in development, so that all cells in the body do not have identical DNA.

In practice phenotypic and genotypic descriptions are not total but partial, referring to some aspect of the genome and phenome that are of particular interest, while ignoring the rest. The partial phenotypic and genotypic descriptions are partial in two senses. First, only some aspect of the morphology or function of the organism is characterized; say the rate of enzymatic splitting of alcohol, a phenotype, and the DNA sequences that code for the enzyme proteins involved in the reaction, a genotype. Second, a decision must be made about the collection of actual partial phenotypes and genotypes that will be regarded as indistinguishable. At the phenotypic level, given the actual variation in enzymatic rates from individual to individual and from time to time, what range of rates will be regarded as belonging to the same phenotypic class? For any phenotypic variable that is continuous, some boundaries of distinguishable phenotypic classes must be laid down in order to make a mapping between phenotype and genotype because genotypes, by their nature, are discrete classes.

The problem of aggregation exists at the genotypic level as well. The relation of DNA differences to differences in the amino acid sequence of a protein is not one-to-one. Because the DNA triplet code is redundant, substitutions of nucleotides in the third position of the triplet often do not change the amino acid that is specified, so for some purposes many different alternative DNA sequences of a gene count as the same genotype. On the other hand, nucleotide substitutions that do not affect the amino acid sequence of a protein may affect the rate of synthesis of that protein and so may, for some questions, be counted as belonging to different genotypes.

The decision about which aspects of total phenotype and genotype are relevant in creating genotypic and phenotypic classes is a deep question of biology that is largely ignored because everything cannot be considered at once. Sometimes biologists act as if the rest of the organism really were constant, sometimes they recognize the variation but claim that it is causally irrelevant to the phenomenon of interest and sometimes they admit its causal relevance but claim that the variation can be treated as background experimental noise that can be averaged out by repeated observations on a large class of individuals. These decisions are usually made either for the sake of observational convenience without a coherent rationale or from a commitment to a reductionist view of the organism that asserts that an understanding of the whole system can be obtained by breaking it down into 'obvious' small components to be studied in isolation.

2. Relationship Between Genotype and Phenotype

It is clearly true that organisms do indeed consist of a collection of quasi-independent genotypic and pheno-

typic subsystems, within which there are strong interactions and between which there are very weak interactions. But the boundaries of these genetic and phenotypic subsystems can only be found by actual experiment and differ from function to function. The hand and eye are, as far as is known, developmentally independent systems for which different gene combinations are relevant during embryogenesis, but hand-eye coordination is essential for reaching out and picking up observed objects and so cannot be regarded as phenotypically or genotypically independent in considering this elementary behavior.

The problem for genetics and development is how the genotypic and phenotypic spaces map into each other. It is now certain that Weissman was correct in claiming that the genotype-phenotype mapping is causally asymmetrical. Genotypic differences may cause phenotypic differences during the process of development, but there is no material pathway by which changes in the phenome can result in related changes in the genome. There is no biological inheritance of acquired characters. This causal asymmetry sets the program for developmental, physiological and behavioral genetics which is to discover the mechanisms by which the reading of DNA by the cell machinery is causally efficacious in influencing the phenotypic state of the organism. This is the problem of making the forward mapping of genotype on to phenotype.

There remains, then, the problem of the backward mapping, how the causal knowledge of epigenetic processes can be used to infer the genotype, given the phenotype. Even if a knowledge of the genotype gave perfect information about the phenotype, the inverse would not be true, as discovered by Mendel. The phenomenon of dominance alone makes it impossible to distinguish between heterozygotes and homozygotes from phenotypic information at an arbitrary level. In addition, the same phenotype may appear as a result of changes at different locations in the DNA. There are, for example, more than 200 different nucleotide substitutions known in the genes coding for blood-clotting proteins, any one of which leads to hemophilia. The major difficulty of mapping between genotypes and phenotypes is, however, in the forward direction. The complete DNA sequence of an organism does not contain the information necessary to specify that organism.

3. Genotype, Phenotype, and Environment

First, the outcome of development and the ongoing functioning of an individual is the consequence of a unique interaction between the molecules contained in the fertilized egg, including the DNA, and the sequence of environments in which development occurs. Figure 1 shows the result of a cloning experiment in a plant, *Achillea*, grown at three different elevations in the

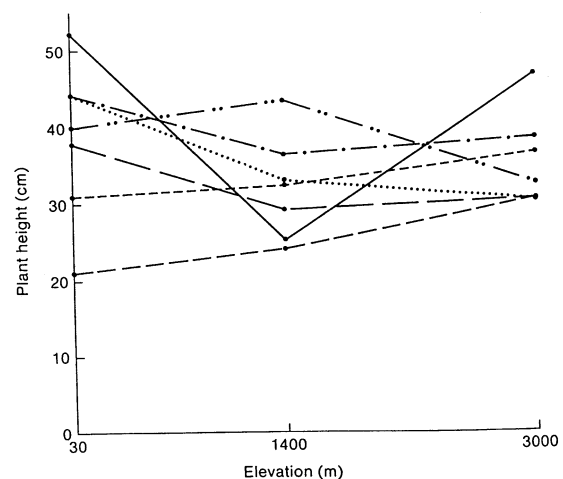


Figure 1

The growth of cloned plants of *Achillea* at three different elevations. The ordinate shows final the plant height and the abscissa the elevation. Each connected line is the pattern of growth for a different cloned genotype. Data of J. D. Clausen, D. Keck and W. W. Hiesey, 1958

Sierra Nevada. Immature plants were sampled from nature, each plant was cut into three pieces, and one piece was replanted at each of the three elevations. The figure shows the height to which the pieces grew in the different environments, each line connecting the three genotypically identical pieces in the three environments. A line, giving the phenotype of a given genotype as a function of different environments is a *norm of reaction*. The norms of reaction in this case show that there is no genotype with unconditionally better growth and, in fact, there is no way to predict the order of growth of the different genotypes in one environment from their order of growth in another environment. Thus, the phenotype is not only contingent on the environment but in an unpredictable and irregular way.

Such irregular and particularistic relations between genotype and phenotype in different environments is characteristic of norms of reaction whenever they have been measured for morphological and behavioral characteristics in experimentally accessible organisms. There are, of course, extreme exceptions; genotypes that result in the same phenotype in any environment and environments that are in no way contingent on genotypes for their effect. Eye color variation among humans is entirely a consequence of genotypic differences, uninfluenced by environment, while regional and social class differences in the phonemic structure of speech have no influence of genetic differences. Moreover, major differences between species are a consequence of genetic differences that are manifest in

all environments. There is no environment in which a gorilla could learn to write this article. Within species, however, the vast majority of phenotypic variations in morphology, physiology, and behavior are a consequence of particular contingent relations between genotype and phenotype.

4. Developmental Noise

Second, even a complete specification of the genotype and the environment is insufficient to specify the organism. All cells contain a very small number of copies of biologically active molecules, spatially segregated within the cell, so that the number of various reactions per unit time vary greatly from cell to cell and time to time. Moreover, when cells divide they do not distribute the molecules exactly evenly, except for DNA which has a special machinery for guaranteeing its equal distribution. The consequence of these small molecular numbers is that there is random variation from cell to cell in time to the next cell division, in cell size, in cell shape and in the movements of cells during development. These random events, random at the level of thermal and quantum chemical molecular states, can have considerable effect on the formation and behavior of the whole organism. (Goss and Pecoud 1998)

So-called 'bilateral symmetries' are not, in fact, symmetrical but vary between sides in an irregular way from individual to individual. The fingerprints on the left hand are not identical, although similar to, those on the right. Such *random developmental noise* is a nontrivial source of variation in phenotype. During the development of the central nervous system of a multicellular organism, superimposed on the major regular patterns of neuronal connection are random variations in the growth and connections of specific neuronal pathways, even in organisms like the roundworm *Caenorhabditis* with a very simple central nervous system. It has yet to be determined how much of the behavioral variation among individuals is a consequence of such random variation in the development of the central nervous system, variation that is inborn, but not genetic.

The environmental and chance contingency of the mapping between genotype and phenotype arises from the nature of organisms as physical objects. They are intermediate in size between particles and planets and they are internally functionally heterogeneous. As a result, unlike simple physical systems like the movement of planets around the sun, their functions are not the consequence of a few strong causal forces whose parameters govern the outcome. Organisms are the nexus of a very large number of individually weakly determining forces subject to noise at all levels. The consequence for the understanding of the physical properties of organisms, including their individual and social behavior, is that we cannot expect any simple

and universal generalizations like Newton's Laws. That is certainly true for the relation between genotype and phenotype.

See also: Body, Evolution of; Brain, Evolution of; Epigenetic Inheritance; Genetics and Development; Genetics of Complex Traits Through the Life Cycle; Human Evolutionary Genetics; Human Genome Project: Japanese Perspective; Intelligence, Genetics of; Cognitive Abilities

Bibliography

- Clausen J D, Keck D, Hiesey W W 1958 *Experimental Studies on the Nature of Species* Vol. 3: *Environmental Responses of Climatic Races of Achillea*. Carnegie Institution of Washington Publication No. 581
- Goss P, Pecoud J 1998 Quantitative modeling of stochastic systems in molecular biology using stochastic petri nets. *Proceedings of the National Academy of Science USA* **95**: 6750-5

R. C. Lewontin

Geoarchaeology

Research that uses the techniques, methods, and concepts of the earth sciences to address archaeological problems is called geoarchaeology. In the important early exposition, *Archaeology as Human Ecology: Method and Theory for a Contextual Approach* (1982), Karl Butzer described geoarchaeology as a collection of techniques for contributing substantive data critical to understanding prehistorical and historical human ecosystems (Butzer 1982, p. xi).

There are at least seven major concerns of geoarchaeology that must be integrated into research programs at the design, excavation and analytical stages and include: (a) site survey techniques that use geochemical, electromagnetic and other remote sensing procedures to locate sites or features within a known site; (b) documentation of site-formation processes and the regional context of a site; (c) using a wide array of approaches to differentiate cultural from natural (nonhuman) sediments and features, including disturbance processes by biological, pedological, and geological processes; (d) the development and correlation of intra- and extrasite temporal contexts by relative and/or absolute dating; (e) integration of ancient landforms and biological information for the reconstruction of palaeolandscapes (after Gladfelter 1981, p. 347); (f) reconstructing palaeoenvironments by the identification of pollen, wood charcoal, leaves, and seeds; and (g) the geochemical analysis of artifacts to document interaction between social groups. Exam-

ined here are a few techniques of geoarchaeology that are geared towards: discovering archaeological sites and documenting their internal structure; surveying site formation and disturbance processes; the analysis of soils and sediments; palaeoenvironmental reconstruction and the impact of humans on the landscape; and the physical analysis of archaeological materials.

1. Site Discovery and Definition

Traditionally, locating archaeological sites was accomplished by asking local residents in conjunction with a pedestrian survey; that is, simply walking over the landscape and looking for artifacts and food remains or larger features such as stone house foundations, terraces, or earthworks. In the 1960s with the advent of public archaeology programs—also known as ‘contract’ archaeology or heritage management studies formulated in response to development projects—archaeological surveys became more systematic, utilizing sophisticated sampling procedures, predictive modeling, and techniques of remote sensing, or archaeological prospecting as it is called in Europe.

1.1 Remote Sensing or Archaeological Prospecting

Archaeological prospecting ‘is considered as an exercise in the deliberate planned measurement and evaluation of physical properties which lead to the detection and mapping of archaeological sites’ (Scollar et al. 1990, p. 27). Since all surface indications of many archaeological sites have been obliterated by modern urban encroachment or been buried under soils, excavation has been the technique of choice for locating subsurface cultural deposits. Remote sensing techniques are nondestructive and include aerial photography, and resistivity, magnetic, electromagnetic, and thermal prospecting (see Scollar et al. 1990 for a detailed explanation of these techniques).

1.2 Aerial Photography

The oldest form of remote sensing, aerial photography is used in archaeology primarily for discovering field monuments, whether earthen or stone constructions. A huge corpus of aerial photographs exist some of which were taken during various wars for reconnaissance purposes, others for urban planning, to record changing shorelines, and to document vegetation patterns. However, aerial photos made exclusively for archaeological site discovery are the most useful since they are taken at the appropriate scale, angle, and time of day for enhancing surface features. Prehistoric agricultural field systems in areas such as Mexico and the Pacific Islands, consist of low terraces and earthen

mounds and embankments which are readily mapped from aerial photographs, saving countless hours of field survey and mapping.

1.3 Other Techniques of Remote Sensing

Once archaeological sites are located, identifying internal structures (intrasite variability) can be accomplished with other techniques of remote sensing including resistivity prospecting, where subsurface anomalies (such as burial pits, compacted ancient floors, or buried house foundations) are mapped by identifying the variation in an electric charge as it passes through the ground (Scollar et al. 1990, pp. 307–421). Magnetic prospecting can locate the shape, size and depth of subsurface anomalies that can then be mapped. Thermal prospecting uses variations in surface temperatures as an indication of subsurface (often cultural) structures.

1.4 Coring Archaeological Sites

Identifying subsurface archaeological site boundaries, internal features, and differential distributions of artifacts and faunal material can be accomplished by coring (Stein 1986), especially in conjunction with the analysis of sediment chemistry. Small cores or bucket augers (about 10cm in diameter) are used to collect sediment samples for the analysis of cultural content and to differentiate cultural from noncultural layers. Since organic matter, carbonates (e.g., marine shells), and phosphates (deposited in the form of urine, excrement and organic refuse) can increase relative to the intensity of human occupation, these residues are good indicators of past site use (Sjöberg 1976) and their distribution is used to identify subsurface cultural deposits with a minimum of site disturbance.

2. Site Formation Processes

The archaeological record is not static, but changes in response to a number of processes—both cultural and environmental (or noncultural). Once deposited in, or on, the surface of an archaeological site, cultural material (artifacts and food remains) and noncultural matter, such as pollen grains and terrestrial sediments, can shift their position or, in some cases, be removed completely. Since archaeological interpretations gain their strength from identifying and analyzing the vertical and horizontal distributions of material within and between sites, it is vital to understand how the positions of these materials may have changed since they were originally deposited—this is the challenge of site formation research defined as the processes, both natural and cultural, that build, disturb, redeposit, and destroy archaeological sites (Schiffer 1987).

2.1 Nonhuman Disturbance Processes

Since the soils and sediments that make up an archaeological site are part of a dynamic open system, archaeological sites are best viewed in a regional context. It is at this level that soil development is best understood because it is influenced by such large-scale processes as precipitation and temperature and erosional agents—including wind and surface water (e.g., rivers, marine shore waves, and sheet wash). Wood and Johnson (1978) have outlined a number of disturbance processes that affect archaeological deposits. Some of the more important processes include faunalurbation, or the mixing of deposits by animals, especially the burrowing kind (rodents, crabs, insects, earthworms, and birds). Digging burrows not only brings buried material to the surface, but open holes are traps for surface material to accumulate beneath the ground. The mixing of deposits by plants that occurs during root growth and decay is known as floralurbation. Uprooted trees can also bring material to the surface, often from a meter or more below the ground.

Discrete layers within archaeological sites can be homogenized by alternating freezing and thawing episodes which sort sediments by size. This is a problem with sites in the higher latitudes that consist of clay soils that swell and contract in response to precipitation and temperature. Gravity can cause mixing and movement of soil and rock debris downslope, as well as settling or subsidence of relatively large areas. Aeroturbation occurs when wind winnows fine sediments from deposits leaving a 'pavement' of artifacts. This is a significant problem in areas with active sand dune systems. Despite the range of disturbance processes that operate on a regional and site-specific basis, most cultural deposits maintain some order of integrity and, after the site formation processes are understood, valid interpretations are possible.

2.2 Human Disturbance Processes

In addition to natural or environmental site formation processes, people intentionally contribute to the formation of archaeological deposits by bringing to a site raw materials such as fine-grained rock to fashion into tools, finished artifacts of bone, shell, wood and stone, cooking fuel, building materials, and food. Many of these items are discarded on site as waste. Even sand grains adhering to feet, after countless thousands of trips to a site, can add significantly to the cultural matrix. Digging pits for earth ovens and trash disposal, small holes excavated for wall or fence posts, or trenches dug for building foundations can displace subsurface cultural deposits during later periods of occupation. Accurate interpretations of the archaeological record can only be made after understand-

ing both noncultural and cultural site formation processes.

3. Definition of Soils and Sediments

Soils and sediments figure prominently in discussions of site formation processes, so it is essential to know how these terms differ. Soils develop *in situ* at or near the ground surface through the various processes of weathering in conjunction with plant growth and decay, other biological elements, time, climate (temperature and moisture), and local topography. Sediments are particulate matter that has been transported from one location to another by one or more processes such as wind, water, or people.

3.1 Principles of Sedimentation

All particles, including artifacts and food remains, found in archaeological deposits can be viewed as sediments (Stein 1987, p. 339) and, because of this, the four principles of sedimentation (Shackley 1975) apply equally to natural and cultural sediments. First, the source of sediments found at an archaeological site can be unconsolidated particles of bedrock as well as humanly-transported stone tools, shells, bones or building stones, to name but a few examples. Second, the transport history of sediments is usually determined by textural and compositional analysis in that biological agents (people, plants, and animals), wind, and water move sediments according to certain physical constraints. For example, wind normally transports evenly-sorted sand-sized sediments, while water—in the form of rivers, sea waves, and glaciers—can move anything from clay particles suspended in its volume to huge boulders that are pushed under enormous power. Third, the environment of deposition, in this example, is the archaeological site, but offsite examples can include sedimentary basins, aggrading shorelines, sand dunes, and till and erratic boulders deposited from melting glaciers. And last, postdepositional alterations, discussed above, are caused by people, other animals, plants, and geologic erosional processes.

3.2 Sampling Soils and Sediments

Soils and sediment samples are usually taken from a profile in an erosional exposure or from the sidewall of an excavation unit or trench. Beginning from the bottom up, samples are taken from separate stratigraphic layers or in a continuous column. The sample volume varies according to the kinds of analyses undertaken and the size of sediments. For example, small samples of about 20 grams are needed to analyze the texture of clays and silts, while several kilograms are necessary for analyzing cultural content or, if

sediments are gravel, to cobble size. The key is to collect a representative sample of the sediment.

3.3 Analytical Techniques

Among other things, particle size analysis permits the detection of the agent of deposition (e.g., wind, river, sea) and the environment of deposition such as beach, flood plain, and dune (Shackley 1975, p. 87). As mentioned above, well-sorted sand is often wind-deposited and certain agents of deposition have particular sediment signatures. Mineral identification of individual grains can suggest a likely source of the sediments, especially if the regional geology is well known. Grain shape (roundness and flatness) can indicate the transport agent, depositional environment, and history. Grains can be, for example, platy or flat; angular, or granular. Roundness refers to the general grain surface curvature and is more a function of mineral composition (certain minerals are harder or softer than others and respond to friction differently), depositional history, and final depositional environment (Shackley 1975, p. 46). Grain surface textures, examined under high-power magnification, can be: unworn and angular, suggesting a fresh or minimally-transported history; rounded and glossy, the result of transport by running water; or matt-surfaced grains, the result of wind transport. A good review of the analytical methods of sediment analysis can be found in Shackley (1975).

4. Palaeoenvironmental Reconstruction and Human Impact

The relationship of people to the landscape (and here I use 'landscape' to mean the physical and biotic environment) is of worldwide interest, not only to archaeologists, but to biologists and politicians alike, to name just a few. How humans have historically managed their resources—whether successfully or with tragic consequences—is of utmost importance to contemporary society—especially as the world becomes increasingly one 'global village.' Archaeology is better positioned than any other discipline to understand the historical details of human-landscape relationships over tens of thousands of years and across an enormously varied set of environmental conditions that posed opportunities and problems for our ancestors.

To document and understand the variation in human-landscape interactions, the analysis of soils and sediments is especially useful when used in conjunction with the differing distribution of floral remains (see Hastorf and Popper 1988 for a review of analytical methods). The analysis of faunal material is also of importance here and is addressed in *Zooarchaeology*. Whether of carbonized wood fragments, pollen, opal phytoliths (minute silica bodies formed in

the cells of living plants that can be diagnostic of certain species), or leaves, plant remains are often indicative of vegetation on or near an archaeological site and, in the case of wind-dispersed pollen, can reveal regional floral patterns. Recovered from dated stratigraphic contexts, vegetation history can be charted over time—before, during, and after human settlement. In essence, archaeological sites hold an historical record of local and regional palaeo-environmental change, often in response to human modifications to the landscape.

5. Physical Analysis of Archaeological Materials

A growing area of interest in geoarchaeology is the elemental analysis of artifacts such as ceramics and stone tools so links can be drawn from the geological source or, in rare occasions, the site of production, to the place of deposition or archaeological site. The emphasis is on establishing lines of interaction or communication between or within social groups. The distribution of exotic artifacts across space and through time (that is, artifacts recovered from dated contexts or layers) is thought to reflect changing aspects of social organization—a topic of great interest to archaeologists.

5.1 Physical Analysis of Ceramics and Stone Tools

Ceramics are manufactured from clay, water, and nonplastic inclusions (temper). Pots were often decorated with an exterior clay slip, incised or stamped with various intricate patterns. Of interest to interaction or 'sourcing' studies is the elemental composition of the clay and temper. If the latter consists of crushed rock or terrigenous (land-based) sands, then chemical analysis of the temper may isolate its source. Stone tools, manufactured from obsidian, basalt, chert, or other fine-grained rocks, also can be geochemically analyzed to determine the artifact's geological source. Economically-important rock sources were generally well known in prehistory and had a limited distribution, thus facilitating sourcing studies.

There are many useful techniques for determining the elemental composition of ceramics and stone tools. (See Parkes 1986 for a description of various techniques.) These include, for example, x-ray fluorescence (XRF) and lead isotope analysis at the newer and more sophisticated end of the techniques (Weisler 1997), to thin-section petrography which has been used in geologic investigations for over a century.

5.2 X-ray Fluorescence Analysis

Simply stated, XRF analysis consists of bombarding a specimen with primary x-rays. The sample emits fluorescent x-rays that are indicative of the kind and

frequency of its elemental composition (oxides, such as iron or silica, are measured in weight percent, while elements are counted in parts-per-million or ppm). Dedicated computers are used for data analysis, while statistical clustering programs can aid in grouping artifacts with possible sources. Once sources are adequately characterized in terms of their geochemical variability, then unknowns or artifacts can be more confidently assigned to a specific source.

5.3 Thin-section Petrography

Thin-section petrography is used extensively today in ceramic sourcing studies when the mineral identification of temper is desired. Samples measuring up to 25 mm by 40 mm are cut from ceramic sherds and ground thin enough to pass light through them. Placed under magnification, minerals (such as olivine and feldspar) are identified and texture measured. Within a geologically well-known region, temper composition can be assigned to a particular river drainage (in the example of placer sands), a marine beach, or a larger region such as a group of hills or a mountain range.

Geoarchaeology uses the techniques, methods, and concepts of the physical sciences to address archaeological questions. In this short article, I have only touched on a few of the techniques and methods applied to research domains that are of worldwide interest. This subdiscipline of archaeology is continually expanding the range of techniques and their applications to innovative questions. The successful melding of interdisciplinary research typical of geoarchaeological pursuits continues to have a bright future.

See also: Bioarchaeology; Ceramics in Archaeology; Geographic Information Systems; Remote Sensing and Geographic Information Systems Analysis; Zooarchaeology

Bibliography

- Butzer K W 1982 *Archaeology as Human Ecology: Method and Theory for a Contextual Approach*. Cambridge University Press, Cambridge, UK
- Gladfelter B G 1981 Developments and directions in geoarchaeology. *Advances in Archaeological Method and Theory* 4: 343–64
- Hastorf C A, Popper V S 1988 *Current Paleoethnobotany: Analytical Methods and Cultural Interpretations of Archaeological Plant Remains*. University of Chicago Press, Chicago
- Parkes P 1986 *Current Scientific Techniques in Archaeology*. Croom Helm, London
- Schiffer M B 1987 *Formation Processes of the Archaeological Record*. University of New Mexico Press, Albuquerque, NM
- Scollar I, Tabbagh A, Hesse A, Herzog I 1990 *Archaeological Prospecting and Remote Sensing*. Cambridge University Press, Cambridge, UK

- Shackley M L 1975 *Archaeological Sediments: A Survey of Analytical Methods*. Wiley, New York
- Sjöberg A 1976 Phosphate analysis of anthropic soils. *Journal of Field Archaeology* 3: 447–54
- Stein J K 1986 Coring archaeological sites. *American Antiquity* 51: 505–27
- Stein J K 1987 Deposits for archaeologists. *Advances in Archaeological Method and Theory* 11: 337–95
- Weisler M I (ed.) 1997 *Prehistoric Long-distance Interaction in Oceania: An Interdisciplinary Approach*. New Zealand Archaeological Association Monograph 21, Auckland, New Zealand
- Wood W R, Johnson D L 1978 A survey of disturbance processes in archaeological site formation. *Advances in Archaeological Method and Theory* 1: 315–81

M. I. Weisler

Geodemographics

1. Introduction

A geodemographic database is a collection of spatially referenced data about individuals or groups, and geodemographics describes a rapidly growing segment of the marketing industry that collects massive amounts of data on consumer characteristics and behavior, constructs statistical models of consumer identity, and maps and analyzes distributions of market-related activity. It is used primarily to identify, or ‘target,’ potential customers, or ‘prospects,’ and in business applications such as retail site evaluation, trade area analysis, market share analysis, sales territory management and micromarketing. It can also, however, be used by nonprofit institutions to enhance fundraising, by governments for administrative purposes, such as school districting, and by political campaigns to identify voters susceptible to customized political messages. The sophisticated techniques of geodemographics depend on two simple assumptions: that ‘you are what you buy’ and ‘you are where you live.’

2. The Development of Geodemographics

The collection of customer information and target marketing have been around since the late nineteenth century, and customer segmentation and micromarketing were introduced in the 1950s. It was not until the 1970s, however, that theoretical, technological and institutional innovations made possible the maintenance of electronic databases on consumer behavior, the statistical analysis of consumer identity, and precise geographical location of consumers upon which geodemographics depends. Its invention is attributed to Jonathan Robbin, a sociologist who, in the early 1970s, saw entrepreneurial potential in the

recently introduced Zone Improvement Plan (ZIP) of the US Postal Service. First, he aggregated demographic data for the 240,000 block groups of the Bureau of the Census into the 36,000 postal delivery areas. Then he performed multivariate analyses of this data to produce 40 lifestyle categories describing consumer potential—hence he called the system PRIZM (Potential Rating Index for ZIP Markets). This innovation allowed marketers to target potential customers through heavily discounted bulk direct mail campaigns.

Today, publicly produced census data are still a vital component of geodemographics, and vendors still adopt clever acronyms like ACORN (A Classification of Residential Neighborhoods developed by CACI Marketing Systems, UK) and PSYTE (from CompuSearch, Canada) to describe proprietary segmentation systems. However, the US Bureau of the Census itself provides ZIP level data as a result of lobbying and financial assistance from the geodemographic industry, and its computerized spatial database called TIGER (for Topologically Integrated Geographic Encoding and Referencing) is the basis of the geography of most geodemographic systems. In Canada and the UK, there is also close cooperation between industry and government—in the UK, for example, Pinpoint Ltd. and the General Post Office have digitized every address that receives mail. In these countries, geodemographic databases incorporate diverse sources of data from official records, reported consumer behavior, lifestyle preferences and even psychological profiling ('psychographics'). Elsewhere, the industry has been slower to develop, largely owing to limitations in quantity and quality of public and private data, but it is truly international, with business marketing publications and software solutions available for many individual countries and most regions of the world. A cross-cultural clustering of consumers is already available for Europe.

While a host of specialist corporations and consultants offer individual services, a series of licensing agreements, acquisitions and mergers have resulted in national, and in some cases, transnational 'information conglomerates' (Curry 1993, p. 27) that purvey complex, integrated geodemographic systems. Such systems typically include three principal components: electronic databases composed of public and private, individual and aggregate information; segmentation routines that identify consumer types through cluster analyses of behavioral and psychographic data; and a geographic information system that provides the means for spatial analysis and mapping.

3. *Electronic Consumer Databases*

The geodemographics industry accumulates data from numerous sources to create its Marketing Customer Information Files: purchasing proprietary informa-

tion on customers from corporations, such as point of purchase surveys, warranty registration questionnaires, customer-preference cards and satisfaction surveys, and from government agencies; purchasing customized data from syndicated national consumer panel surveys, media surveys, opinion polls and credit bureaus; obtaining public transaction records of institutions such as banks, insurance companies and health services, which are kept in compliance with government regulations; and 'mining' public records on births, deaths, marriages, property transactions, crime, and mailing address changes. The scale is enormous, encompassing the vast majority of all of US households, and what were once databases are now 'data warehouses' with more than one trillion entries. Specialist firms and software programs provide means to sift and sort data to eliminate redundancies and match entries, particularly in terms of street addresses, and consumer information is now updated instantly and can be downloaded on line from most of the geodemographic vendors.

Such data are often collected and sold with only the implicit consent of consumers, who are unaware of the further uses to which data gathered in one transactional context might later be put. As a result of informational campaigns and legal actions, consumers are becoming increasingly wary of invasions of privacy and suspicious of 'consumer espionage' (Larson 1992, p. 5), and the industry has adopted voluntary controls. Even individuals who explicitly opt out when presented the option, however, are likely represented in official databases, and their consumer 'profile' can be inferred accordingly. Privacy is a notoriously tricky issue, and surveillance may be a necessary quid pro quo of participation in capitalist society (Robbin, cited in Larson 1992, p. 55). Perhaps public concern should be not so much about specific instruments of surveillance technology nor their applications to particular individuals, but the manner in which social subjects are constituted and represented in the discourses of social and spatial sciences (Foucault 1979). Thus, geodemographics is replete with 'strategic' metaphors of military operations, sexual conquest, and psychological manipulation, based upon reduction of identity and social difference to consumer profiles and residential location through its segmentation schemes and 'market mapping' (Goss 1994).

4. *Consumer Profiling*

Contemporary segmentation schemes go well beyond the abstract demographic variables of the census to offer 'Character: not just characteristics,' and this system, called Focus, for example, offers 'multidimensional portraits of real people,' revealing 'lifestyle and mindset' and promising marketers will 'see how customers view themselves,' and 'meet their customers on a first name basis' (National Demographics and Lifestyles 1993, p. 2). Segmentation schemes draw

upon information on brand preferences, media habits, credit rating, lifestyle and values, to divide neighborhoods in the US, typically into 30–40 categories such as ‘Pools and Patios,’ ‘Shotguns and Pickups,’ ‘Fur and Station Wagons,’ and ‘Hard Scrabble’ (PRIZM) or ‘Cautious Young Couples,’ ‘Sustaining Ethnic Families,’ and ‘Young Accumulators’ (MicroVision). Marketers ‘get acquainted’ with these potential customers through visual scenarios representing consumers engaged in a typical activities in front of their home and/or detailed imaginary vignettes describing their typical names and everyday lives—particularly their consumption practices. Although analysts argue their schemes imply no social judgments, the rankings, weightings and premiums charged for names of affluent consumers, and description of segments give them away: to the marketing industry, social identity and value is lifestyle, and ‘you are what you buy’ (Piirto 1991, p. 23).

5. *Geographic Information Systems*

The ‘geo’ of geodemographics is essential for several reasons: first, because the map transforms complex data into simple ‘actionable’ visual images—red typically represents a ‘hot’ market—ideal for busy executive officers; second, because addresses of households provide a convenient way to match database records; third, because addresses are means by which consumers are reached with marketing messages or products, whether delivered by mail, conventional media or retail outlet; fourth, because a range of analyses and transformations can be applied to spatial data to create ‘custom’ geographies such as retail trade areas, sales territories and media footprints; and finally because marketers infer characteristics of individual consumers from aggregate data and substitute missing values in their databases based on residential location. This last function is important because in some forms of data (census and vehicle licensing, for example) individual records are not available for confidentiality reasons, and where records are missing, the costs of collecting such data are generally prohibitive.

The importance of the geographical basis of marketing information is attested by the lengths to which the industry goes to infer residential location even for users of the World Wide Web, constructing user profiles from ‘search engine queries; online shopping patterns; requests for localized information, such as movies, concerts, lectures and stores; and interested in regional websites such as *yourcity.com*’ (Melville 1999, p. 34). Inference from ‘neighboring’ records is based on the assumption, according to Jonathan Robbin, that ‘humans group themselves into natural areas ... They create or choose established neighborhoods that conform to their life-style of the moment’ (cited in Burnham 1983, p. 92). Not surprisingly, Claritas’ promotional material similarly claims that it

is ‘based on the fundamental sociological truism that “birds of a feather flock together”’ (Claritas/National Planning Data Corporation (NPDC) 1992). The geographical basis of consumer profiling is also a result of the so-called neighborhood effect, such that, for example, ‘when one moves into a suburban neighborhood in which everyone has a gas barbecue there is a strong motivation to buy one’ (Lea and Heumann 1999, p. 26). There is, of course, an undeniable spatiality to everyday life, but to infer individual characteristics from those of an area is to fall for the ecological fallacy (to assume that individual characteristics can be inferred from group characteristics, which is only true if the population is homogeneous) and the question is how much such an effect is natural or intrinsically motivated, and how much it is the product of the marketing enterprise itself.

6. *Conclusion*

Geodemographics is particularly hype-ridden and the success of its applications limited, for example, to very modest ‘lifts’ in customer response by 5–10 percent in direct mail campaigns, even if this represents considerable savings in a large-scale operation. It is, nevertheless, the fastest growing segment of the information industry, presumably because of its promise of intimate and actionable knowledge about consumers. Geodemographics first reduces the subjectivity of individuals to objective and quantifiable characteristics, then reconstructs composite ‘soap opera’ identities (Goss 1995, p. 187), perhaps more consistent with marketers prejudices than social reality—hence one geodemographic system promises that ‘as you get to know [our cohorts], you’ll discover these are all people you know: your friends, neighbors, relatives, and—most importantly—your customers’ (National Demographics and Lifestyles 1993, p. 2). There is a second, more sinister, self-fulfilling promise, however, in that the definition and targeting of consumers based on residential location may effect redlining of consumption, such that marketing campaigns and retail services are provided only to addresses—with known names, inferred names, or merely ‘Current Resident’—with the ‘right’ statistical and behavioral profiles. This is not to argue that there are not cleavages in consumption based on class and residential location, but if, for example, magazine issues are differentiated in terms of advertising content for subscribers in different ZIP codes, as some already are in the US, consumption of these goods and services in turn will reproduce, and perhaps actually produce, further segmentation and segregation of society. Such is the instrumental reason behind the hype of geodemographics.

See also: Economic Geography; Market Areas; Market Research

Bibliography

- Burnham D 1983 *The Rise of the Computer State*, 1st edn. Random House, New York
- Claritas/NPDC 1992 *Prizm Clusters*. Brochure, New York
- Curry D J 1993 *The New Marketing Research Systems: How to Use Strategic Database Information for Better Marketing Decisions*. Wiley, New York
- Foucault M 1979 *Discipline and Punish: The Birth of the Prison*. Vintage Books, New York
- Goss J D 1994 Marketing the new marketing: the strategic discourse of geodemographic information systems. In: Pickles J (ed.) *Ground Truth: The Social Implications of Geographic Information Systems*. Guilford Press, New York
- Goss J D 1995 'We know where you are and we know where you live': the instrumental rationality of geodemographic information systems. *Economic Geography*. **71**: 171–98
- Larson E 1992 *The Naked Consumer: How our Private Lives Become Public Commodities*, 1st edn. Holt, New York
- Lea T, Heumann D 1999 Key in on variables to improve databases. *Business Geographics*. **7**: 24–6
- Melville R 1999 Online marketing goes geographic. *Business Geographics*. **7**: 33–5
- National Demographics and Lifestyles 1993 *Character. Not Just Characteristics*. Brochure, Denver, CO
- National Demographics and Lifestyles 1993 *NDL FOCUS: Knowing your Customers*. Brochure, Denver, CO
- Piirto R 1991 *Beyond Mind Games: The Marketing Power of Psychographics*. American Demographics, Ithaca, NY

J. Goss

Geographic Education

Despite the commonly used name, 'geographic' education is more appropriately thought of as 'geography' education. The distinction is significant. The use of 'geographic' as a modifier might suggest a special type of educational theory or practice. For example, it could imply distinctive and particular pedagogical procedures and therefore unique methods of training for teachers of geography. While the interaction between content and pedagogy in any domain of knowledge requires discipline-specific understandings and skills, there is nothing inherent in the nature of geography that warrants a special type of education, fundamentally different in nature from that practiced in disciplines such as physics, psychology, and history. Geography education is about the teaching and learning of geography (Bednarz et al. in press).

Both geography education and geographic education must be distinguished from the geography of education. As a major sector of the modern economy, the education system employs significant numbers of people and involves substantial capital investments in terms of school buildings, university and college campuses, buses, computers, textbooks, etc. The geography of education is concerned with understanding the spatial organization of the education

system. It encompasses, for example, the relationships between patterns of tax revenues and levels of education expenditures, the diffusion of colleges and universities, the optimal design of bus routes to link students with schools, and the equity and efficiency of the provision of education to different socioeconomic and/or racial and ethnic groups. The focus is on understanding the spatial manifestation of the K-22 education system as part of the infrastructure of society.

Although an understanding of the educational infrastructure is not part of geography education as such, the state does have a major impact on the nature of geography education. In the USA, for example, the fundamental belief is that education is best controlled at a local, sub-state level. In Britain, and in France to an even greater degree, education control is more strongly vested in the national government. The ideological and structural differences between decentralized and centralized systems affect the nature of the curriculum, the training of teachers, the selection of educational materials, and the relative emphases on geography education.

1. The Nature and Process of Geography Education

Geography education is concerned with fostering an understanding of the domain of geographic knowledge (its subject matter content, skills, and perspectives). The intent is not necessarily to produce geographers as such but to provide the essential geographic knowledge and skills for an educated person.

1.1 The Nature of Geography

As the science of space and place on the earth's surface, geography seeks to describe and explain the distribution, pattern, and nature of places and regions as seen at scales ranging from the local through to the global. The understanding of places and regions is set in multiple contexts: spatial, environmental, and temporal. The character of places and regions is seen as shaped by the interrelationships between physical phenomena, such as climate, terrain, the hydrologic cycle, biosystems, etc., and human phenomena, such as demographic characteristics, economic functioning, culture, politics, urbanization, etc. The shaping process is understood in a variety of ways, in terms of:

(a) the coalescence of physical and human phenomena to create the particular character of a place (e.g., local topography, microclimates, level of urbanization, ethnic composition, etc.);

(b) the impacts of global environmental and economic changes on local places (e.g., economic restructuring in labor-intensive industries, with the loss of local employment to cheaper 'offshore' production facilities); and

(c) the interdependencies between places on local, regional, and global scales (e.g., chemicals in runoff from farms in upstream areas polluting regional water tables and watersheds).

Therefore, the focus of modern geography is on understanding the evolving character and organization of places and regions on the earth's surface. Geography's approach is distinguished by (a) a sensitivity to context, (b) an appreciation of interconnections between spatial patterns and processes, (c) an emphasis on understanding systemic interrelationships at different spatial scales, and (d) an awareness of the reciprocal relations between physical and human systems.

The skills necessary for geographic understanding are of three types. There are general problem-solving or critical thinking skills which ask and answer geographic questions by means of acquiring, organizing, and analyzing geographic information. There is the use of broadly applicable methods such as spatial analysis (quantitative and qualitative), geographic information systems (GIS), field methods, and cartography. Finally, there is the application of discipline-specific technologies such as satellite image interpretation, computer-based geographic visualization, and global positioning systems (GPS). Perspectives in geography start with the essential spatial and ecological perspectives and have been extended to include, for example, feminist, historical, economic, social and critical theoretic, and humanist perspectives.

1.2 The Process of Geography Education

The process of geography education involves the teaching and learning of geography in formal (e.g., preschool, school, college) and informal (e.g., familial, societal, workplace) contexts. Ideally, the process is lifelong in the context of an individual person—preschool, K-12, and perhaps beyond into higher education (13–16 or even 17–22) and the workplace—and continuous in the context of developments in the discipline of geography. In practice, however, there have been significant historical changes in:

(a) the ages at which geography education has been offered;

(b) its links to other components of the formal education system (e.g., geography as a standalone or as an infused subject);

(c) the role of the state vs. informal organizations in developing materials, training teachers, and setting the educational agenda; and

(d) geography's role in facilitating societal imperatives (e.g., support for exploration and imperialism, mass education, and the development of workforce skills). The field of geography itself has also changed its emphasis on and approaches to geographic education.

2. Approaching the History of Geography Education

Livingstone (1992) characterized the history of geography as a 'contested enterprise.' His strategy for 'telling geography's story' was to focus on the 'intellectual and social context within which geographical knowledge was produced' (Livingstone 1992, p. 23). If we affix 're' to 'produced,' then we can also tell the story of geography education.

While the production and reproduction of knowledge are necessarily linked, they do not move in some elegant *pas de deux*. The field of geography—the social organizations by which domain knowledge is generated, preserved, and transmitted—is complex. As a discipline, geography has an existence as an academic and as a popular subject, the latter exemplified by the activities of the National Geographic Society, which at first were confined to the USA but now have an increasingly international reach. The two parts of the discipline are characterized by approaches, agendas, and audiences that are more often competitive rather than complementary. Moreover, internal, disciplinary views and external, societal views of the purpose and value of geography are different and sometimes at odds with one another. Superimposed on these differences are the valuations placed by professional geographers on the roles for geography education.

3. Geography Education as a Contested Enterprise

The history of geography education is also a contested enterprise. It has been shaped by complex interactions among a variety of factors:

(a) individual and societal valuations of geographic knowledge;

(b) the current state and past history of geography as a discipline (in terms of dominant approaches, concepts, models, and theories);

(c) the state of the formal education system (in terms of societal goals, physical infrastructure, resources, teacher training, textbooks, etc.); and

(d) the state of pedagogical theory and practice in education in general, and geography in particular.

While all of these factors vary across time and space, they coalesce into changing national approaches to geography education.

4. The Rationales for Geography Education

As with many disciplines, the rationale for the importance of geographic knowledge has been intertwined with the definition of an educated person. That definition has been subject to changes in the nature of

geographic understanding, and to the shifting emphases of the larger projects of society.

While it is impossible to unearth the roots of geography education, we can trace the emergence of some of the rationales for expecting people to understand geography. In late-Medieval Europe, for example, one grounding of formal education was in the teaching and learning of the seven sciences of the *artes liberales*: the *quadrivium* or advanced division (arithmetic, geometry, music, and astronomy) and the *trivium* or lower division (grammar, logic, and rhetoric).

5. Two Geographic Themes

Geography, as in knowledge of the location and character of places and the creation and use of maps, fell partially within the rubrics of geometry and astronomy. Place knowledge was an expected and necessary part of the knowledge base of an educated person and map use was a valuable practical skill in travel, navigation, surveying, business, warfare, and diplomacy. As both foundational knowledge and applied skill, geography therefore had elements of what we would today recognize as a liberal arts subject and as a science discipline, thus accounting for the discipline's varied placement within the college and faculty structures of universities in North America.

The two themes of foundational knowledge and applied skill are interwoven throughout the history of geography education in all countries. On the one hand, geography education is seen as essential to the project of creating an educated person, whether that project is couched in terms of (a) a broad liberal education, (b) citizenship education (in local, patriotic and national, or world contexts as expressed in the slogan 'think globally, act locally'), or (c) geographic literacy (e.g., reading and understanding the contents of a newspaper). Participation in local community affairs, voting, and support for national and increasingly international policies all depend upon a level of geographic literacy. On the other hand, geography education is seen as providing essential skills for operating in society in general (e.g., map-reading skills) and for participating in the workforce in particular (e.g., using geographic information systems). Careers in planning, business, and government agencies all depend upon significant levels of geographic skill.

5.1 Geography as Foundational Knowledge

The foundational knowledge approach to geography education has many variants. Geographic knowledge is seen as a way of counteracting and overcoming the natural and/or socially induced human tendencies towards egocentrism, parochialism, and ethnocentrism. Knowledge of the diversity of patterns of

human occupancy of the earth's surface reduces the chances of privileging one's own place in the world and increases the appreciation of cultural diversity and human creativity. Similarly, geographic knowledge underpins an understanding of physical environments, especially important as those environments are under threat from human-induced changes at local and global level. It can reinforce an understanding of the interdependency between human and physical systems. Geographic knowledge is seen as important for the teaching of other subjects, especially history, sociology, anthropology, economics, business, and political science. Without an appreciation of the intersections between space and place and between environment and society, the subject matter of those other disciplines cannot be appropriately grounded and understood. And finally, geographic knowledge need not be seen as exclusively instrumental or utilitarian: '(t)he power and beauty of geography lie in seeing, understanding, and appreciating the web of relationships among people, places, and environments' (Geography Education Standards Project 1994, p. 3).

5.2 Geography as an Applied Skill

The grounding of geography education in skills also has multiple bases. The science, art, and craft of cartography has been fundamental to many social projects, including surveying, documenting land ownership systems, exploration and navigation, colonization and military conquest, regional and local planning, and environmental science. As technologies for data gathering (e.g., satellite remote sensing), ensuring data accuracy (e.g., global positioning systems), data processing (e.g., GIS), and displaying geospatial data (e.g., geo-visualization tools) have become pervasive, so too have been the needs for geographically educated people to produce and interpret an increasing range of graphic displays. Those technologies and their products are central to the daily operations of government (local and national), the military, business and logistics, international politics, and so on. For example, knowledge of the spatial structure of national and world economies, together with the tools of spatial analysis, are seen as essential for individual participation in, and national success in, the business of the increasingly interdependent global economy.

Geography education has, therefore, been based on an intermixing of four principal rationales:

- (a) the existential value of understanding one's place in the world;
- (b) the ethical imperative of appreciating the fragile human and physical systems upon which we depend;
- (c) the intellectual power of understanding the interplay between two pairs of ideas—space and place, and environment and society—in shaping the patterns of the earth's surface; and

(d) the utilitarian motivation of using geography to solve individual and societal problems. These rationales have motivated the process of geography education and have been differentially emphasized during the history of geography education (Douglass 1998).

6. Geography Education: a History

While there had been a strong classical tradition of geography in the eastern Mediterranean as far back as 2000 BC, the efflorescence of geography was contemporaneous with the so-called Age of Exploration and the origins of the Scientific Revolution beginning in fifteenth-century Europe. Built on the maps and texts surviving from the Greek, Egyptian, Roman, and Arab cultures, geography was both a statement about the world and a way of recording the new information about diverse places. Maps were both tools for navigation and means of recording the results of travels. Texts captured descriptions of the location and properties of places encountered.

6.1 Varenius and the Two Geographies

These two roles of maps and texts are a convenient starting point in an account of the history of geography education. In 1649, Bernhardus Varenius (Bernhard Varens), a German doctor living in the Netherlands, published *Descriptio Regni Japoniae et Siam*, followed in 1650 by his *Geographia Generalis*. In the latter work Varenius systematized an understanding of the logic of geography, distinguishing between general and special geography. Although the terms were not original to Varenius, his statement of the distinction was crucial to the development of geography as a discipline in general and to the history of geography education in particular.

General (or universal) geography was explanatory in nature, seeking to understand the patterns and processes of the world via general (or natural) laws. Echoing the links of geography to geometry and astronomy in the *quadrivium*, general geography was linked to the 'mixed' (or applied) mathematics. Its focus on the terrestrial globe was paralleled by astronomy's focus on the celestial globe. General geography sought to understand global patterns of climate, topography, hydrography, etc., by means of forces and mechanisms (the laws of motion, heat, erosion, etc.). It moved from global explanations to accounts of regional variations, emphasized physical over human phenomena, and incorporated both geodesy and cartography.

Special geography was descriptive in nature, seeking to depict the characteristics of particular places. Although Varenius did not offer a detailed description of special geography in *Geographia Generalis*, his

outline description matched the contents of his earlier *Descriptio Regni Japoniae et Siam*. While the basis for general geography was demonstration from general laws, special geography derived from experience, systematized in terms of the terrestrial (physical) and human properties of places.

While there is a necessary link between special and general geography, there is also a recognition that general geography is more challenging, more intellectually demanding, and therefore in a sense, more advanced than special geography. General geography seats the discipline squarely within the tradition of natural philosophy and eventually the mathematical and empirical sciences. Varenius' work was influential throughout Europe for over two centuries (used by Newton in England and Kant in Germany, for example) and in the USA (Warntz 1964, 1982).

For geography education, Varenius' distinction between general and special geography establishes the two tracks along which geographic teaching and learning developed. It also helps to account for the trajectory of success, failure, and eventual re-emergence of geography instruction in educational systems. The book itself is important for two additional reasons. First, Varenius was concerned with restoring what he felt was a needed balance between general and special geography, the balance being decidedly in favor of special geography. To Varenius the imbalance meant that students were ignorant of the foundations of their discipline and this lack of attention to general geography placed into question the status of geography as a science. Second, the book was written as a textbook for advanced students, not for scholars. The changing balance between special and general geography, the relative intellectual status of the discipline, and the role of textbooks are themes that recur through the history of geography education.

From the sixteenth to the early nineteenth centuries, geography was part of both the school and college education systems in Europe and North America. In the USA, for example, admission to the colonial colleges was based in part upon knowledge of geography. It was expected that school and college students would know about the world in general and their national worlds in particular, and that they would have some minimal understanding of the maps and globes which were instruments for conveying and using knowledge about the world. Geography instruction incorporated both general and special elements, although there was a preponderance of the latter in schools and the former in college. Geography was a standalone subject in the curriculum of schools and the faculty of colleges and universities. It was supported by a rich infrastructure of textbooks, atlases, wall maps, and globes. Geography instruction led, for example, to new tools. In England in the 1760s, John Spilsbury, a London mapmaker, created 'dissected maps' as a way of teaching geography. In an interesting twist, dissected maps became jigsaw puz-

zles, a major source of entertainment named after the tool used for their construction, and they lost their meaning as a tool of geographic instruction.

During the late-eighteenth and early-nineteenth centuries, geography instruction responded to larger societal projects. In European countries, the twin forces of nationalism and imperialism reshaped territorial relationships throughout the continent and the world. Geography became a way of inculcating an understanding of national identity and international aspirations. Thus there are numerous literary allusions to children being forced to learn maps of the British Empire in which imperial possessions were boldly portrayed in vast expanses of red. Children were subjected to a world being carved into places that were named and owned, demarcated by sharp lines and differentiated by rich physical and human detail. This detail was the cumulative result of several centuries of exploration, conquest, colonization, and settlement. There was also an ideology for structuring this world understanding. Places were ordered as being populated by savages, barbarians, half-civilized, and civilized and enlightened people. Progression along this typology was allegedly a response to the civilizing influences of European society. This ideology was the basis for many eighteenth and nineteenth century textbooks and atlases.

In the USA, two interlocking forces shaped the nature of geography education and led to its gradual demise as a school subject during the first half of the twentieth century. The first force was universal or mass education, a force which applied equally to European societies. The second force was the expansion and integration of the national territory. As the nation grew in size and complexity, there were increasing demands that students should understand the geography of the emerging nation. Universal education brought with it pressures towards access for all students and the fostering of a spirit of national identity. Political independence meant also independence in terms of what was taught and the material on which teaching was based. All of these imperatives lead to an increasing focus on special geography, with the USA being the focal point of the world view and with a sense that understanding one's own country was essential, easy, and accessible to all (Bruckner 1999).

Beginning in 1784, Jedidiah Morse wrote a series of textbooks that shaped American school geography. An 1814 edition was entitled *Geography Made Easy* and it was developed 'for the use of schools and academies in the United States of America.' These books spawned a series of progenitors throughout the nineteenth and twentieth centuries. The texts made all places, but especially the nation and the student's state of residence, seem special and perhaps unique. Children learned by rote, working from the ubiquitous trio of the globe, the textbook, and the wall map of the world. As geography, particularly special geography,

literally became easy, it lost respect in the eyes of many in the academy. While the knowledge gained was useful and even essential, it did not challenge children's minds, except in the most trivial sense of factual learning.

From the middle of the nineteenth century to the beginning of the twentieth century, major geographic scholars were centrally involved in the process of geography education at all levels, K-12 included. For example, William Morris Davis in the USA and Halford MacKinder in the UK exemplified a seamless approach to the development of geography as an academic discipline and as a school subject. In both countries, there were organizations, the National Council for Geographic Education and the Geographical Association, whose primary function was the fostering of geography education in society. The National Geographic Society pursued its goal of increasing and diffusing geographic knowledge for mass audiences in as many countries as possible. Geography remained a popular subject, as exemplified by two bestselling books. V. M. Hillyer's (1929) *A Child's Geography of the World* and Hendrik van Loon's (1932) *Van Loon's Geography* remained in print for over 20 years and both were translated into many languages. There was a popular interest in geographic knowledge and therefore modest levels of support for K-12 geography instruction.

7. Geography Education in the Modern Era

During the middle to late twentieth century, geography education played different roles in different countries. In the UK, it remained a standalone subject that was part of the curriculum at all school levels and it was available as one of the Ordinary and Advanced level exit exams in high school. In this context, geography was traditionally grouped with the arts rather than the sciences. In both France and Germany, geography also played a comparably steady but modest role in the school curriculum. In the USA, geography languished at all school levels. The integrity of the discipline as a whole was destroyed as a consequence of being split apart as a subject. Human geography was subsumed within the social studies and physical geography within the earth sciences. These curricular emphases were mirrored in the structure of the education training system, with preservice teachers being explicitly trained as specialists in geography in many countries but not in the USA. In all cases, geography was taught as a cross between knowledge of the fundamental properties of places and regions in the world and, to some limited extent, a set of practical skills (e.g., map and atlas use, survey methods, fieldwork, etc.)

There are three principal frameworks underpinning the scope and sequence of geography in the K-12 curriculum. Curricula were place, scale, or concept-

based. The place-based curriculum structure is the oldest structure, dating back to the eighteenth century focus on special geography. Various known as the 'capes and bays' or the 'principal products of Peru' approach, it employed a simple classificatory framework to capture the characteristics of places around the world. Students learned about the location, boundaries, physical geography, and human geography of places, often in a rote fashion and using simple maps and atlases. In effect, younger children learned less, as in fewer and simpler things, and older children learned more, as in more detail, about the world. Geography education became an A to Z world tour, organized around continents, countries, and regions.

The scale-based structure is also known as the expanding environments or expanding communities approach. It begins in the early grades with a focus on the familiar and on the local world of the child, progressing to increasingly distant places with increasing age of the students. The local-to-global progression is premised on teaching about what is readily accessible to the understanding of children at different ages and levels of cognitive understanding. The third approach is concept-based, developing a systematic understanding of sets of concepts that can be applied to a range of spatial contexts at different scales. Geography is seen as a way of asking and answering questions about what is where in the world, and about why and how it got there.

These three curricular structures fall loosely into a historical sequence, if not progression, and reflect significant changes in the view of the discipline of geography. The essential transition is from seeing geography as a property of the world (the geography of a place) to seeing geography as a perspective on the world, as a way of understanding patterns and processes on the earth's surface. In the place-based approach, there are lists of things to be known about the world; in the concept-based approach, there are alternative ways of knowing about the world. Beginning with the scale-based approach, there is a sense of active involvement with the world, an involvement based on field trips, on maps as tools for recording and investigating places and environments, and students as explorers and creators of their own understanding.

Underpinning these three approaches are, therefore, approaches to child development. Many of the great educational theorists (Pestalozzi, Montessori, Piaget, etc.) treat issues that are central to the child's understanding of the world, although they do not necessarily do so within the formal domain of geography. Thus, for example, Piaget discusses the development of spatial understanding, the sense of nation, the understanding of basic physical phenomena (e.g., clouds, rain, etc.) and he makes use of maps and spatial models in his research designs. Geography education, while not necessarily basing its approaches explicitly on such theorists, is of necessity informed by their ideas.

In the early to middle 1980s, there was a resurgence of interest in many countries in the success—and often failure—of education systems in generating knowledgeable and competent students. The context for this interest was the intersection among several factors, including the effects of globalization on national economies, the desire to remain competitive in a rapidly changing world, and the increasing demands for technical skills in the workforce. As typified in the seminal US report *A Nation At Risk* (National Commission on Excellence in Education 1983), there were significant concerns about the structure of education in terms of teacher training, teacher support, curricula, textbooks, assessment, accountability, etc.

In the case of geography, the focus, especially in the USA, was on unacceptably low levels of geographic literacy among recent high school graduates. International comparisons showed relatively low levels of geographic literacy in a range of countries, but the performance of American students was particularly disappointing. As a consequence of the renewed national interest in education, geography education became the focus of attention from national and private organizations in many countries.

That attention paralleled national concerns about the knowledge and skills of school graduates and it led to a normative focus on what students should know and be able to do. The focus gave rise to national documents specifying the content of the curriculum at various grades and/or ages. Often these documents specified learning outcomes that were expected of students and therefore the documents became the basis for the assessment of student learning. In the case of geography, both the UK and the USA produced national guidelines (or standards).

8. *The Current State of Geography Education*

Geography education remains a contested enterprise. As K-12 educational systems place an increasing emphasis on immediately relevant skills—mathematics, writing, and to some extent science—other subjects are of necessity given less attention in terms of time in the classroom, teacher training, material support, and access to computing resources. Where geography education is infused in the curriculum, this lack of attention is a serious problem and results in geography being squeezed out or at best subsumed within another subject such as history. While there is public support for a level of geographic knowledge, that support is often in terms of place location knowledge, a necessary but insufficient base for geography education.

Geography education is changing in ways that move it away from the negative image associated with 'old-fashioned' rote learning of geographic facts. Several

transitions are underway. The first is a recognition that the process of geography education can be set within a novice–expert framework. This entails understanding the different types of knowledge that underpin increasing expertise: (a) declarative or propositional (e.g., factual and conceptual understanding), (b) procedural (e.g., how to perform some operation), and (c) metacognitive (when, where, and why to link facts, concepts, and operations to answer geographic questions). It involves understanding the organization of the mental models that children develop (e.g., Earth–Sun relations, the water cycle, urban growth), the ways in which different skills are available at different ages (or cognitive developmental levels), and the role of peers and teachers in supporting the student's developing understanding (Downs 1994).

A second transition is a switch from rote learning to active, inquiry-based learning. Geography becomes a way of finding out about the world rather than simply a set of things to be known about the world. Computers and geographic databases allow students to ask and answer geographic questions for themselves. A third transition is from tests to authentic assessments. Assessment is built into the structure of the ongoing learning process and is focused on problems and situations that are likely to be encountered in real-world contexts. Students can demonstrate their developing skills by building portfolios of their work.

9. The Future of Geography Education

All of these changes in geography education mirror changes that are occurring in the nature of education in general. The short-term future of geography education lies in the successful linking of the case for geography education with the increasing capacity to do geography. In the same way that geography was fundamental to past societal projects (e.g., exploration and imperialism), it is also fundamental to the current development of an information society. Much of the data that is available can be assigned geospatial coordinates and therefore it can be handled efficiently and effectively via geographic information systems. The larger parent field of geographic information science requires an understanding of maps and graphics, of space and place, of the interactions between environment and society. Those understandings, while not exclusive to geography, are central to the subject and can be taught through geography education. Given the growing importance of information science and technology to all aspects of modern society and economy, there is a potential to re-establish an essential role for geography education in the contest for space—and time—in the educational system (Rediscovering Geography Committee 1997).

See also: Cartographic Visualization; Cartography; Geographic Learning in Children; Geographic Thought, History of; Knowledge, Education, and Skills, Geography of

Bibliography

- Bednarz S, Downs R M, Vender J in press Geography education. In: Gaile G L, Wilmott C J (eds.) *Geography in America at the Dawn of the 21st Century*. Oxford University Press, Oxford, UK
- Bruckner M 1999 Lessons in geography: Maps, spellers, and other grammars of nationalism in the early Republic. *American Quarterly* **51**: 311–43
- Douglass M P 1998 *The History, Psychology, and Pedagogy of Geographic Literacy*. Praeger, Westport, CT
- Downs R M 1994 Being and becoming a geographer: An agenda for geography education. *Annals of the Association of American Geographers* **84**: 175–91
- Geography Education Standards Project 1994 *Geography for Life: Executive Summary, National Geography Standards*. National Geographic Research & Exploration, Washington, DC
- Hillyer V M 1929 *A Child's Geography of the World*. The Century Co., New York
- Livingstone D N 1992 *The Geographical Tradition: Episodes in the History of a Contested Tradition*. Blackwell, Cambridge, MA
- National Commission on Excellence in Education 1983 *A Nation At Risk: The Imperative for Educational Reform*. US Government Printing Office, Washington, DC
- Rediscovering Geography Committee, National Research Council 1997 *Rediscovering Geography: New Relevance for Science and Society*. National Academy Press, Washington, DC
- Van Loon H W 1932 *Van Loon's Geography: The Story of the World We Live In*. Simon and Schuster, New York
- Wartzt W 1964 *Geography Now and Then*. American Geographical Society Research Series No. 25, New York
- Wartzt W 1982 Geographia generalis and the earliest development of American academic geography. In: Blouet B (ed.) *The Origins of American Geography in the United States*. The Shoe String Press, Hamden, CT, pp. 245–63

R. M. Downs

Geographic Information Systems

A geographic information system (GIS) is best defined as a software system designed to provide services related to geographic data. Such services include storage, analysis, transformation, maintenance, editing, visualization, modeling, and many more. That somewhat dry definition masks a record of rapid growth and widespread applications, beginning some 30 years ago. GIS is now the focus of a substantial software industry and boasts a user community that probably numbers in the hundreds of thousands or

even millions worldwide. While many of those applications lie in areas of social and behavioral science, they also extend across the full range of sciences, industries, and government agencies that deal in one way or another with information about the surface of the earth, the subsurface, and the atmosphere.

1. Overview

GISs are one of several geographic information technologies that have appeared or evolved rapidly in recent years. Geographic information is also handled in image processing systems, in the form of aerial photographs or images from space satellites, and many GISs include image-processing capabilities. The Global Positioning System (Kennedy 1996) and its Russian equivalent use satellites to allow users to measure their positions on the earth's surface, with accuracies that range from millimeters to tens of meters depending on the complexity of the measuring device. Earth-observing satellites (Jensen 1996) capture images of the surface and are now widely used to map, detect change, discover and inventory resources, forecast weather, and measure variables of importance to earth science. The Internet and the World Wide Web (WWW) are widely used to distribute geographic data from data archives and digital libraries, and to provide GIS services to remote users.

Today, the potential user of GIS might take advantage of any of a number of hardware and software configurations. The examples that follow relate to systems available in the year 2000 and are ordered by increasing cost and power.

(a) A simple WWW browser, running on a laptop or office machine, and used to access WWW sites that offer a range of GIS services. Such services include generation of standard map products, and simple analysis and mapping of data.

(b) A desktop system, such as a PC or Macintosh, running a relatively simple GIS such as ArcView (Environmental Systems Research Institute (ESRI), www.esri.com) or MapInfo (MapInfo Corporation, www.mapinf.com). Data would be resident on the system's hard drive, and might have been created locally or obtained from a data supplier. The operating system would likely be a Microsoft or Apple product.

(c) A networked workstation, linked to other workstations and servers in a local area network, perhaps under the Windows NT or Unix operating system. The software would be more complex than in the first option, perhaps ESRI's ArcInfo or Spatial Database Engine.

A social scientist 'doing GIS' might use any one of these configurations, depending on personal skills and needs. The system would provide services that include preparation of simple maps, manipulation of data to overcome issues of map projections, running of statistical tests, and fitting of complex models to data.

Courses in the use of GIS and in methods of spatial analysis and modeling are now offered in almost all universities and colleges, and in many high schools. A wide range of educational materials are available in the form of textbooks (e.g., DeMers 1997; Clarke 1997; Heywood et al. 1998; Burrough and McDonnell 1998), WWW sites (e.g., www.esri.com and www.geoplace.com), and WWW-based courses (e.g., www.unigis.com). GIS software vendors maintain WWW sites with much useful information, and trade magazines (*GeoWorld* and *GeoInfo Systems*) carry information about new products and initiatives.

The history of GIS (Foresman 1998) originated in the 1960s, with the development of the Canada Geographic Information System (CGIS). Analysis of paper maps is notoriously difficult, because of the awkwardness and unreliability of techniques for measuring simple properties such as length or area, and because of the difficulties associated with overlaying different maps of the same area. Computer technology was applied in the late 1960s and early 1970s to the editing of maps by producing agencies such as the US Geological Survey, and to the automation of various functions associated with taking the census. But early efforts were crude and expensive, and GIS became widely available to government agencies and corporations only in the early 1980s. Today, the general public is increasingly aware of GIS through WWW sites such as www.mapquest.com, which generates several million maps a day for users who need driving directions, and through simple mapping packages such as Microsoft's MapPoint 2000.

2. GIS Essentials

General computer packages to support analysis of data began to appear in the 1960s, led by BMD, SAS, and SPSS. The table is a powerful model for organizing data, by arraying cases in the rows and devoting each column to some property of each case, and early packages took advantage of this commonality by organizing housekeeping and analytic functions around it. Thus, means and variances could be computed by column; regression could be readily applied to two columns; and data could be aggregated by computing the mean value of a column for every unique value of some other, categorical column. In essence, these packages exploited an economy of scale in software production: once the housekeeping functions for a simple data model like a table had been built, it became very cost-effective to add new functions based on the same data model.

GIS similarly exploits economies of scale, but in this case by supporting not one but several data models, all of them effective ways of representing the information content of a map or image. The real geographic world is enormously complex, and there are many different ways of capturing its contents in digital form, all of

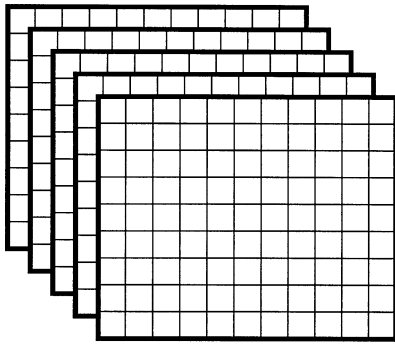


Figure 1

The concept of layers. Geographic reality is represented by a set (five in this case) of co-registered rasters, each representing the spatial pattern of one variable. Example layers might include soil type, vegetation cover type, ownership, mean annual temperature, or elevation

them involving some form of approximation. Remotely sensed images use a raster model, capturing the earth's reflectance in each of an array of square cells. Because of its simplicity, the raster model has been used as the basis for many GISs, each of them representing data as a series of layers of co-registered raster cells (Fig. 1), with each layer capturing some distinct variable, such as land elevation, county name, average income, population, or class of land use.

In the vector model the world is viewed as occupied by collections of points, lines, and areas. Points are located by coordinates, often latitude and longitude or some universal coordinate system such as the Universal Transverse Mercator (UTM). Lines are represented as polylines by connecting an ordered sequence of points with straight lines. Areas are similarly represented as polygons or figures with point vertices connected by straight edges. Each point, line, or area

object in the vector model is associated with one or more attributes which capture the object's significant characteristics. A set of objects of the same type is associated with an attribute table by assigning one row of the table to each object, and one column to each characteristic.

Figure 2 shows a simple example of raster and vector data. The same set of census tracts is shown in Fig. 2(a) in the form of a raster, and in Fig. 2(b) in vector form. Most GISs include support for both models, and for transformations between them. They also support various kinds of map projections, allowing the user to switch readily between latitude and longitude, UTM, and other coordinate systems, and to display any part of the earth's surface in the form of a flat paper map.

One of the most compelling arguments for GIS relies on the ability to integrate information around a common geographic location. A project might involve many different kinds of information about different themes or variables for a geographic area. Every point, line, area, or raster cell in a GIS database has an associated location, so it is possible to determine the distance between two locations or to place locations correctly on a map. But more significantly, it is also possible to identify all of the information available about a location, or to link together two data sets through common location. GISs include functions to determine how much overlap exists between two areas, or whether a given point lies inside a given area, and functions to select all of the points lying within a given distance of a line. To return to a point made earlier, the economics of software production provide great advantages to GIS designers once a basic structure has been laid down, because new functions that work on the same structure can be added very easily.

Certain key GIS functions are particularly relevant to social and behavioral science applications.

(a) Geocoding, or the ability to convert street addresses to geographic coordinates, and vice versa.

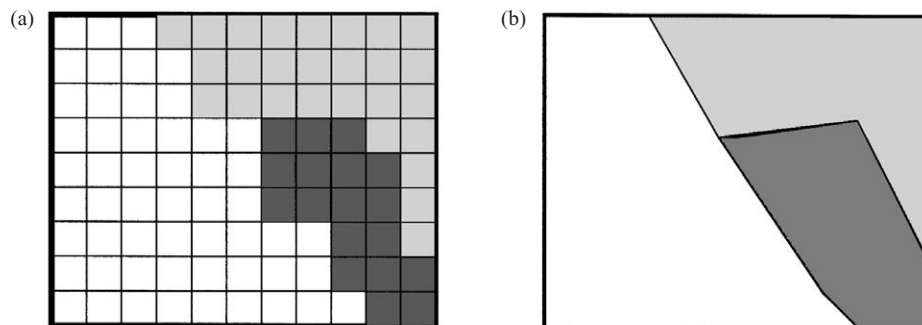


Figure 2

Representation of three adjacent census tracts using raster cells (a) and vector polygons (b). In the raster case, cell size determines the accuracy of boundary positions. In the vector case, each tract is located by specifying the coordinates of its vertices in order

This function might be used to associate coordinates with surveyed households, allowing locations to be mapped, integrated with other data, or subjected to various kinds of analysis.

(b) The point-in-polygon function, or the ability to compare a set of points to a set of polygons, to identify the polygon containing each point, and to combine the characteristics of the containing polygon with those of the point. For example, this function might be used to associate census tract properties with those of the individuals who live there.

(c) The polygon overlay function, or the ability to combine two sets of polygons, creating a new polygon for every unique area of overlap, and associating it with the attributes of the two polygons that overlap to create the area. By overlaying a map of county boundaries on a map of school districts, it would be possible to estimate the number of children living in each school district based on knowledge of the population density of each county (subject to the assumption that population densities are uniform within each county).

(d) The buffer function, which identifies the area lying within a specified distance of a given point, line, or area, and gives it the attributes of this object. By combining this with the geocoding and point-in-polygon functions, it would be possible, for example, to determine the addresses of households within a certain distance of a shopping center.

3. GIS as Research Infrastructure

GIS software first became widely affordable in the late 1980s, when personal computers powerful enough to manipulate geographic data became available for a few thousand dollars. Social and behavioral scientists began to invest in its capabilities, particularly in those disciplines that had traditionally adopted a spatial perspective. Archaeologists and anthropologists, for example, began to use GIS software to map and store artifacts at the scale of the dig, to build digital maps of reconstructed landscapes, and to make use of spatial analysis to understand how early societies operated in space (Aldenderfer and Maschner 1996). Criminologists use GIS to map crime, correlate crime rates with social variables, and examine detailed spatio-temporal patterns of criminal behavior. Political scientists use GIS to evaluate districting schemes, and to explore the spatial diffusion of political movements (Eagles 1995). Market researchers and location analysts use GIS to map market potential, model the spatial behavior and choices of consumers, and search for optimal locational strategies (Martin 1996).

Of all the social and behavioral disciplines, geographers have long claimed a special interest in space, and much of the early work on GIS design and application was strongly associated with the discipline.

Geography is often the home department of GIS, though many other arrangements exist, especially at institutions lacking a geography department. GIS courses are frequently offered by geography departments, and geography is often the sponsor of campus GIS infrastructure, in the form of laboratories, software licenses, and data repositories.

In recent years efforts have been made to develop wider linkages. In the USA the University Consortium for Geographic Information Science (www.ucgis.org) now links over 60 member institutions in an organization dedicated to advancing GIS and its applications through research and education. In Europe, the Association of Geographic Information Laboratories in Europe (AGILE) has a similar mission, and the European Science Foundation has sponsored a major study of GIS in the social sciences through its GISDATA program. National centers have been funded in several countries to pursue GIS research: in the USA, the National Center for Geographic Information and Analysis (www.ncgia.org) was founded by the National Science Foundation in 1988, while in Canada the GEOIDE network was founded by the three national research councils in 1998 (GEOIDE, *La géomatique pour des interventions et des décisions éclairées*; see www.geoide.ulaval.ca).

Several journals are devoted to advances in GIS, including the *International Journal of Geographical Information Science*, *Cartography*, and *Geographic Information Systems*; *Geoinformatica*, the *Journal of Geographical Systems*; and *Transactions in GIS*. Descriptions of research using GIS can be found in major journals in many application disciplines.

Comparisons have often been drawn between GIS and statistics, since both can be seen as generic tools and concepts used widely across the sciences (e.g., Goodchild 1987). It is argued that GIS provides the implementation of methods of spatial data analysis in easy-to-use computer tools, just as the statistical packages implement methods of statistical analysis. In this context spatial data analysis can be defined as a set of analytic methods whose results are not invariant under changes in the locations of the objects of analysis—in other words, location matters. This serves to distinguish these techniques from more familiar numerical and statistical analyses, such as the calculation of a mean, or the preparation of a scatterplot, that do not normally include location or respond to it. Spatial analysis predates GIS, and major texts include those by Berry and Marble (1968), Haining (1990), and Bailey and Gatrell (1995). However, the advent of GIS has greatly improved access to these methods and, moreover, it has come at a time when there is increasing interest in visual, intuitive, and exploratory techniques, all characteristics that are readily compatible with GIS. For discussions of the impact of GIS on spatial analysis see, for example, Fotheringham and Rogerson (1994), Goodchild and Longley (1999), and Fischer et al. (1996).

4. Sources of Data

Maps are highly compressed representations of large amounts of information, and extremely tedious to convert to digital form. Two methods are available, with both advantages and disadvantages:

(a) Digitizing, a process in which a human operator traces information from a map, capturing the locations of features and their attributes using a pointing device and a keyboard. The process is very expensive, since it is difficult to capture complex geographic features rapidly and accurately.

(b) Scanning, in which an automated device captures the information from a map. Extensive editing and error correction is necessary, especially if the map is of low quality, such that scanning is typically as expensive as digitizing.

In the social and behavioral sciences, data are often highly specific to applications and must be rendered into digital form using one of these methods. Thus an archaeologist may have to digitize the locations of sites or artifacts from manuscript maps or drawings. In other cases, however, the process of database creation will take advantage of the widespread availability of framework data, or geographic data sets of generic use that already exist in digital form. These include:

(a) Topographic maps, in both raster and vector form (produced in the USA by the US Geological Survey and known as digital raster graphics, or DRGs, and digital line graphs, or DLGs. Various mapping scales are commonly available (in the USA the most detailed scale is, 1:24,000).

(b) Elevation data in raster form (produced in the USA by the US Geological Survey and known as digital elevation models, or DEMs). These data sets vary according to the spacing of the grid: for the entire earth the best spacing available is several kilometers, but 30m spacing is available for much of the USA.

(c) Street centerline data in vector form. These data sets record the locations, names, address ranges, and other information on streets. The TIGER (Topologically Integrated Geographic Encoding and Referencing) database created and maintained by the US Bureau of the Census covers all of the USA, and includes boundaries of statistical reporting zones. These data sets represent streets by single lines and have positional accuracies of tens of meters.

(d) Imagery from aerial photography. The Digital Orthophoto Quad (DOQ) is typical of these data sets, with a 1m ground resolution and a 6m positional accuracy, and available from the US Geological Survey for much of the USA.

(e) Imagery from satellites. These sources include images collected by the Landsat and SPOT satellites, with cell sizes down to 10m.

(f) Municipal databases created by local government agencies, often in collaboration with utility companies. These are in vector form and often include

property boundaries, locations of buildings, and streets represented by double lines. In some countries, notably the UK, they are produced by government for the entire land area. In the USA, coverage is patchy. Positional accuracies are typically in the meter range.

(g) Gazetteers, the digital versions of the indexes found in atlases, linking placenames to geographic coordinates. These indexes make it possible to translate information recorded by placename to map form, for analysis and display.

Many of these data sets are created and maintained by longstanding programs of national mapping agencies. Recently, many countries have attempted to place them within broader policy frameworks that emphasize sharing of data, avoidance of duplication, and the setting of standards. In the USA, the National Spatial Data Infrastructure (National Research Council [NRC] 1993) is a leading example.

These data sets provide the framework on which social and behavioral scientists can often hang their own features and attributes. For example, a criminologist might use a municipal database, adding records of crime locations; a sociologist might use reporting zone boundaries from the TIGER database to create maps and analyze data at the census tract level; and a planner might use imagery to monitor urban growth. By using data that are already in digital form the cost and tedium of database creation are greatly reduced.

Unfortunately, the conditions under which such framework data sets are accessible vary widely. In the USA, policy requires that geographic data produced by the federal government be distributed to anyone at no greater cost than that of dissemination. It is often argued that this policy has stimulated growth in the use of GIS and in the geographic information industry generally. But while US courts have held that geographic data are facts and therefore not subject to copyright, in other countries national governments have adopted cost-recovery policies, protected by copyright. Academic scientists in such countries have found various ways of accessing geographic data at lower cost, through bulk purchase agreements. Lower levels of government in the USA are also not subject to the cost-of-dissemination policy, and have made efforts at cost recovery under trademark protection. For more discussion of these issues, see *Spatial Data Infrastructure*.

While the public sector has traditionally supplied many of the geographic data used by the scientific community, a growing private sector may have substantial impact in the coming years. Very large geographic databases are being created for internal purposes by insurance, retail, banking, and other corporate sectors. Point-of-sale systems now routinely capture data on grocery purchases, and direct-mail and delivery companies make use of detailed geographic data for scheduling and targeting. These activities have raised privacy concerns, and it is clear

that the levels of detail now available to corporations, together with the ability to integrate and merge data based on geographic location, have very significant privacy implications. Much of this activity remains largely outside any regulatory framework.

5. Networks

The Internet and WWW have had major impact on GIS. They have made it much easier to share data and to access data remotely. Data can be published and disseminated easily, as can the more sophisticated results of analysis. Indeed, the entire process of scholarly communication and publication is being radically restructured by networking technologies.

It is useful to distinguish two forms of processing of geographic information. The term 'geographic information system' suggests a combination of hardware and software that may or may not be connected to a larger network. It is probably controlled directly by a single user, and provides capabilities and functions directly related to the user's needs. On the other hand the term 'geographic information services' is used to describe WWW sites that offer specific GIS functions in response to user requests. Instead of sending the data to a GISystem, a GIService retains its data, sending only the results of analysis. The Mapquest site (www.mapquest.com), for example, provides a range of services from the generation of maps at user request to the geocoding of addresses supplied by the user.

GIServices are growing rapidly, and are expected to grow much further in the coming years. In addition to avoiding lengthy transmission of data, they allow the custodian of a sophisticated analysis tool to retain control, and to upgrade the tool, rather than undertake a complex process of distribution to user sites. Pricing based on services is very different, and perhaps simpler, than pricing based on data (Guenther and Mueller 1999). GIServices also assume a level of adherence to standards, and encourage the development of interoperable tools using common formats and protocols (the Open GIS Consortium, www.opengis.org, is a major influence toward greater interoperability in GIS).

Several major initiatives are making geographic data more easily found and accessed via the WWW. In the USA, the Federal Geographic Data Committee sponsors the National Geospatial Data Clearinghouse (www.fgdc.gov), a catalog of sites offering geospatial data, linked together using the Z39.50 protocol. The Alexandria Digital Library (alexandria.ucsb.edu) offers the services of a map library, including search and retrieval of data, over the WWW. Other sites offer more specialized data, for example, a site at the Massachusetts Institute of Technology (ortho.mit.edu) offers DOQs for the Boston area. The term 'geolibrary' has been coined to describe WWW sites that allow search for data based on geographic location, including search for information that is associated with

a geographic area, but which is not a map or an image, such as a planning report. The implications of geolibraries are discussed in a recent report of the US National Research Council (NRC 1999).

6. Space and Scientific Explanation

Ultimately, the value of GIS and related technologies in the social and behavioral sciences will be intimately tied to the role of space in scientific explanation and understanding. Before expanding on that point, however, it is useful to distinguish between normative and explanatory applications. Many applications of GIS may be scientific in their objectivity and rigor, but nevertheless aimed at supporting decisions, developing policy, or solving day-to-day problems. Such normative efforts often apply the principles and knowledge of social and behavioral science. Moreover, many do so in a spatial context, by combining principles about behavior and society with the specific conditions existing in a geographic area. Thus, there are abundant normative applications of GIS.

In this section, however, the focus is on those applications that advance scientific knowledge about the world, through observation, experiment, hypothesis-testing, and deduction, because such explanatory efforts lie at the heart of the social and behavioral sciences. To what extent can GIS be used as a tool to advance scientific knowledge?

Ideally, scientific knowledge advances through controlled experiment, by varying the conditions under which society operates and observing the response. But although such experiments are common in areas such as agriculture, and GIS is often used as a tool to manage experimental plots, they are unlikely to be possible in the case of social and behavioral processes operating at geographic scales. Instead, science must advance through inference, either inductively by inferring the nature of a process from its geographic or temporal expression, or deductively by comparing the predictions of theory with reality.

GIS is most strongly associated with cross-sectional approaches to observation and inference. A phenomenon is compared across a geographic area, and its spatial variation is compared with the variation in relevant conditions. Inferences are drawn about process, but since the process is not observed directly it is impossible for inference to be definitive. Alternatively, patterns may be used to provide confirmation of hypothesized processes, though again confirmation is never definite. The tradition of using maps as a basis for inference has a very long history, one of the most compelling historic examples being the map made by Dr. John Snow of a cholera outbreak in the Soho district of London in the 1850s, which provided evidence of the role of drinking water as the disease vector.

It is often argued that longitudinal observation and inference based on change through time is more

definitive than cross-sectional inference. But cross-sectional data are widely available from sources such as the census, and powerful technologies like GIS increasingly make it possible to examine data from both cross-sectional and longitudinal perspectives.

Geographers draw a distinction between the properties of a site, and the site's situation in spatial context. Many social processes respond both to the conditions at a place, and to the conditions in the place's surrounding area. For example, choice of residential location is affected both by the properties of the available sites, and by their surrounding communities and neighborhoods. GIS makes it possible to observe and analyze a site's situation, in the form of maps and images, and also to link one property of a site to other properties through their common geographic location. Further discussions of the spatial dimension of explanation, and of the role of GIS, are available in many sources (e.g., Abler et al. 1992, NRC 1997a).

7. GIS in the Future

There is every indication that the use of computers to handle geographic information will continue to increase dramatically in the future, and that the user base will continue to broaden. But trends in the software industry, and in society more generally, suggest that the nature of GIS will continue to evolve.

In 1997 the US National Research Council sponsored a workshop on the Future of Spatial Data and Society (NRC 1997b). It found strong support for continued increases in technical power, in the form of computer speed, network bandwidth, and storage capacity, at constant or diminishing cost. But it also identified a large number of societal trends that are likely to affect GIS, and also fundamental trends in the nature of software.

The report found a discernible trend in recent years towards the solution of problems at the local level. It is reflected, for example, in the popularity of place-based decision making in the US Environmental Protection Agency, a policy that encourages the solution of problems locally, by local stakeholders. It is also reflected in efforts to devolve power to lower levels in the administrative hierarchy.

In the 1960s, the so-called quantitative revolution in geography emphasized the nomothetic search for general laws affecting spatial phenomena, and saw traditional geography's approach as idiographic, with its stress on cataloging the unique properties of places. Today, there is much interest in quantitative techniques that accommodate spatial variation in parameters, and emphasize the role of spatial analysis in exploring local variation. Fotheringham, for example, has identified this as one of the most important themes in contemporary spatial analysis (Fotheringham 1997). GIS provides an ideal platform for such methods, since it makes it easy to analyze spatial data, and to display spatial variation.

At the same time, there is increasing interest in several disciplines in the incorporation of spatial effects directly into theory, on the grounds that a system that is spatially distributed, and forced to communicate over distance, is fundamentally different from one that is infinitely connected. Spatial economics and spatial ecology are now significant areas within the theoretical wings of those disciplines, while geography has a long tradition of explicit theorizing about space. Again, GIS provides an ideal platform for subjecting such spatially explicit theory to empirical test.

Finally, the spatially explicit nature of GIS provides an ideal environment for linking normative and explanatory science. The implications of theories for real geographic areas are readily evaluated, and models that incorporate social and behavioral processes can be combined with geographic boundary conditions to provide readily visualized predictions. Thus, GIS offers an ideal environment for bridging the gap between theory and practice, and a significant clue in the eternal search for relevance.

See also: Computers and Society: Critical Approaches; Spatial Analysis in Geography; Spatial Pattern, Analysis of; Spatial Thinking in the Social Sciences, History of; Technological Innovation

Bibliography

- Abler R F, Marcus M G, Olson J M (eds.) 1992 *Geography's Inner Worlds: Pervasive Themes in Contemporary American Geography*. Rutgers University Press, New Brunswick, NJ
- Aldenderfer M, Maschner H D G 1996 *Anthropology, Space, and Geographic Information Systems*. Oxford University Press, New York
- Bailey T C, Gatrell A C 1995 *Interactive Spatial Data Analysis*. Longman Scientific and Technical, Harlow, UK
- Berry B J L, Marble D F 1968 *Spatial Analysis: A Reader in Statistical Geography*. Prentice Hall, Englewood Cliffs, NJ
- Burrough P A, McDonnell R A 1998 *Principles of Geographical Information Systems*, 2nd edn. Oxford University Press, New York
- Clarke K C 1997 *Getting Started with Geographic Information Systems*. Prentice Hall, Upper Saddle River, NJ
- DeMers M N 1997 *Fundamentals of Geographic Information Systems*. John Wiley, New York
- Eagles M (ed.) 1995 *Spatial and Contextual Models in Political Research*. Taylor and Francis, London
- Fischer M, Scholten H J, Unwin D J (eds.) 1996 *Spatial Analytical Perspectives on GIS*. Taylor and Francis, London
- Foresman T W (ed.) 1998 *The History of Geographic Information Systems: Perspectives from the Pioneers*. Prentice Hall PTR, Upper Saddle River, NJ
- Fotheringham A S 1997 Trends in quantitative methods I: stressing the local. *Progress in Human Geography* 21: 88–96
- Fotheringham A S, Rogerson P A (eds.) 1994 *Spatial Analysis and GIS*. Taylor and Francis, London
- Goodchild M F 1987 A spatial analytic perspective on geographic information systems. *International Journal of Geographical Information Systems* 1: 327–34

- Goodchild M F, Longley P A 1999 The future of GIS and spatial analysis. In: Longley P A, Goodchild M F, Maguire D J, Rhind D W (eds.) *Geographical Information Systems: Principles, Techniques, Management, and Applications*, Wiley, New York, pp. 567–80
- Guenther O, Mueller R 1999 From GISystems to GIServices: spatial computing on the Internet marketplace. In: Goodchild M F, Egenhofer M J, Fegeas R, Kottman C A (eds.) *Interoperating Geographic Information Systems*. Kluwer Academic, Norwell, MA, pp. 427–42
- Haining R P 1990 *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press, Cambridge, UK
- Heywood I, Cornelius S, Carver S 1998 *An Introduction to Geographical Information Systems*. Addison-Wesley Longman, New York
- Jensen J R 1996 *Introductory Digital Image Processing: A Remote Sensing Perspective*, 2nd edn. Prentice Hall, Upper Saddle River, NJ
- Kennedy M 1996 *The Global Positioning System and GIS*. Ann Arbor Press, Chelsea, MI
- Martin D J 1996 *Geographic Information Systems: Socio-economic Applications*, 2nd edn. Routledge, London
- National Research Council 1993 *Toward a Coordinated Spatial Data Infrastructure for the Nation*. National Academy Press, Washington, DC
- National Research Council 1997a *Rediscovering Geography: New Relevance for Science and Society*. National Academy Press, Washington, DC
- National Research Council 1997b *The Future of Spatial Data and Society*. National Academy Press, Washington, DC
- National Research Council 1999 *Distributed Geolibraries: Spatial Information Resources*. National Academy Press, Washington, DC

M. F. Goodchild

Geographic Information Systems: Critical Approaches

Geographic information systems are computerized information systems, computer hardware and software, used to store, manipulate and map geo-referenced databases of information about the physical and human world. Growing rapidly since their development in the 1970s, GIS are now used in a wide variety of private and public organizations, and increasingly by individuals and grassroots organizations, for analysis, decision-making, and the display of geographical information. In response to this growing influence, critical assessments of GIS have emerged. These have analyzed their positivist bias, their potential to reinforce social inequality, their capacity for surveillance and ethical and legal implications, and have sought to examine these issues and to pursue alternative approaches to contemporary GIS. Critical approaches refer to analyses of GIS that are based in critical theory, i.e., that assess GIS in terms of their relationship to power hierarchies and wealth inequalities in society (Habermas 1984). This article:

- traces the origins of such critiques of GIS;
- discusses the engagement between critical theorists and GIS researchers that resulted in development of a GIS and society research agenda;
- examines critical assessments of the impact of society on GIS, and of GIS on society; and
- summarizes attempts to seek alternatives to GIS that may reduce some of the problems identified through these critical approaches.

1. Early Critiques

Broad-based claims were made about GIS as the key to a new integrated scientific geography at the end of the first wave of GIS research. This was the point in time when the technology had stabilized around a layer-based GIS architecture, and when the US National Science Foundation had recognized the importance of GIS through establishment of the National Center for Geographic Information and Analysis (cf. Dobson 1983). Written at a time when human geographers were moving away from positivist and critical rationalist approaches to research, such statements catalyzed a series of critical responses (Pickles 1999).

These criticisms focused on both epistemological and practical implications of the spread of GIS. Epistemologically, GIS was argued to be a Trojan horse for the reassertion of broadly positivist approaches within geography, because of its quantitative and empirical nature. Critics argued that geography was too complex and varied a discipline to be analyzed with this approach alone, criticized the claims that positivist science is foolproof, universal and objective, and claimed that the use of GIS would marginalize critical theoretic and hermeneutic approaches to geography, thereby limiting its ability to make sense of the world. They also argued that this kind of scientific approach would not promote human emancipation, and would not leave space for investigation of other possible worlds than those observed and experienced. The benefits for geography of using GIS to enhance its status as a science were seen to be outweighed by the costs associated with a consequent narrowing of geographic inquiry.

It was also argued that embedded within GIS are certain conceptions of space (particularly a geometric and relative space) and certain forms of reasoning (particularly Boolean logic) that are unable adequately to represent non-European conceptions of space and communicative rationality (Habermas 1984, Rundstrom 1995, and Sheppard 1995). Practically, it was argued that increased use of GIS in society would enhance current social and geographical inequalities. It was noted that social actors have differential access to GIS; wealth, gender, race, and geographic location are all strongly correlated with access to the necessary equipment, access to training, and comfort in working

with GIS. Consequently, GIS was seen as facilitating practices of surveillance, social engineering, opinion formation, and warfare by those with access to the technology. These arguments closely parallel more general analyses of the digital divide. These critical attacks provoked equally sharp responses by GIS specialists.

These specialists saw the critiques as simplistic, unduly pessimistic, and even paranoid, and as reflecting a lack of understanding of and experience with GIS and/or a lack of patience or aptitude for the rigors of science. They also resented the implication that GIS specialists are unconcerned with the social issues raised, and/or unaware of the social implications of science. At the same time, they saw such criticisms as a minor irritation that could be ignored given the growing interest in GIS among both students and practitioners. As a consequence, after some initial high profile exchanges, from 1983 to 1993 there was little communication between the two 'cultures of indifference': those critical of and those specializing in GIS (Pickles 1999).

2. Engagement: GIS and Society

This intellectual divide was challenged in 1993, when the National Center for Geographic Information and Analysis sponsored a conference organized by both GIS specialists and social theorists who sought a more constructive engagement. This and a subsequent conference in 1995 brought together prominent researchers (Poiker and Sheppard 1995 and Harris and Weiner 1996). Notwithstanding some early tensions, caricatures broke down as participants gained a greater appreciation for the breadth of skills and interests of those coming from the other side. A common desire among those present to learn from one another stimulated development of a GIS and society research agenda composed of the following themes:

- the social history of GIS as a technology;
- the relevance of GIS for community and grassroots perspectives and lifeworlds;
- issues of privacy, access to spatial data, and ethics;
- the gendering of GIS;
- GIS, environmental justice and political ecology;
- GIS and the human dimensions of global change;
- and
- alternative approaches to GIS.

Formal collaboration continued, as collective research catalyzed by these meetings, in jointly organized sessions at GIS and geography conferences, and within the 'Geographies of the Information Society' research initiative workshops between 1997 and 1999 (Sheppard et al. 1999). In addition, critical theorists were asked to participate in key GIS events, and GIS theorists were given voices in critical theory projects (cf. Pickles 1995). By the end of the 1990s, the constructive engagement between GIS specialists and

their critics had substituted overlapping cultures of respect for separate cultures of indifference, notwithstanding the continuing reluctance of many critical geographers or GIS specialists to engage with one another.

Tensions between different perspectives remain. For example, the renaming of GIS as Geographic Information Science revived concerns about attempts to reassert an agenda of normal, positivist science with all its trappings (Pickles 1997). Yet the tone of debate had shifted to a search for common ground; much to the relief of graduate students who had been uncomfortable about having to choose between GIS and critical geography (Sheppard et al. 1999). With the shift towards more constructive engagement has come a shift from critique, in which possibilities were suggested and individual cases cited to bolster their plausibility, to analysis of the GIS and society nexus, which attempts to trace the complex ways in which each influences the other in specific contexts. While the GIS and society research agenda consistently has pointed to the dialectical relationship between the two, more research effort has been put into the impact of GIS on society than vice versa.

3. From Society to GIS

The starting point of this research has been the proposition that, as for technological change in general, the evolution of GIS has followed only a restricted number of the many paths that could have developed (cf. Arthur 1994). On the one hand, research has investigated the evolution of the software and hardware itself, the rise to dominance, for example, of a layer-based GIS over object-oriented GIS. This research has closely examined processes of technological development at key sites and moments when paths of future development were set, and has speculated on the 'paths not taken'; the alternative designs that never got off the ground or never became popular. The roles of key individuals, technical barriers, disciplinary tradition, and broader social forces in shaping paths of development have been analyzed (Chrisman 1988 and Curry 1998). Critical research has also stressed that the practice of GIS includes not just the technology itself, but the whole set of practices through which it is utilized, adapted, and connected with other related activities on a daily basis in the particular places where it is used. Drawing on insights from actor network theory and science studies, this approach has focused on how users, as well as inventors, go beyond learning-by-doing to shape the technology (Harvey and Chrisman 1998).

4. From GIS to Society

Critical approaches to examining the societal implications of GIS have focused on case studies examining

how some of the hypothesized implications are realized in particular contexts. A number of the implications raised in critiques of GIS have been examined:

- limits to GIS representations of the world;
- limits in access to and in the appropriateness of GIS technologies;
- legal and ethical implications of GIS use; and
- the applicability of GIS for redressing social and geographical inequalities.

Those examining the limits of GIS representations of the world have taken seriously the criticism that GIS, as we know it, undergoes a process of representational stabilization, foregrounding representations which prioritize a Cartesian view of the world, reify geometric spatial objects, and rely on institutional databases for their information. More generally, such criticisms contrast the instrumental rationality of GIS with the communicative rationality of lifeworlds, and conclude that GIS has difficulty capturing the latter. Rundstrom (1995), for example, argues that American Indian conceptions of space are inconsistent with those used in GIS, concluding that it is inappropriate for representing Indian lifeworlds. Harris et al. (1995) experiment with incorporating into conventional GIS sketch maps made by local peasant farmers in South Africa, in order to determine whether the capabilities of GIS can be extended to incorporate lifeworld-based information. The sketch maps represent African farmers' views of the landscape and their roots in local narratives about land alienation under apartheid. Goss (1995) examines GIS-based representations used in geo-demographic marketing, and how these may shape lifeworlds and the places in which they are pursued.

Those examining limits in access to and in the appropriateness of GIS have often begun with the observation that GIS was developed initially for large organizations, universities, corporations, state agencies, and the military, raising questions about its appropriateness for other (more marginal) social actors. Public participation GIS (PPGIS) research has been examining whether GIS, as we know it, is accessible to grassroots organizations; whether it is appropriate to the goals of such organizations, and thus to their empowerment; and whether it enhances or undermines the participatory decision-making processes characteristic of such organizations (Obermeyer 1998). Barriers to the use of GIS by grassroots organizations have been identified, and strategies to overcome these have been suggested. Studies of the impact of GIS use on the empowerment of, and on democracy within such organizations are beginning. Research into legal issues has focused on undesirable social implications of the spreading use of GIS and geo-referenced databases, for which legal remedies may be desirable. It has also focused on how the laws governing information policy and intellectual property shape the social implications of GIS use. Prominent social implications include:

violations of privacy resulting from the capability to map individual actors and events (such as AIDS cases);

the consequences of inappropriate GIS use, and the question of who bears legal responsibility; and

reduced public access to publicly funded spatial databases, as a consequence of intellectual property laws and the cost recovery strategies of local government agencies (cf. Curry 1998 and Sheppard et al. 1999).

Legal questions frequently revolve around ethical issues, but ethical implications of GIS are far broader than this. Curry (1998) suggests not only that there are significant ethical issues revolving around GIS use, but indeed that a close examination of GIS reveals that conventional models of ethical behavior are themselves problematic. The use of GIS to redress social inequalities has been dominated to date by GIS-based analyses of environmental equity (Nyerges and McMaster 1997). Much of this research employs GIS, as we know it, in fairly conventional ways to gather empirical information about the significance, pervasiveness, and correlation of social and geographical inequities in the potential exposure to toxic hazards. Considerable complexities in the relationship between exposure, social characteristics, and location have been revealed, and measures for determining their significance explored. Research into the historical processes behind such patterns has helped unravel the roles of social actors, and of race and class in creating such patterns (Pulido 2000). Some research is beginning to go beyond this, exploring ways to make information about exposure available to communities, and examining the utility of GIS as a tool for neighborhoods to carry out environmental inventories and develop sustainability indicators.

5. Prospects

Critical approaches to GIS are an even newer field of endeavor than GIS themselves, although they draw their inspiration from longer standing traditions of critical social theory and critical geography. The growth of the information society, and the central role of GIS and GIS-related information technologies (such as navigation systems, geo-referenced cellular phones, geographical positioning systems, and on-line real time mapping systems) in shaping the geography of the information society, mean that the questions raised in this research will remain important. Yet, the novelty of this field and its considerable potential for growth make it hard to predict major areas for future research. In addition to further elaboration of current research, however, two topics are likely to become increasingly important. First, the path-dependent nature of how GIS developed, and concerns about its limitations, suggest that an important issue, raised in the early 1990s but as yet little examined, is the potential for using digital technologies to convey

geographic knowledge in ways that are very different from those instantiated in GIS as we know it (Harris and Weiner 1996). Possibilities of alternative GISs have been broached, based perhaps on forms of rationality and information and opinion exchange that have more in common with the communicative logic of the Internet than with the instrumental logic of current GIS (Harris and Weiner 1996). Little research has been done, however, examining how such systems might be developed, and whether their social implications are likely to be more emancipatory than current GIS designs.

Second, we know little about reasons for inequalities in access to and use of GIS as yet. While the great majority of GIS researchers and users are white males, there has been little research into why this is the case as yet, and how it might be mitigated. Research taking a critical perspective on GIS has matured from postulating significant potential problems and limitations, to some selected case studies examining actual interrelationships between GIS and society in individual concrete cases. Such case studies are a vital step forward, and much more of this research is necessary. Yet a fuller understanding of these interrelationships will require comparative case study analysis, to tease out which kinds of contextual conditions affect which kinds of outcomes. Research designs should compare both case studies of similar organizational contexts in different places, *inter alia*, and case studies of different organizational contexts in similar places.

See also: Computers and Society; Geographic Information Systems; Spatial Analysis in Geography; Spatial Pattern, Analysis of; Spatial Statistical Methods; Spatial Thinking in the Social Sciences, History of

Bibliography

- Arthur W B 1994 *Increasing Returns and Path Dependence in the Economy*. The University of Michigan Press, Ann Arbor, MI
- Chrisman N R 1988 The risks of software innovation: A case study of the Harvard Lab. *The American Cartographer* **15**: 291–300
- Curry M R 1998 *Digital Places: Living with Geographic Information Technologies*. Routledge, London
- Dobson J E 1983 Automated geography. *The Professional Geographer* **35**: 135–43
- Goss J 1995 We know who you are and we know where you live: The instrumental rationality of geodemographic systems. *Economic Geography* **71**: 171–88
- Habermas J 1984 *The Theory of Communicative Action, Vol. 1: Reason and the Rationalization of Society*. Beacon Press, Boston, MA
- Harris T, Weiner D 1996 GIS and society: The social implications of how people, space and environment are represented in GIS. *Report for the Initiative 19 Specialist Meeting, March 2–5, 1996*, Koinonia Retreat Center, South Haven, MN. Technical Report 96–7, National Center for Geographic Information and Analysis, Santa Barbara, CA
- Harris T M, Weiner D, Warner T, Levin R 1995 Pursuing social goals through participatory GIS: Redressing South Africa's historical political ecology. In: Pickles J (ed.) *Ground Truth: The Social Implications of Geographic Information Systems*. Guilford, New York
- Harvey F, Chrisman N R 1998 Boundary objects and the social construction of GIS technology. *Environment and Planning A* **30**: 1683–694
- Nyerges T, McMaster R (eds.) 1997 GIS and environmental risk assessment. A special section of *Cartography and Geographic Information Systems* **24**(3)
- Obermeyer N (ed.) 1998 Public participation GIS. A special issue of *Cartography and Geographic Information Systems* **25**(2)
- Pickles J (ed.) 1995 *Ground Truth: The Social Implications of Geographic Information Systems*. Guilford, New York
- Pickles J 1997 Tool or science? GIS, technoscience, and the theoretical turn. *Annals of the Association of American Geographers* **87**: 363–72
- Pickles J 1999 Arguments, debates and dialogues: The GIS-social theory debate and concerns for alternatives. In: Longley P A, Goodchild M, Maguire D, Rhind D (eds.) *Geographical Information Systems: Principles, Techniques, Applications, and Management*. Wiley, New York
- Poiker T, Sheppard E (eds.) 1995 GIS and society. A special issue of *Cartography and Geographic Information Systems* **22**(1)
- Pulido L 2000 Rethinking environmental racism: White privilege and urban development in Southern California. *Annals of the Association of American Geographers* **90**: 12–40
- Rundstrom R A 1995 GIS, indigenous peoples, and epistemological diversity. *Cartography and Geographic Information Systems* **22**: 45–57
- Sheppard E 1995 GIS and society: Towards a research agenda. *Cartography and Geographical Information Systems* **22**: 5–16
- Sheppard E, Couclelis H, Graham S, Harrington J R, Onsrud H 1999 Geographies of the information society. *International Journal of Geographic Information Science* **13**: 797–823

E. Sheppard

Geographic Learning in Children

1. Early Childhood Development and Geographic Learning

Most research on early childhood geographic learning has been concerned with the ability of the developing child to understand space, and in particular, to comprehend spatial location. Children begin to be capable of spatial perception—the perception of small-scale spaces they can see at once—shortly after birth. Not long thereafter, they begin to crawl around, and slightly later to toddle through rooms filled with obstacles that impede perception. It is at these slightly later ages that geographic learning really begins, when a child goes from merely perceiving to actively searching the environment. Between 12 and 18 months

of age, children can find objects once seen and now hidden (by a screen, for example) from view. Children between 18 months and two years of age can also deal with 'invisible displacements,' in which a toy, for example, is not just concealed but 'invisibly' transported to another hiding place. Two- and three-year olds can carry out much more complicated searches. However, at these age levels most research has been limited to the scale of an ordinary room, or slightly larger (Matthews 1992).

There is now strong evidence that children acquire basic map-reading skills and the ability to engage in certain aspects of map-reading somewhere between the ages of three and five (see Blades and Spencer 1990, DeLoache 1989, Newcombe and Huttenlocher 1992). Studies have focused on two general levels of spatial ability: (a) micro- or meso-spatial contexts—i.e., table-scale or room-scale representations (e.g., Cohen 1985), and (b) macro- or geographical-scale environments. Micro-spatial cognition involves understanding spatial relationships among objects that can, for the most part, be directly perceived as themselves and at their own scale. For example, micro-spatial cognition is required to put together a jigsaw puzzle. Macro-spatial cognition, on the other hand, involves representations at a perceptual scale of objects and object arrays that occupy much larger spaces in the 'real world,' objects that, together, would not ordinarily be perceived at once on the ground, for example, reading a map to navigate through a city.

The most interesting aspects of spatial development and geographic learning concern the ability of children to understand macro-spatial environments, which are large in extent, and their representation in the form of maps (Blades and Spencer 1990). Almost all research in this area has involved miniature representations of such environments in the form of aerial photos, maps, and modeled landscapes. At issue is how to ascertain very young children's understanding of the connection between these miniature representations and full-scale geographic reality. Children younger than five have limited vocabularies, and efforts to use verbal responses as indices of this understanding have not been very successful.

However, ethnographic studies done in several cultures have begun to document the universality of the use of toys or natural objects to model large-scale environments (Blaut 1999, Stea et al. 1996). It has been noted that some three-and-a-half-year-old children are able to build model landscapes and to solve navigation problems on these landscapes. Such navigation problems involve getting from one place to another on the map in a 'realistic' manner, that is, by a route on the representation that would be reasonable in the 'real world'. However, a series of studies have found that while some spatial skills (micro-spatial puzzle and block-building ability; macro-spatial navigation ability) are present as young as age three-and-a-half, they do not necessarily appear and develop at

the same rate. This indicates that while spatial skills are related, they likely reflect different dimensions of cognitive development. Beyond age four-and-a-half, a majority of children studied appear capable of the above tasks, including navigation.

Studies done with children in a number of cultures have concluded that five-year-olds can verbally identify objects in an aerial photograph at scales of from 1:1000 to 1:3000 with minimal errors (although attempts to identify landscape objects at scales larger than 1:10,000 do result in some errors of scale). Further, such children can trace 'maps' from these photos, and can use these maps to solve simple navigation problems.

Slightly older children, in kindergarten through second grade, are capable of drawing quite complex maps. While these maps are plan views, overall, children typically draw the elements of such maps in elevation rather than in plan (e.g. house facades instead of house roofs), leading to the conclusion on the part of some psychologists that these children are unable to 'coordinate perspectives'. However, when seven-year-olds are asked to tell why they have drawn their maps in that way, they show full recognition that theirs is not a literal representation of an 'airplane view' of the world. On the contrary, they say that they are purposely trying to show what's most interesting about each object: a house facade is obviously more interesting than the view of the same house from above. That this is not simply a failure to coordinate perspectives is further indicated by the fact that many maps drawn by adults in the Middle Ages had the same characteristics.

2. Theoretical Controversies Concerning Spatial Aspects of Geographical Learning

In determining when children are ready for the introduction of geographic concepts in school, geographic education has been more concerned with the sequence of children's development than with their learning, in a psychological sense. The dominant theoretical paradigm in geographic education is that provided by the noted theorist and researcher in child development, Jean Piaget. In Piaget's view, all cognitive development, including spatial cognition, goes from lower to higher levels of conceptual and symbolic functioning, part of a more general set of competencies acquired in a stagewise, lock-step fashion (see Piaget and Inhelder 1956). Thus, such development is said to progress stepwise through four stages, which he termed sensorimotor, pre-operational, concrete operational, and formal operational, loosely associated with certain age ranges. Children at the sensorimotor and preoperational stages view the environment in relation to themselves and their own actions. Thus, according to this developmental view, children begin to be capable of true geographic learning only upon

reaching the concrete operational stage of development (Liben and Downs 1989). While Piaget stressed that his developmental sequence is to be measured in 'stages, not ages,' some influential educators, more comfortable with ages, have proposed that map concepts cannot be understood by children younger than seven.

An alternative theoretical perspective is termed 'universal mapping' (Stea et al. 1996). Universal mapping, while not directly challenging the general concept of developmental stages, suggests that the kinds of spatial skills that eventually result in the production and use of geographical maps transcend cultural differences and begin to appear very early in the life of the child. Universal mapping states that, throughout the development of the human species, people have had a need to construct mental representations of the environment at a geographical scale. Comprehension of maps requires perspective taking ability, in particular, moving from a three-dimensional real-world perspective to a two-dimensional, micro-level perspective. Thus, the theory proposes that young children have a 'natural' ability to represent the world in symbolic form at a reduced scale, rotated to the overhead perspective. Some proponents of universal mapping theory in its strongest form suggest that children may have an inborn mapping acquisition device analogous to the language acquisition device proposed by the linguistic theorist Noam Chomsky.

There is still some question as to whether and in what way mapping skills and spatial development are related. Piagetian theory has been interpreted by many educators to imply that children are incapable of engaging in the form of geographic learning called mapping at school-entering age. Universal mapping suggests, to the contrary, that some geographic abilities are so fundamental that their mastery by pre-school children can form a base for the teaching of other subjects in school.

Deciding between these two theoretical positions has been made more difficult by marked differences in methodologies used by their respective proponents. Studies have differed in dependence on verbal or nonverbal responses by children, interpretations of errors made in verbal responding, degree of generalization from results obtained in meso-scale (room-level) environments to macro-scale (geographical) environments, and widely divergent scales of material used in aerial photos studies, to name but a few. At present, it seems that while cognitive development certainly takes place, and probably in stages, this staged sequence is neither quite as stepwise nor as irreversible as originally proposed. It may be that transition between stages occurs earlier in the development of spatial abilities at the geographic scale than in other cognitive realms, indicating its importance to both human development in general and geographic learning in particular.

3. *Geographic Learning, Geographic Education, and Future Directions*

As children begin to master perspective taking and part-whole relations, they should become more adept at attending to elements that will aid in their learning map information. However, while learning about space at the geographical scale is one important aspect of geographical learning, learning about other characteristics of geographical environment, including its flora, fauna, landscapes, and geology, is also important. Thus, learning to map, while in many ways the keystone of geographic learning, is not all of it. With increased demand in US school systems to develop a systematic program of geography instruction, there is a need for standardized, cost-effective methods for delivering this information. Instruction via electronic media (e.g., videotapes, computer programs) can be both appealing and educational to young viewers. The success of such television shows as *Where in the World is Carmen San Diego?* provides one demonstration that programs can be produced which capture children's attention and at the same time stretch the boundaries of their geographic knowledge.

Geography instruction at the primary school level usually involves, at least in part, the comprehension of pictographic information, in particular, learning the spatial locations and boundaries of continents and geo-political regions such as countries and states. Because of its inherent visual and spatial content, geography lends itself well to instruction via electronic media. In fact, the types of visual and spatial content that are included in televised instruction can determine how readily that information is encoded and later retrieved from memory. For example, the addition of unique visual images can serve as a mnemonic aid for learning geographic shapes. Recent studies have demonstrated that visual icons, such as the Blue Ridge Mountains in the shape of Virginia, provided strong recall cues for state shapes (see Piñon 1999). Similar studies have shown the facilitative effects of icons and imagery in recalling map content (see Webb et al. 1994).

See also: Cartography; Cognitive Maps; Geographic Education; Knowledge (Explicit and Implicit); Philosophical Aspects; Navigation in Spatial Environments; Neural Basis of; Space; Linguistic Expression; Spatial Cognition

Bibliography

- Blades M, Spencer C 1990 The development of 3- to 6-year-olds' map using ability: the relative importance of landmarks and map alignment. *Journal of Genetic Psychology* **151**: 181-94
- Blaut J M 1999 Maps and spaces. *The Professional Geographer* **51**: 510-15
- Cohen R (ed.) 1985 *The Development of Spatial Cognition*. Erlbaum Associates, Hillsdale, NJ

- DeLoache J 1989 The development of representation in very young children. *Advanced in Child Development and Behavior* **22**: 1–39
- Liben L S, Downs R M 1989 Understanding maps as symbols: the development of maps concepts in children. *Advances in Child Development and Behavior* **22**: 145–201
- Matthews M H 1992 *Making Sense of Place*. Barnes & Noble, Lanham, MD
- Newcombe N, Huttenlocher J 1992 Children's early ability to solve perspective-taking problems. *Developmental Psychology* **28**: 635–43
- Piaget J, Inhelder B 1956 *The Child's Conception of Space* [trans. Langdon J F, Lunzer J L], Routledge, London
- Piñon M F 1999 Video imagery and children's recall of televised geographic information: Affecting more than visual content. *Journal of Instructional Psychology* **26**: 24–32
- Stea D, Blaut J M, Stephens J 1996 Mapping as a cultural universal. In: Portugali J (ed.) *The Construction of Cognitive Maps*. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 345–60
- Webb J M, Saltz E D, McCarthy M T, Kealy W A 1994 Conjoint influence of maps and aural prose on children's retrieval of instruction. *Journal of Experimental Education* **62**: 195–208

D. Stea

Geographic Thought, History of

1. General Conditions

1.1 Vernacular, Corporate, and Scientific Geographies

As a scientific discipline, geography has a long history. Its practice goes back to the time of Herodotus (484–425 BC) in the 5th century BC. The term itself was used in *De Mundo* attributed to Aristotle in the 4th century BC. Geography deals with orientation, the location of places, the surface of the Earth, the features which characterize it, and the way people conceive and experience places, the Earth, and the Cosmos. It means that geography existed as a practice of daily life before the development of scientific thought. People had to find places, evaluate the value of soils and the best way to plant them, detect and exploit mineral ores, etc. Many of these vernacular practices and knowledge coexisted simultaneously with the formal development of the discipline.

Since geography gives information on distant places, the ways to reach them, their economy, population, and culture, it is important for rulers. Governments rely either on empirical knowledge accumulated by local persons or on objective evidence gathered by their cartographic, statistical, and ethnographic services for the management of a territory. In the first case, the local agents of the sovereign know

the country they are in charge of so well that they can develop policies of their own. Hence, there is a tendency for all governments to develop centralized forms of control, for which they need special forms of geographic knowledge. Modern enterprises also have to develop geographic skills. They have to locate resources, exploit them, and distribute goods. A part of modern geography is linked with their activity.

The history of scientific geographic thought has always been situated in this context, with an eye to the type of practices used by ordinary people as well as to the skills developed by administrative services in gathering and interpreting spatial data. The discipline was institutionalized for several reasons:

(a) the will to provide the state (or other power institutions) with valuable knowledge for domestic control or overseas expansion;

(b) a curiosity for the low culture forms of geographic lore and their preservation; and

(c) the will to provide the citizens of modern nation states with an image of their country and the rest of the world in order to build or strengthen their national identity and explain how the contemporary world is working.

1.2 Geographic Narratives and the History of Geographic Thought

Geographic thought was expressed through standardized forms of narratives. The history of the discipline is partly that of the transformation of geographic discursive fields. What were the publics aimed at, the rhetoric used to attract or convince them, and the type of knowledge judged convenient for them? In Greek and Roman times for instance, according to one of the established geographic genres, geography was often presented as the account of a journey by Icarus.

From the Renaissance onwards, geography developed, as narratives on the newly discovered lands and the strangeness of their inhabitants, universal geographies, i.e., pictures of all the places and countries of the Earth, or quantitative assessments of the population, activities, wealth, and military assets of the provinces of kingdoms in order to facilitate their control. During the nineteenth and twentieth centuries, geographers developed other types of narratives: the description of exotic places for tourists and the analysis of national territories for the civic instruction of young people.

Geography is a complex discipline. Its changes are partly explained by the ways it was institutionalized and partly by the evolution of the discursive fields geographers developed in order to gain a large readership. However, its transformation mainly reflects the dynamics of its scientific purposes. It is convenient to distinguish five different scientific programs that were transformed into traditions later.

2. *The Five Approaches in Geography*

2.1 *Geography as a Science of Terrestrial Locations. Cartography*

The first program was linked with the problems of orientation. In order to build a reliable system of place locations, Greek geographers discovered that the best solution was to rely on an abstract grid of reference (Jacob 1991). The idea appeared with the first forms of mapmaking. For the Ionian mapmakers in the late sixth and early fifth centuries BC, all territories had geometric shapes. The idea of spherical coordinates, which appeared in the third century BC, was conducive to the first scientific cartography. At the same time, it provided a coherent grid of reference for locating places. As expressed by Ptolemy (AD 98–168), geography became ‘the sublime science which reads in the sky the image of the Earth.’

The Earth is a sphere and terrestrial positions, i.e., longitudes and latitudes, can be measured through the observation of the Sun and stars. Eratosthenes (273–192 BC) was the first to develop this line of research. Hipparchus (second century BC) proposed the idea of geographic coordinates. In the second century AD, Ptolemy synthesized the accumulated data in this field. His contribution was preserved and improved by Arabic geographers during the Middle Ages. The rediscovery of his texts in the fifteenth century marked a new start in the history of geographic thought. In the seventeenth century, Bernhardus Varenius (1622–1650) modernized the cosmographic approach through the introduction of the Copernician view of a Sun-centered Universe.

The major problem for this conception of geography came from the impossibility of keeping time, so that astronomical observations were unable to give a measure of longitude—a weakness which subsisted from Antiquity to the discovery of the marine chronometer by John Harrison (1693–1776) in 1735. Until the mid-eighteenth century, one of the main tasks of geographers was to compile data relative to distances in travel accounts. From the mid-eighteenth century, mapmaking became the task of specialized engineers and scientists. They provided geographers with the topographical maps they needed.

Geographers ceased to be responsible for the basic cartography, but the problems of location and mapmaking remained important for them. Their role changed; they learnt in the nineteenth century to use statistical sources to produce thematic maps. During the twentieth century, remote sensing facilitated the gathering of data on land use, vegetation, temperatures, atmosphere, etc. Thanks to computing facilities, the building of geographical information systems has recently become easy; one of the most prized skills of contemporary geographers is to provide administrations or firms with this new cartographic aid to decision-making.

Thanks to cartography, geographers were the first able to describe the Earth as seen from the sky—the Icarus’ viewpoint, according to Greek geographers. It was essential for getting an objective description of the face of the Earth and its regional divisions.

2.2 *Geography as Description. The Regional Approach*

The second agenda developed by geographers was the description of the Earth’s surface and its regional differentiation. During Antiquity, it was illustrated by Herodotus, Marinus of Tyre (first or second century AD) and Strabo (64 BC–AD 20). Arab geographers improved much of the knowledge of the Middle East, India, the Far East, and Africa south of the Sahara. With the great discoveries, the program became really worldwide.

The main problem was: What to describe and how to describe it? Thanks to the map and the vertical perspective it offered, geographic description was different from the travelers’; it was not structured along itineraries, but on the apprehension of regional units as seen from above. Such a perspective allowed for the contraction of much information, but it destroyed the link between direct observation and description, hence the dryness of many descriptions until the eighteenth century.

Because there was no standardized vocabulary to speak about the details of topography, rocks, plants, animals, and human activities, geographers were unable to grasp the real world in their narratives. Sometimes, they preferred to provide their readers with information on local mythology (in the Greek and Roman times), the history of Christianity (during the Middle Ages), and the local coats of arms and heraldry (in feudal Europe). In a way, the travel accounts of the Middle Ages and Renaissance are generally more interesting for today’s geographers than the geographical publications of the time.

Everything changed in the eighteenth century, with the development of natural history and a new curiosity for human activity. The work of Linné in botany and zoology, a few geologists for rocks, and the first agronomists for rural life, created the technical terms geographers needed in order to write precise descriptions of the world. Thanks to Jean-Jacques Rousseau (1712–1778) and Johann-Heinrich Pestalozzi (1746–1827), geographers began to offer descriptions on two scales: one was based on both fieldwork and direct observation, and the other on the synthetic vision of the map-reader. Carl Ritter (1779–1859), one of the pioneers of modern geography, has been trained in a Pestalozzian school, just as Elisée Reclus (1830–1905). Alexander von Humboldt (1779–1859) discovered the value of first-hand contact with landscapes through his training as a natural scientist and his contacts with

Georg Forster (1754–1794), who was on Cook's second voyage around the world.

The quality of geographic description kept increasing with the development of new scientific fields: pedology, meteorology, and climatology on the natural side of the discipline, and agronomy, demography, and social and cultural anthropology on the human one. Many of these research fields were initially explored by geographers and then turned independent.

Because of the improvement of geographical descriptions, geographers conceived their discipline in a different way: they began to consider that one of its major advantages was to deal with the real world. The idea that geography was a science of concreteness appeared. Geographers did not consider themselves only as scientists; description was an art, so that they had to display a literary bent.

During all of the nineteenth century and a good part of the twentieth, when writing descriptions, geographers felt to provide impartial views of the world since their work was based on objective observations. They considered that the descriptions and modes of regionalization developed by laymen were subjective and as such, not reliable. They preferred to ignore them.

2.3 Geography as Explanation: The Perspective of Natural Philosophies and Natural Sciences

As soon as description progressed in the eighteenth century, a third task appeared for geographers: they had to explain the distributions they observed. A precise description was necessarily conducive to questions about the causes responsible for the patterns read on the maps or the features discovered through fieldwork. Geographers had to give an explanatory dimension to their analyses.

During the first half of the nineteenth century, the main interrogation was about the interaction of natural features, their mutual conditioning, and their influence on the destiny of human groups. In the seventeenth century, Varenus had opened the way through his analysis of position. In order to explain terrestrial distributions, he stressed the role of latitude and of the major topographic accidents. He opposed inland and coastal locations. In the first half of the nineteenth century, this approach was transformed into a coherent paradigm: the analysis of position, well exemplified by Alexander von Humboldt and systematized by Carl Ritter. The major factors in the shaping of the Earth and the explanation of regional distributions were latitude, altitude, continental versus coastal locations, and the patterns of mountain ranges and coastlines. This kind of analysis had to be developed at different scales.

Geographers used these tools to explain the distribution of particular natural features: the altitudinal patterns of vegetation in tropical mountains for Alexander von Humboldt. Ritter wished to explain

how the destiny of peoples was shaped by the countries they inhabited. Behind these projects, the influence of early nineteenth century philosophies was obvious. Humboldt stressed the harmony between the different levels of the geographic reality according to the natural philosophies of his time. Ritter followed Herder (1744–1803) when exploring the influence of environment on cultures.

This first type of explanation, linked with the German philosophies of late eighteenth and early nineteenth centuries, was displaced by a natural science perspective when Charles Darwin (1809–1882) stressed the role of the environment in the transformation of life. Geographers of the late nineteenth century, Friedrich Ratzel (1844–1904) in Germany and Paul Vidal de la Blache (1845–1918) in France, built human geography on this evolutionist, either Darwinist or neo-Lamarckian basis. Their ambition was to explain human distributions through the analysis of the mainly 'vertical' (or local) relations people had with their milieu. In fact, they combined the analysis of position inherited from Ritter with an ecological approach. They stressed both the significance of local environments (the ecological approach) and of circulation (the analysis of position). The risk with the ecological approach was to accept a deterministic interpretation of human beings/nature relationships. Circulation offered a way to escape this danger.

Vidal de la Blache used the idea of *genre de vie* as the main concept of the new approach. In order to exploit an environment, people developed sets of techniques and practices that were soon transformed into a way of life because of the force of habits. The main interest of the *genre de vie* notion was to give a social dimension (the force of habits) to human geography and to stress the role of techniques in the mediation between human groups and the environment.

Even if Ratzel and Vidal de la Blache were conscious of the significance of circulation, their human geography tried mainly to explain the influence of physical geography on human distributions and the way people took advantage of natural endowments. The role of social relations and of distance as an obstacle to interaction was practically ignored. Human geography was at its best when it used the *genre de vie* paradigm to explain regional structures through the interaction between physical forces and social groups. Unfortunately, this approach could not work in urbanized and industrialized societies, where *genre de vie* became too complex to be a useful means of investigation.

Because human geography was unable to explain satisfactorily the distribution of social groups and artifacts on the surface of the Earth, some geographers, specially in Germany, preferred to define geography as the science of landscape, which avoided splitting the discipline between natural and social parts.

The analysis of position played a central role in the first efforts to investigate political geography. Alfred T. Mahan (1840–1914) and Halford Mackinder (1861–1947) stressed the opposition between inland locations and coastal ones. This theme became central to geopolitical studies in the first half of the twentieth century.

Thanks to the natural sciences, an explanatory dimension had been introduced in geography. Geographers succeeded in organizing efficiently physical systematic disciplines: geomorphology, climatology, biogeography, pedology, hydrology, oceanography, etc. On the human side of the discipline, the results were less impressive. Geographers capitalized more on the analysis of position than on the ecological paradigm. Even cultural geographers like Carl Sauer (1890–1975) did not try to explore systematically the role of techniques in the distribution of human groups. By the mid-twentieth century, a feeling of dissatisfaction had developed in human geography.

2.4 *Geography as Explanation: The Social Science Approach*

Many geographers tried to give a new start to their discipline. The most successful was Edward Ullman (1912–1976). His diagnosis was clear. Human geography failed because it was conceived as a natural science. Overemphasis had been given to the study of human beings/milieu relationships. Circulation, the other face of human geography as defined at the end of the nineteenth century, had been neglected. In order to build a modern geography, emphasis had to be given to social interaction (Ullman 1980).

Geographers discovered that economists had developed an interest for spatial economics since the beginning of the nineteenth century. They had explored the influence of distance upon the location of farming, industrial activities, and services. Geographers decided to take advantage of the results of spatial economics to give a more systematic dimension to human geography. It was the basis of the new geography of the late 1950s and 1960s.

In a way, this new geography translated Ritter's position into social terms. Instead of focusing on the opposition between inland and coastal locations, for instance, it focused on the dialectics between center and periphery. The center was defined in terms of social physics as the point or the area that enjoyed the highest accessibility to the totality of a social group. Procedures were invented to measure centrality and to map it. The idea of population or income potential changed the traditional vision. Instead of the center as a point, it presented the center as an area. In the United States, a central core, the manufacturing or industrial belt, appeared as the main feature of the territorial structure.

Geographers relied on spatial economics to develop their new understanding of human geography, but they went far ahead of economists in their analysis of spatial organization. The regional geography of the first half of the twentieth century fell short of a systematic formulation. Thanks to the works of Edward Ullman, Torsten Hägerstrand (born 1916) and a few other geographers, the way space was organized—through specialization, the development of networks, the dynamics of scale and external economies, and the opposition between center and periphery—became clear.

During the 1970s and 1980s, a few geographers, for instance, Paul Claval (born 1932), tried to extend the analysis of spatial interaction to non-economic aspects of social life, either social relations or political structures. Instead of focusing on transport costs, they analyzed communication, the structuring influence of commutation costs, the role of medias, and the significance of symbolism. In this way, they provided a new understanding of the dynamics of globalization, so evident from the 1960s.

The emphasis on social interaction had positive effects on geography in many respects, but it reduced the interest in the ecological approach at a time when ecology came really of age and pollution became a major problem at the local, regional, and global levels.

2.5 *Geography and Postmodernity: The Cultural Approach and the Experience of Nature, Place, and Society*

From the early 1970s, a strong critical movement developed in geography. Young colleagues struggled against the conception of their discipline as a natural science and the interpretation that was given to its social dimension. The positivist approach considered social relations as natural facts. It ignored their historical dimensions and their creation through human agency. As a result, neo-positivist geographers did not criticize the hierarchical structures and spatial organizations of the groups they observed and explained, even if they were conducive to exploitation and oppression.

From the eighteenth century, a different conception of geography had developed. In his *Géographie politique*, Turgot (1727–1781) wished to promote new geographies much more than to explain the existing one (Turgot 1808–1811). It was one of the main features of the geographic projects of the Enlightenment. The critical tradition was represented during the nineteenth century by anarchists like Elisée Reclus and Peter Kropotkin (1842–1921), who made it illustrious, but the prevailing option for a natural science approach restricted their influence.

At the beginning of the 1970s, David Harvey (born 1935) played a decisive role in the development of a new interest for social justice (Harvey 1973). However,

the break with the neo-positivist positions of the 1950s and 1960s had other correlates. If social realities are the product of human agency, it is urgent to study the ways people experience places and conceive nature and society, the dreams they have about their future, and the significance they give to their lives. Geographers can no more ignore the ways people organize their existence. The rigid division between vernacular or corporate geographies on the one side, and scientific ones on the other, fades out. Scientific geographers have to explore the geographies of lay people in order to understand the landscape they shape and the institutions they create. Ethnogeography becomes a new field of investigation.

There was an old but limited interest in people's conceptions of place thanks to the analysis of choronyms. The development of new attitudes among geographers had not been possible without the influence of Heidegger (1889–1979). Eric Dardel (1902–1968), in France, was well aware of the significance of his philosophy, as Yi-Fu Tuan (born 1930) or Edward Relph (born 1944) a generation later.

When societies are conceived as a product of human activity, i.e., in an historical perspective, it becomes important to understand the practices, attitudes, knowledge, and beliefs shared by social groups, the techniques they master in the fields of material activity and social life, and the significance they give to their lives. The cultural approach, which had been mainly developed by Carl Sauer in the first half of the twentieth century and had experienced a crisis in the 1960s and early 1970s, benefited from a new popularity thanks to Denis Cosgrove (born 1948), James S. Duncan (born 1945), Joël Bonnemaison (1940–1997), and Augustin Berque (born 1942). It was one of the main frontiers in human geography from 1980 onwards. Culture ceased to be considered as a collective heritage. It was analyzed as a continuous process of communication, adaptation, evaluation, and refurbishing.

The criticism of the neo-positivist conceptions of social sciences was by no means restricted to geography. The transformations of the discipline were parallel with new orientations in sociology, anthropology, and history. Geographers discovered the interest of the Clifford Geertz's type of thick description as a tool to explore ethnogeographies and enter the complexities of cultures.

Under the aegis of French deconstructionist philosophers, Jacques Derrida (born 1930) and Michel Foucault (1926–1984) for instance, social knowledge was increasingly analyzed as a tool to legitimize the Establishment and gain support for governmental policies, whatever their content. Geographers began to scrutinize geographical narratives and look for their hidden significance. They got increasingly aware of the role of geography as an instrument to build efficient forms of social control. Thanks to Derek Gregory (born 1951), the study of geographical imaginations

and of their role in social life and power situations became popular (Gregory 1994).

The new epistemologies are increasingly critical of the great narratives of the past. Geographers are reluctant to rely on macro-theories or philosophies—historical materialism for instance. They develop a growing interest in the local and privileged micro- or meso-theories. They explore realms that had been previously neglected: gender geography, the geography of young and aged people, the varied forms of segregation, and exclusion or sexual differences that most societies build.

Men and women do not live in a standardized, sterilized, and acclimatized space. They inhabit specific places, develop territorial identities, and use landscape markers as symbols of what unite them and provide them with a sense of shared community. Consequently, geographers are increasingly concerned with the religions, philosophies, or ideologies people create in order to give significance to their lives. They explore the moves the human mind performs in order to discover perspectives of society, nature, and collective destiny, and build values. Transcendence and immanence are wrongly considered as purely philosophical concepts. They have a spatial dimension; geographers have to explore them.

Men and women are part of nature; they live in nature, with nature (Harvey 1996). The ecological perspective has been completely recovered. Geographers are increasingly involved in the study of the energy and material relations which develop between human beings and their environment, and in the analysis of the attitudes towards nature, the significance given to natural landscape, and the management of the planet Earth.

Contemporary geography goes much deeper into the analysis of social and psychological processes than was the case in the 1970s. At the same time, it has become a more critical and modest discipline. For many of the young radicals, scholars have to refrain from building mega-narratives, as in the past. It is better to concentrate on limited problems.

3. The Coexistence of Perspectives Instead of the Displacement of Paradigms

Kuhn's interpretation of scientific revolutions was very popular among geographers in the 1970s. They interpreted the new geography of the 1960s as a new paradigm. In fact, the history of geographic thought does not fit well into the Kuhnian model. We have distinguished five traditions, but when a new one appeared, it did not displace the older ones. The interest in location and map making is less central today than in Antiquity or modern times, but it is still fundamental to the discipline. Description as an intellectual venture worth pursuing for its own sake has long ceased to be fashionable, but progress in the

analytical instruments offered by other natural or social sciences allows for new readings of landscapes and maps, and the discovery of new patterns. It makes geographers aware of new problems.

Geography as a natural science and geography as a positivist social sciences are still alive. They provided geographers with many of the tools they use, the analysis of position both relative to physical factors along Ritter's views, or social distributions according to Ullman's conceptions, and the ecological approach to human beings/milieu relationships. Many colleagues try to develop more performing models of economic or social distributions. The emphasis on communication is a new development of an old curiosity.

When building a periodization of the history of geographic thought, the development, fading out, and renewal of geographical traditions constitute only one component of the whole picture.

Greek geography was born out a Kuhnian revolution, the invention of the grid of meridians and parallels, but its successes came partly from its role as a repository of Greek culture. With the development of empires, first in Alexandria and then in Rome, it became a useful tool for the new bureaucracies because of the maps it produced and the descriptions it provided.

The decline of the Classical tradition in the Middle Ages reflected the disappearance of imperial polities in Western Europe and their displacement by feudal structures, which did not involve explicit corporate geographies. Maps were used as symbols of the monarch's power, not as a tool for his or her action. Because of the development of trade from the eleventh century, new cartographic and geographic instruments were needed—hence the invention of Portolano charts. In the Arab world where imperial polities played an important role, the fate of medieval geography was far better.

The new phase of development of geography which started from the Renaissance was linked with the rediscovery of Ptolemy and the cartographic perspective, the voyages of discovery, the rise of the modern state with its proto-bureaucracies, and new forms of corporate economic activities. For two centuries, geography remained based on the two perspectives of cartography and regional description.

The eighteenth century was as a major turning point. The cartographic limitations disappeared thanks to an easier measure of longitudes. Because of the progress of natural sciences and the birth of social sciences, geographic description improved much and ceased to be problematical. This kind of description necessarily led to an interest in explanation. The transition from pre-modern to modern forms of geography was a difficult one and lasted for about a century, since it meant a change in the institutionalization of the discipline: from state cartographic services to universities.

The naturalist orientation that was born at the beginning of the nineteenth century, prevailed until the mid-twentieth. From that time, geography became first a neo-positivist social science, and then a discipline, partly scientific, partly humanistic, or philosophical, based on human experience and the significance of the human presence on Earth. Institutionalization began once again to change. Geography is increasingly dependant on research institutions and private funding.

Geography offers the intellectual tools that are needed to understand the contemporary acceleration of globalization. It explains why technical uniformity is conducive to new forms of regional organization and does not entail the destruction of all spatial differentiation.

What will be the future of the discipline? As in the past, a part of the specialty fields it nurtured will develop as fully-fledged sciences of their own. The technicalities of contemporary physical geographies will deepen the gap between them and human geography. The future of geography as a unitary discipline is perhaps more on the reflective side than on the analytical side of human knowledge.

Bibliography

- Berdoulay V 1981 *La Formation de l'Ecole française de géographie*. Bibliothèque nationale, Paris
- Capel H 1981 *Filosofía y ciencia en la Geografía contemporánea. Una introducción a la Geografía*. Barcanova, Barcelona, Spain
- Claval P 1998 *Histoire de la géographie française de 1870 à nos jours*. Nathan, Paris
- Dunbar G 1991 *Modern Geography. An Encyclopedic Survey*. Garland, New York
- Godlewska A 1999 *Geography Unbound. French Geographic Science from Cassini to Humboldt*. The University of Chicago Press, Chicago
- Gould P 1985 *The Geographer at Work*. Routledge & Kegan Paul, London
- Gregory D 1978 *Ideology, Science and Human Geography*. Hutchinson, London
- Gregory D 1994 *Geographical Imaginations*. Blackwell, Oxford, UK
- Hartshorne R 1939 *The Nature of Geography*. The Association of American Geographers, Lancaster, PA
- Harvey D 1973 *Social Justice and the City*. Arnold, London
- Harvey D 1996 *Justice, Nature and the Geography of Difference*. Blackwell, Oxford, UK
- Holt-Jensen A 1988 *Geography. History and Concepts*. Paul Chapman, London
- Jacob C 1991 *Géographie et ethnographie en Grèce ancienne*. A. Colin, Paris
- James P 1972 *All Possible Worlds. A History of Geographic Ideas*. Bobbs-Merrill, Indianapolis, IN
- Livingstone D N 1992 *The Geographical Traditions*. Blackwell, Oxford, UK
- Schultz H D 1980 *Die deutschsprachige Geographie von 1800 bis 1970*. Selbstverlag des Geographischen Instituts der Freien Universität Berlin, Berlin

- Stoddart D R 1986 *On Geography*. Blackwell, Oxford, UK
- Turgot A R 1808–1811 Sur la géographie politique. In: Dupont de Nemoun P S (ed.) *Plans et projets d'ouvrages, Oeuvres complètes*, Paris
- Ullman E L 1980 *Geography as Spatial Interaction*. University of Washington Press, Seattle, WA

P. Claval

Geography

The practice of geography—literally ‘earth writing,’ from the Greek *geo* and *graphein*—goes back at least 2000 years, but it has a much shorter history as an academic discipline, with little formal presence in universities before the twentieth century. There have been many attempts to define it: one widely accepted in the mid-twentieth century saw it ‘concerned to provide accurate, orderly, and rational description and interpretation of the variable character of the Earth surface’ (Hartshorne 1939, p. 21). Its focus remains areal differentiation, but it has moved beyond cataloging and description to *analyzing and providing explanatory accounts for the nature and implications of the material character of the earth’s surface and its use by people and societies*. Its key concepts are space, place, and environment, each studied at a variety of scales.

Although most definitions and university departments of geography embrace both physical and human worlds, studies of the two are now substantially separate components of the discipline. This article focuses on human geography, that part normally affiliated to the social and behavioral sciences.

1. Intellectual Origins

As a source of information about the earth, geography has been important to all societies. It was part of the curriculum at Cambridge and Oxford Universities by the late sixteenth century, for example, taught for its practical value with two main themes. The provision of locational information involved cartography, on which a wealth of early material in many cultures has been identified. Alongside was chorology, the description of physical environments and their inhabitants. Geographers portrayed the different parts of the world, meeting the needs of traders, diplomats, and the military and also whetting the curiosity of the ‘educated classes.’

Although geographical material was widely disseminated and used, there was no formal discipline with institutional status in universities. This was countered in the nineteenth century by societies founded to promote the dissemination of geographical knowledge (Livingstone 1992): many were closely linked to imperialist projects and actively promoted exploration of extra-European *terrae incognitae*. The

first university chairs and departments were established by royal decree in Germany in the 1870s; they were to prepare school teachers, as part of the creation of an educated citizenry, and to sustain colonial and military enterprises. Similar successful campaigns elsewhere included the Netherlands, where the first established chair was in ‘colonial geography,’ and in England, where the Royal Geographical Society provided financial assistance to establish departments at Oxford and Cambridge, although the first honours degrees in geography were awarded at Liverpool, in 1917. The departments were small but provided the foundation for substantial expansion after World War II, when there was a geography department in nearly every British university and a graduate school in 27 United States’ universities.

2. The Creation of a Discipline

The subject matter taught in these new departments focused on the inter-relations among peoples and environments, as conceived by several important German and French geographers (such as von Humboldt, Ritter, Ratzel, and de la Blache: Livingstone 1992). Geographers sought a unique intellectual rationale within the academic division of labor, increasingly so as universities adopted the German tradition of marrying teaching with research. Of the six trends that Freeman (1961) identified up to c.1950 (encyclopedic, educational, colonial, generalizing, political, and specialization), the generalizing trend became increasingly important, with individual specialization as a necessary consequence.

The emerging practice of geography was strongly influenced by its milieu (Livingstone 1992). Debates in the first half of the twentieth century crystallized around the concept of the region. Hartshorne (1939, p. 462) presented geography as:

... a science that interprets the realities of areal differentiation of the world as they are found, not only in terms of the differences in certain things from place to place, but also in terms of the total combination of phenomena in each place, different from those at every other place.

Regions possess unique characteristics based on their inter-related physical and human environments, and regional geography aimed (Wooldridge 1956, p. 53):

... to gather up the disparate strands of the systematic studies, the geographical aspects of other disciplines, into a coherent and focused unity, to see nature and nurture, physique and personality as closely related and interdependent elements in specific regions.

This identifies two long-recognized strands in geographical practice: the systematic, focusing on particular subject matter, and the regional, concentrating on specific areas.

There were common elements to geographical practice in those countries where it was established early:

Hartshorne's 1939 essay exemplified the strong German influence on American geography, for example, and much regional geography was based on classic French work. The links between the different language realms weakened after World War II, however (Johnston and Claval 1984): the remainder of this article focuses on the English-speaking world.

2.1 A 'New' Geography

Systematic specialization increasingly dominated the expanding discipline in the mid-twentieth century, with two significant intradisciplinary divides: a split between physical and human geography; and the creation of separate specialisms within each. Regional geography declined in importance as a research field, but remained important in the discipline's teaching. The early systematic strengths were in historical geography, particularly strong in the UK (led by Darby and his major work on the Domesday Book: Darby 1977), and cultural geography in the USA, led by Sauer (1952) and with strong links to anthropology. Economic, political, and social geography were small enterprises until the second half of the twentieth century.

Two strongly linked developments in the 1950s/1960s—termed 'quantitative and theoretical revolutions'—generated an explosion of interest in both physical and human geography. Quantitative methods were adopted to handle spatial data, but the 'revolutionary' component within human geography involved explicit consideration of theory, notably location theories whose key element was a geographical variable—space, usually represented as the cost involved in moving people, goods, and information. Classic works on the location of agricultural land uses industry and retail establishments (notably Christaller's central place theory) were appropriated and formed the basis for a 'new geography,' a spatial science whose goal was identifying and accounting for spatial order. This was largely North American in provenance, initially focused on a few departments (Johnston 1997), but the approach soon infiltrated other parts of the discipline, assisted by two seminal British anthologies (Chorley and Haggett 1965, 1967). Further spatial theories were added—on the diffusion of innovations and the widely-applied gravity model for studying flow patterns—and the entire corpus was synthesized in influential text books, notably Haggett (1965) and Abler et al. (1971).

The expansion was accompanied by increased specialization. Economic geography spawned subdivisions of industrial, agricultural, transport, and development geography, for example; urban geography grew a substantial subsector in urban social geography and was paralleled by interest in rural geography; social geography begat population and medical geography; and political geography was both

revived and spawned electoral geography. The shared goal was the discovery of spatial order—in both individual behavior patterns and their outcomes in the built environment.

Although diverging systematic interests reduced links between physical and human geographers, they explored common ground in techniques of spatial analysis and approaches to knowledge. The switch from an ideographic focus on unique characteristics of places to a nomothetic search for spatial order largely occurred in a philosophical vacuum, but some geographers promoted the positivist approach to knowledge production (Bunge 1966, Harvey 1969).

These shifts occurred when the discipline was expanding rapidly within universities. Their promoters achieved major intradisciplinary changes by portraying spatial science as meeting contemporary needs among the burgeoning, increasingly quantitative and theoretical, social sciences. They presented the regional approach as outdated: the concepts of areal differentiation and region were retained—though employed in different ways.

2.2 '-Isms,' Schisms, and 'Turns'

Although some promoted this 'new geography' as surpassing its predecessor, there was no complete revolution to a new hegemonic paradigm in the Kuhnian sense. Not all geographers were won over, and some continued to teach and research in their established ways—sustaining the traditions of historical, cultural, and regional geography which were foundations for later responses to the 'shock of the new.' Those responses criticized spatial science on two main grounds. The first attacked location theories for narrow economism, for privileging profit-maximizing (translated geographically as transport-cost-minimizing) as a determinant of decision-making and so denigrating the wide range of cultural and other influences on behavior. Spatial science, it was claimed, removed free will by assuming that humans react to stimuli in predetermined ways while its elevation of space as the predominant influence on locational decisions led to a derogation of place. Some critics explored alternatives to positivism—such as phenomenology, idealism and existentialism—in work categorized as humanistic geography.

The second set of criticisms concerned spatial science's approach to explanation. Location theories, it was argued, could not account for the major contours of the geography of development and underdevelopment at various spatial scales, of inequality and injustice, wealth, power, and discrimination. Explanation required understanding how economic systems, especially capitalism, work and involved more than adding the costs of transport and communications to neoclassical economic models: such explanation could be achieved through Karl Marx's writings. Harvey's (1973) collection of essays initiated

a series incorporating space to Marxian thought (notably Harvey 1982) and stimulated others during a period of social tensions over civil rights and the Vietnam War. Their work became known as radical geography.

Although both critiques attracted adherents, many were as dissatisfied with each as they were with spatial science. As Gregory (1978) demonstrated, whereas some spatial scientists strayed very close to naive economic determinism, much humanistic geography approached another extreme of voluntarism, tearing the individual out of context, while some radical geography offered little more than an alternative economic determinism. Much effort was spent in the late twentieth century integrating aspects of humanistic and radical geography.

3. *The Contemporary Discipline*

Sheppard (1995) divided human geographers into two main groups—spatial scientists and social theorists; strands of contemporary geographical scholarship that operate somewhat independently (Johnston 1997). Some recent textbooks in the ‘social theory mold’ dismiss spatial science as part of a discredited past (as in Peet 1998), whereas a volume promoting geography within the US scientific community concentrated on it (plus physical geography; NRC 1997). Attempts at comprehensive overviews include anthologies which lack integration (e.g. Gaile and Willmott 1989, 2001) while the editors of a volume on the ‘pervasive themes’ and ‘common elements within a discipline whose practitioners are in danger of forgetting their shared heritage and ideals’ concluded that not only had they failed to bridge the human–physical divide but they also feared a similar divide was opening within human geography (Abler et al. 1992).

3.1 *Spatial Science*

Spatial science remains a substantial component of contemporary human geography. It is strongly quantitative, but the formal (geometrical) location theories based on a single causal variable (space) have largely been abandoned: the search for spatial order neither anticipates the discovery of regular structures nor seeks universal laws of spatial behavior. Sayer (1984) drew an important distinction between extensive and intensive research: the former seeks empirical regularities whereas the latter explores the causal chains responsible for particular outcomes. Much spatial science is extensive research, a necessary precursor to many detailed investigations; by not eschewing empirical generalisations, it identifies significant features and trends in the mass of numerical data which characterize modern societies.

In the 1950s–1970s geographers assumed that standard statistical procedures could be applied unprob-

lematically to analyses of point, line, flow, and area patterns. This was challenged by work on spatial autocorrelation, which identified a range of problems and proposed new methods of spatial data analysis (e.g. Haining 1990). Other issues identified included the modifiable areal unit problem. Some geographical analyses study the characteristics of populations aggregated by areas (such as census administrative units), but there is an extremely large number of ways in which such places can be defined, involving both scale (how large are the areas?) and aggregation (how many different ways can areas of the same size be created?) effects. Openshaw (1985) showed that different aggregations can produce divergent statistical results, creating problems in deciding which to employ. There are also related geographical examples of ecological fallacies—assumptions that results for a particular dataset can be generalized to others at different scales and/or aggregations (including individuals). Various procedures for attacking these problems and providing robust solutions are exemplified by essays in Longley and Batty (1996), many significantly assisted by developments in computing power (including applications of artificial intelligence) and stimulating the conception of geocomputation to describe such work in both physical and human geography (Longley et al. 1998).

The most significant technological developments have been in geographical information systems (GIS: Longley et al. 1999), combined hardware and software for the organization, integration, analysis, and display of spatially-referenced data, with the powerful display media underpinning the growth of visualization strategies. These systems have revolutionized spatial analysis and led to the identification of a geographical information science: traditional studies can be undertaken much more readily and quickly; exploratory studies are increasingly feasible, and large-scale modeling strategies integrating datasets collected on different spatial templates made possible.

Much contemporary spatial science, including GIS, is applied in a wide range of public and private sector contexts—as in spatially-targeted niche marketing strategies based on small-area classifications (geodemographics). This has partly been in response to changes in the economic context for universities: the pressure to increase nonstate income has stimulated ‘applied research’ and has seen the development of such skills as a major selling point in the attraction of students to read for geography degrees (see NRC 1997).

3.2 *Social Theory*

Although spatial science remains a substantial component of the discipline, to many human geographers approaches involving social theory form the disciplinary mainstream (as in Massey et al. 1999). Humanistic and radical geography have metamorphosed

into new forms, although some maintain allegiance to the latter—notably Harvey (1989) and his continued reworking of Marxian approaches for understanding the contemporary world.

Work within this umbrella category focuses on understanding difference. Place now takes precedence over space among the main geographical concepts (on the tension between the two, see Taylor 1999): places provide contexts within which differences are produced and reproduced. But places are not fixed containers; like spaces, they are produced and continually reproduced by human action so that whereas in spatial science space is an exogenous independent variable (though modifiable, as in transport network developments), in social theory it is both independent and dependent variable (as stressed in Martin's 1999 critique of the 'new geographical economics' which relates more to the spatial science of the 1970s than to either strand of contemporary human geography). The interactions between people and places/spaces continually transform both, introducing a geographical component to the structure-agency dialectic for which Giddens (1977) sought a resolution in his structuration theory, very influential in human geography in the 1980s. People make their own histories, but within constraining/enabling spatial contexts—locales, or settings for interaction. (Giddens 1984 was one of the few social theorists who not only used spatial metaphors in his writings but was also influenced by geographical work, notably time-geography—e.g. Hägerstrand 1982—which draws attention to the geographical constraints on how individuals pursue personal and collective projects.)

A very important stimulus to recognition of the role of context/locale in the production and reproduction of difference came from feminist geography. Initially launched to highlight the discrimination against women within a male-dominated discipline and the consequential many silences in what geographers studied, this expanded into a much wider critique of 'single-factor' approaches (such as Marxist's concentration on class and the economic infrastructure) that ignored important other societal cleavages (Women and Geography Study Group 1984, Rose 1993). People viewing society from different positions create separate identities of self and other, which are potential foundations for identity politics. The class struggle is not the only one within capitalist society, and others (based on sexual orientation, race and religion, for example, as well as gender) cannot be folded within it.

Other stimuli came from postmodernism, which also emphasized difference and rejected the search for grand theories. Dear (1988) and introduced it to geographers as a way of addressing the contemporary condition, exemplified in the argued transition from a Fordist regime of accumulation, characterized by mass production of uniform commodities for mass consumption, to post-Fordism, characterized by small

niche markets and consumption for status as well as use (sometimes called flexible accumulation). Harvey (1989) countered by arguing that this shift, involving the production of new spaces and places, is but the latest response to a crisis of capitalism.

More important has been the impact of the various strands of poststructuralist thought derived from the humanities and social sciences. Poststructuralists argue that human subjectivity is not given but rather produced in contexts. So much of that production process is spatially and temporally contingent that truth and knowledge are contingent too, varying through time and over space in a range of situated and local knowledges that form the resource foundations for structuration dialectics and identity politics. Representation of these contingencies is crucial to social conduct, and poststructuralists stress the importance of language not only as the communication medium involved in creating human subjectivity but also as a social construction itself, with variations that are fundamental to the creation and understanding of difference (i.e. in the relationship between the signified and the signifier and in its interpersonal transmission—the hermeneutic process). Language is just one means of communicating understandings, however, and as well as interpreting written texts geographers have focused on other images (works of art, films, landscapes) implicated in the creation of situated knowledges—as in Harley's (1989) seminal work on deconstructing maps.

Said's (1978) book on the Western creation of a 'mythical East' was an important stimulus to geographers' studies of the roles of images of places and spaces in the creation and transmission of meanings and identities. They have embraced the postcolonial movement which challenges not only the political and economic components of colonialism and imperialism but also their roles in creating imaginative geographies at variance from those of the peoples who experienced that external power and are creating alternative geographies, knowledges, and identities through their own writings.

Many of these arguments were introduced through cultural geography but not confined to it and have been instrumental in reducing some of geography's fragmentation. Several parts of the discipline have experienced a 'cultural turn': the perceived interdependence of cultural and economic processes, for example, has stimulated broader studies of production, exchange, and consumption (Lee and Wills 1998); critical geopolitics emphasizes the ways images of the world are created, transmitted, and underpin geographies of power and conflict (O'Tuathail 1996); and writing on nonrepresentational theory focuses on the transitory, the 'billions of happy or unhappy encounters' (performances) which characterize the daily interactions among people, places, environments, and 'things' but are never recorded/represented in the traditional source materials for geographical schol-

arship (Thrift 1996). At an extreme of spatial scale heretofore ignored, geographers are also exploring the body as a site of identity-formation, as the most intimate of 'personal spaces' or 'the geography closest in' (Pile 1996).

Research within this area is both catholic and eclectic in its substantive interests. Space and place are being theorized in a wide range of contexts and scales. There are increasing links with other social science and humanities disciplines experiencing a 'spatial turn,' both in their use of spatial metaphors and in their appreciation of the roles of space and place in the structuration of people and societies.

Such work is very far from most spatial science and there is little methodological or philosophical common ground: one area of interaction, however, has emerged from critiques of the representations of the world created within GIS (Pickles 1995) and, more broadly, in cyberspace (Kitchin 1998). Social theory also lacks an 'applied' element as generally appreciated by those promoting geography as a 'relevant' or 'saleable' discipline (hence the absence of much on social theory in NRC 1997). Critical human geography is increasingly used to categorize work committed to emancipation, however, clarifying ways in which space and place are manipulated in the creation and maintenance of power relationships. This has a strong moral/ethical component, concerned not only with how geographical research and teaching are conducted but also with evaluating what is right and wrong, good and bad (Smith 1994, Sack 1997).

One area of traditional geographical interest revived by social theorists is the study of people-environment interactions, on which there has been a continuing stream of studies on various aspects of environmentalism, human impacts on the environment (Turner et al. 1990), environmental policy, and political economy approaches to environmental problems and issues. Poststructuralists argue that nature is a social construction—work on nature-culture interrelationships must take account of the cultural constructions of nature even the binary nature: culture categorization sets up an unreal opposition since each is part of the other. Writing on society-nature interrelationships has thus been substantially broadened, expanding from the technical issues involved in the production of 'environmental problems' and the political ones in their 'solution/resolution' to raise fundamental questions regarding the role of humans in nature. Some argue that studies of the environment can integrate physical and human geography (Simmons 1997) and present a major opportunity for geographers to undertake relevant research.

4. Conclusions

Human geography is a large, expanding, and vibrant discipline, characterized like most others by fragments and tensions: its extent and the nature of the expansion

can be appreciated by comparing the four editions of *The Dictionary of Human Geography* (Johnston et al. 2000). It has an established niche without the academic division of labor; its core concerns remain those crystallized over a century ago; and its basic concepts remain place, space, and environment. Within those general characterizations, however, the discipline has changed markedly in recent decades, not only in what it studies but also in how and why. Buttimer (1993) has suggested that throughout their discipline's history geographers have regularly shifted among four foundational metaphors (mosaic, machine, organism, and arena)—without the academic revolutions that are central to Kuhn's (1970) paradigm model of scientific progress. Contemporary spatial science employs the mosaic metaphor with its emphasis on spatial order (without the earlier determinist overtones when the machine metaphor was employed to 'explain' those mosaics), whereas social theory concentrates on organism and, especially, arena—the spatial contexts, or places, which are involved in the production of human-ness while being produced and reproduced by humans.

See also: Cultural Geography; Environment and Development; Geographic Thought, History of; Historical Geography; Place in Geography; Regional Geography; Rural Geography; Space and Social Theory in Geography; Spatial Analysis in Geography; Urban Geography

Bibliography

- Abler R, Adams J S, Gould P 1971 *Spatial Organization: the Geographer's View of the World*. Prentice-Hall, Englewood Cliffs, NJ
- Abler R F, Marcus M G, Olson J M (eds.) 1992 *Geography's Inner Worlds: Pervasive Themes in Contemporary American Geography*. Rutgers University Press, New Brunswick, NJ
- Bunge W 1966 *Theoretical Geography*, 2nd edn. Gleerup, Lund
- Buttimer A 1993 *Geography and the Human Spirit*. Johns Hopkins University Press, Baltimore, MD
- Chorley R J, Haggett P (eds.) 1965 *Frontiers in Geographical Teaching*. Methuen, London
- Chorley R J, Haggett P (eds.) 1967 *Models in Geography*. Methuen, London
- Darby H C 1977 *Domesday England*. Cambridge University Press, Cambridge, UK
- Dear M J 1988 The postmodern challenge: reconstructing human geography. *Transactions, Institute of British Geographers* **13**: 262–74
- Freeman T W 1961 *A Hundred Years of Geography*. Duckworth, London
- Gaile G L, Willmott C J (eds.) 1989 *Geography in America*. Merrill Publishing, Columbus, OH
- Gaile G L, Willmott C J (eds.) 2001 *American Geography at the Millennium*. Oxford University Press, New York
- Giddens A 1977 *Studies in Social and Political Theory*. Hutchinson, London
- Giddens A 1984 *The Constitution of Society*. Polity Press, Cambridge, UK

- Gregory D 1978 *Ideology, Science and Human Geography*. Hutchinson, London
- Hägerstrand T 1982 Diorama, path and project. *Tijdschrift voor Economische en Sociale Geografie* **73**: 323–39
- Haggett P 1965 *Locational Analysis in Human Geography*. Edward Arnold, London (2nd edn., with Cliff A D, Frey A E 1977)
- Haining R Q 1990 *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press, Cambridge, UK
- Harley J B 1989 Deconstructing the map. *Cartographica* **26**: 1–20
- Hartshorne R 1939 *The Nature of Geography: A Critical Survey of Current Thought in the Light of the Past*. Association of American Geographers, Lancaster, PA
- Harvey D 1969 *Explanation in Geography*. Edward Arnold, London
- Harvey D 1973 *Social Justice and the City*. Edward Arnold, London
- Harvey D 1982 *The Limits to Capital*. Blackwell, Oxford, UK
- Harvey D 1989 *The Condition of Postmodernity*. Blackwell, Oxford, UK
- Harvey D 1996 *Justice, Nature and the Geography of Difference*. Blackwell Publishers, Oxford, UK
- Johnston R J 1997 *Geography and Geographers: Anglo-American Human Geography Since 1945*, 5th edn. Arnold, London
- Johnston R J, Claval P (eds.) 1984 *Geography Since the Second World War: an International Survey*. Croom Helm, London
- Johnston R J, Gregory D, Pratt G, Watts M J (eds.) 2000 *The Dictionary of Human Geography*, 4th edn. Blackwell Publishers, Oxford, UK
- Kitchin R 1998 *Cyberspace: The World in the Wires*. Wiley, Chichester, UK
- Kuhn T S 1970 *The Structure of Scientific Revolutions*, 2nd edn. University of Chicago Press, Chicago, IL
- Lee R, Wills J (eds.) 1998 *Geographies of Economies*. Arnold, London
- Livingstone D N 1992 *The Geographical Tradition: Episodes in the History of a Contested Enterprise*. Blackwell Publishers, Oxford, UK
- Longley P, Batty M (eds.) 1996 *Spatial Analysis: Modelling in a GIS Environment*. GeoInformation International, Cambridge, UK
- Longley P A, Brooks S M, McDonnell R, Macmillan B (eds.) 1998 *Geocomputation: A Primer*. Wiley, Chichester, UK
- Longley P A, Goodchild M F, Maguire D J, Rhind D W (eds.) 1999 *Geographical Information Systems: Principles, Techniques, Applications and Management*, 2nd edn. John Wiley, New York
- Martin R 1999 The new 'geographical turn' in economics: some critical reflections. *Cambridge Journal of Economics* **23**: 65–91
- Massey D, Allen J, Sarre P (eds.) 1999 *Human Geography Today*. Polity Press, Cambridge, UK
- NRC 1997 *Rethinking Geography: New Relevance for Science and Society*. National Research Council, Washington, DC
- Openshaw S 1983 *The Modifiable Areal Unit Problem*. CATMOG 38, GeoBooks, Norwich, UK
- O'Tuathail G 1996 *Critical Geopolitics: The Politics of Writing Global Space*. Routledge, London
- Peet R 1998 *Modern Geographic Thought*. Blackwell Publishers, Oxford, UK
- Pickles J (ed.) 1995 *Ground Truth: The Social Implications of Geographical Information Systems*. Guilford Press, New York
- Pile S 1996 *The Body and the City*. Routledge, London
- Rose G 1993 *Feminism and Geography*. Polity Press, Cambridge, UK
- Sack R D 1997 *Homo Geographicus: A Framework for Awareness, Action and Moral Concern*. Johns Hopkins University Press, Baltimore, MD
- Said E W 1978 *Orientalism*. 1st edn. Pantheon Books, New York
- Sauer C O 1952 *Agricultural Origins and Dispersals*. American Geographical Society, New York
- Sayer R A 1984 *Method in Social Science: A Realist Approach*. Hutchinson, London (2nd edn. 1992 published by Routledge, London)
- Scott A J 1988 *New Industrial Spaces*. Pion, London
- Sheppard E S 1995 Dissenting from spatial analysis. *Urban Geography* **16**: 283–303
- Simmons I G 1997 *Humanity and Environment: A Cultural Ecology*. Longman, London
- Smith D M 1994 *Geography and Social Justice*. Blackwell Publishers, Oxford, UK
- Stoddart D R 1986 *On Geography, and its History*. Blackwell, Oxford, UK
- Taylor P J 1999 Places, spaces and Macy's: place-space tensions in the political geography of modernities. *Progress in Human Geography* **23**: 7–26
- Thrift N 1996 *Spatial Formations*. Sage, London
- Turner B L, Clark W C, Kates R W, Richards J F, Matthews J T, Meyer W B (eds.) 1990 *The Earth as Transformed by Human Action: Global and Regional Changes in the Biosphere Over the Last 300 Years*. Cambridge University Press, Cambridge, UK
- Women and Geography Study Group of the IBG 1984 *Geography and Gender: An Introduction to Feminist Geography*. Hutchinson, London
- Wooldridge S W 1956 *The Geographer as Scientist*. Thomas Nelson, London

R. J. Johnston

Geopolitics

The idea of geopolitics evokes mixed reaction. It is understood variously as: a set of geographically determined 'laws' governing a state's strategic destinies; the geographical underpinnings of *Realpolitik*; and a scholarly analysis of the geographical factors underlying international relations and guiding political interactions. Widely embraced by policy makers, geopolitics has encountered scholarly criticism for its sometimes deterministic doctrines and lack of moral considerations.

1. Definitions

Geopolitics is a product of its times, and its definitions have evolved accordingly. Johan Rudolph Kjellén, who coined the term in 1899, described geopolitics as 'the theory of the state as a geographical organism

or phenomenon in space' (Kjellén 1916). For Karl Haushofer, the father of German Geopolitik, 'Geopolitics is the science of the conditioning of political processes by the earth ... of political space organisms and their structure ... [and] aims to be equipment for political action' (Haushofer et al. 1928). In 1939, Derwent Whittlesey, the American political geographer, dismissed geopolitics as 'a dogma ... the faith that the state is inherently entitled to its "place in the sun".' In contrast, Edmund Walsh espoused an American geopolitics based upon international justice, and 'a combined study of human geography and applied political science ... dating back to Aristotle, Montesquieu and Kant' (1944).

Saul Cohen defined geopolitics as a mode of analysis, relating diversity in content and scale of geographical settings to exercise of political power and identifying spatial frameworks through which power flows (Cohen 1973). Recently, Gearoid O'Tuathail argued that 'geopolitics does not have a singular, all-encompassing meaning or identity ... it is discourse, a culturally and politically varied way of describing, representing and writing about geography and international politics' (1998).

Because geopolitics straddles two disciplines—geography and politics—its approaches vary according to frameworks of analysis common to each. While most early theories and concepts in geopolitics grew out of geographical thought, historians and political scientists often failed to adapt their later theories to the dynamic, complex nature of geographical settings.

2. Stage One—The Race for Imperial Hegemony

Geopolitical thinking can be traced back to Aristotle, Strabo, Bodin, Montesquieu, Kant, and Hegel. Nineteenth century precursors include von Humboldt, Guyot, Buckle, and Ritter. Modern geopolitics has developed through four stages—the race for imperial hegemony, German Geopolitik, the Cold War era, and the post-Cold War years. The theories of its founders, Ratzel, Mackinder, Kjellén, and Mahan, reflected their era of intense nationalism, state expansionism, and overseas empire-building. Their principles and 'laws' reflected national perspectives and experiences, as well as the influence of social Darwinism.

2.1 Ratzel

Friedrich Ratzel (1844–1904), German natural scientist and 'father' of political geography, was the first to treat space and location systematically in his comparative studies of states (1897). He provided successor geopoliticians with a 'scientific' basis for state expansionist doctrines that reflected Germany's nineteenth century experiences and ambitions for the future. During the last half of the century, Germany emerged as continental Europe's chief economic and military

power. Unified under Bismarck's leadership and victorious in its wars with Austria and France, Germany had enlarged its territory, expanded its heavy industries, and enacted social reform. With the aid of a new powerful naval fleet, Germany posed a serious threat to Britain and France as it acquired an overseas empire in Africa and the Western Pacific, and commercial footholds in East Asia.

Ratzel based his system upon principles of evolution and science (1882/1891). He viewed the state as an organism fixed in the soil, whose spirit derived from mankind's ties to the land. His geographical 'laws' focused on space (*Raum*) and location (*Lage*), the former dependent upon and contributing to the political character of groups living in the space, the latter providing space with its uniqueness. He introduced *Lebensraum* (living space) as the law driving states to expand. Ratzel's 'organic' theories of state growth fitted Germany's view of its future as a youthful, aggressive, capitalist 'Giant State.'

2.2 Mackinder

Halford Mackinder (1861–1947), who established geography as a university discipline in Britain, was concerned with safeguarding the British Empire's political, commercial, and industrial primacy, in an era when command of the seas no longer guaranteed world supremacy owing to the wider reach and speed of rail. For Mackinder, geographical realities lay in the advantages of centrality of place and efficient movement of ideas, goods, and people. He theorized that the inner area of Eurasia, characterized by interior or Arctic drainage and impenetrable by sea power, was the 'pivot area' of world politics (1904). He warned that rule of the heart of the world's greatest landmass could lead to world domination. A Eurasian land-power (be it Russia, Germany, or even China, and especially an alliance of the first two) that gained control of the pivot area would outflank the Maritime world. Eleven years later, the English geographer James Fairgrieve (1915), who coined the term 'Heartland' opined that China was in an excellent position to dominate Eurasia.

In *Democratic Ideals and Realities* (1919) Mackinder, now using the term 'Heartland,' and taking into account advances in land transportation, population increases, and industrialization, enlarged his map to include Eastern Europe from the Baltic through the Black Sea as Inner Eurasia's strategic annex. This became the basis for his dictum, 'Who rules East Europe commands the Heartland; who rules the Heartland commands World-Island; who rules World-Island commands the world.' The warning to Western statesmen was clear—the key to world domination lay in the middle tier of German and Slavic states, or Mitteleuropa. Mackinder's world-view guided Lord Curzon's imperial strategies in South Asia and South Russia (Blouet 1987), German Geopolitik between

World War One and Two, and for Western containment strategies of the post-World War Two era. Foreseeing Britain's decline as the world's leading power (Mackinder 1924), he advocated that Western Europe and North America constitute a single community of nations—a forerunner of the North Atlantic community concept.

Whereas Ratzel's theories of the large state were based on concepts of self-sufficiency, closed space, and totalitarian controls, Mackinder was strongly committed to cooperation among states, democratization of the Empire into a Commonwealth of Nations, and preservation of small states. He bridged the academy and politics, serving as a Conservative and Unionist member of Parliament (1910–1922), and as British High Commissioner for South Russia (1919–1920). Initially a Liberal Imperialist and proponent of free trade, he eventually became committed to a preferential tariff system to protect British imperial unity.

During World War Two, Mackinder presented a global strategy based upon two equal power units—Heartland and Midland Ocean (the North Atlantic)—separated from two other regions (Monsoonal Asia and South Atlantic lands) by a mantle of empty deserts and mountains. He felt that, ultimately, Monsoonal Asia (India and China) would emerge to challenge a combined Heartland/Midland Ocean (1943).

2.3 Kjellén

Johan Rudolph Kjellén (1864–1922), the political scientist who, in 1899, coined the term 'geopolitics' was influenced by both his Swedish background and Germany's growth into a giant state. He viewed the impending breakdown of the Concert of Europe and the drift towards war and chaos as the death-knell for small states like Sweden (Parker 1998). Adopting Ratzel's organic state concept, he considered Germany's emergence as a great power inevitable and desirable. Sweden's needs would be fulfilled within the framework of a new Mitteleuropean bloc from Scandinavia and the Baltic, through Eastern Europe and the Balkans, dominated by ascendant Germany.

A Conservative member of the Swedish parliament, Kjellén viewed geopolitics as the 'science of the state,' whereby the state's natural environment provided the framework for a power unit's pursuit of 'inexorable laws of progress.' The dynamic organic approach led Kjellén to espouse the doctrine that political processes were spatially determined (1916). Moreover, since giant states in Europe could be created only by war, he viewed geopolitics as primarily a science of war (Gyorgy 1944).

2.4 Mahan

The global perspective of Admiral Alfred Thayer Mahan (1849–1914) was also Eurasian-centered

(1900). For Mahan, the northern land hemisphere, connected by Panama and Suez, was the key to world power; within that hemisphere, Eurasia was the most important component. Mahan recognized Russia as the dominant Asian landpower whose location made it unassailable. However, he felt that its landlocked position put it at a disadvantage, because, contrary to Mackinder's opinion, sea-movement was superior to land-movement (1890).

Mahan argued that world dominance could be held by an Anglo-American alliance from key bases surrounding Eurasia. The critical zone of conflict lay between the thirtieth and fortieth parallels in Asia where Russian landpower and British seapower met. He viewed the USA as an outpost of European power and civilization, its Pacific shoreland and islands extensions of the Atlantic-European realm. Indeed, he predicted that an alliance of the USA, Britain, Germany, and Japan would one day hold common cause against Russia and China.

3. Stage Two—German Geopolitik

German Geopolitik emerged in reaction to Germany's devastating defeat in World War One. Humbled by the Treaty of Versailles, Germany was stripped of its overseas empire and important parts of its national territory, and the social cohesion forged by Bismarck's policies was shattered. The Socialist Weimar Republic was beleaguered by class warfare and attempts to overthrow it by Communists on the Left and racist militant nationalists and aristocratic conservatives on the Right. Unemployment was rampant and inflation raged. Within this setting, Haushofer and his colleagues established the *Zeitschrift für Geopolitik* (1924–1939) and the Institute for Geopolitics at the University of Munich. Undoing Versailles by restoring the lost territories and rebuilding Germany as a world power undergirded the pseudo-scientific 'laws' and principles of Geopolitik that served Nazi Germany.

3.1 Haushofer

Karl Haushofer (1869–1946), the general who became a political geographer, was not an original thinker. The Geopolitik of the German geopoliticians whom he led (Otto Maull, Erich Obst, Ewald Banse, Richard Hennig, Colin Ross, Albrecht Haushofer) was based essentially upon the writings of Kjellén, Ratzel, and MacKinder. Mahan, Fairgrieve and the American geographical determinist Ellen Churchill Semple were also influential. The organismic-Hegelian philosophy of Geopolitik came from Ratzel directly or via Kjellén.

Lebensraum and autarchy became slogans for doctrines whose consequences were conflict and total war. Three geographical settings permeated the literature

of Geopolitik: Ratzel's large states, Mackinder's World-Island, and Pan-Regions. The organic growth of Germany to its west and east was regarded as 'inevitable.' To gain mastery over World-Island, it was necessary to dominate the USSR and destroy British seapower. The geopoliticians posited that German control over Pan-Europe (including Eastern Europe) would force the Soviet Union, regarded as an Asian power, to come to terms.

During most of the 1920s and 1930s, Haushofer espoused continental pan-regionalism based upon complementarity of resources and peoples: Pan-America, Pan-Eurafrica, and Pan-Asia, with the USA, Germany, and Japan as respective cores. His position on the USSR was ambiguous. He proposed variously a German-Russian alliance, a Pan-Russia-South Asia, and a Japan-China-Russia bloc. His call for Germany, the USSR, and Japan to form a Eurasian pan-region that would dominate World-Island, influenced the German-Soviet pact of 1939, made moot by Hitler's subsequent invasion of the Soviet Union. The principles of Geopolitik were designed to fulfill German national and imperial aims, without concern for contradictions.

Haushofer's extraordinary influence derived from his close ties to Rudolph Hess, his aide-de-camp in World War One and subsequent student at the University of Munich. Through Hess, he gained access to Hitler from 1923 to 1938. Doctrines such as *Blut und Boden* (blood and soil), *Rasse und Raum* (race and space), and *Lebensraum* were incorporated into *Mein Kampf* and became ideological tenets of the murderous Nazi regime. He also advised Hitler at Munich 1938. With Hess's flight to England in 1941, the influence of the Geopoliticians upon Hitler ended. Haushofer was imprisoned briefly at Dachau (ironically, he had a Jewish wife), and his son Albrecht, also a geographer with links to aristocrat military circles, was killed by the SS for involvement in the 1944 Generals' Plot to assassinate Hitler. Haushofer and his wife committed suicide in 1946.

4. Stage Three—The Cold War: State-centered vs. Universalistic Approaches

Onset of the Cold War re-awakened Western interest in geopolitics by historians, political scientists, and statesmen. Geographers distanced themselves from geopolitics because of the taint of German Geopolitik.

4.1 State-centered 'Political' Geopolitics

American Cold Warriors embraced geopolitics as a basis for national policy aimed at confronting the

Soviet Union and International Communism. Building on early geographically derived geopolitical theories, and holding static interpretations of global and regional spatial patterns, they introduced into the lexicon of Cold War geopolitics such political-strategic doctrines as containment, domino theory, balancing linkages, and linchpin state.

Germany and Japan emerged as the key military bases for the USA and its allies in the struggle to contain Soviet and Chinese continental power, thrusting the Allies' World War Two enemies into the strategic forefront. Military alliances forged as part of containment produced mixed results. NATO, established in 1949, has proved an enduring strategic instrument. However, SEATO, formed in 1954 to oppose the spread of Communism in Southeast Asia after France's withdrawal from Indochina, collapsed with the US defeat in Vietnam, and was disbanded in 1977.

In the Middle East, the Baghdad Pact (later called the Central Treaty Organization) was even shorter-lived (1955–1959). Its members included Turkey, Iraq, Iran, Pakistan, and Britain. Iraq's subsequent withdrawal to pursue a non-alignment policy left the pact with no Arab representation and therefore limited capacities to counter Soviet influence in Egypt, Syria, and South Yemen.

Western foreign policy was not long confined to containing the Eurasian continental power along Heartland's borders, but adopted a strategy of checking the spread of Communism throughout the Third World. The idealistic vision that had prompted the USA to support freedom and democratization among colonial peoples quickly gave way to expedient realpolitik—propping up right-wing dictatorships in order to stop the threat of Communism.

Another popular geopolitical doctrine, 'domino theory,' was first proposed by William Bullitt (1947), who feared that Soviet Communist power would spread, via China, into Southeast Asia. It was adopted by both the Kennedy and Nixon administrations, rationalizing American intervention in Vietnam to 'save' the rest of Southeast Asia (O'Sullivan 1982).

Domino theory was an important argument for extending Western containment policy well beyond the Southeast Asian and Middle Eastern shatterbelts into the Horn of Africa and Sub-Saharan Africa, Central America and Cuba, South America, and South Asia. These regions became battlefields for both superpowers, each supporting local surrogates militarily, politically, and economically. The goal was to protect or gain sources of raw materials and markets, while denying the enemy overseas military bases.

The imagery of 'dominos' persists. Threat of the spread of Kosovar Albanian irredentism to Macedonia, Bulgaria, and Greece precipitated NATO's air war against Yugoslavia. Ironically, the result has been Serbian ethnic cleansing of the Kosovars, further destabilizing the region.

A third doctrine, 'linkage,' introduced into geopolitics by Henry Kissinger in 1969, was based upon the theory of a network that connected all parts of the world's trouble spots to the Soviet Union, and that US involvement in any single conflict would have an impact upon overall superpower balance (Kissinger 1979). Linkage rationalized continued US fighting in Vietnam long after the war had been lost.

Linkage was applied to detente with the Soviet Union and accommodation with China. To maintain the balance of power, the Nixon administration sought Moscow's agreement on strategic arms limitations and mutual nuclear deterrence, and tried to play China off against the USSR. The logical consequence of this policy was non-intervention in the 1968 Czech uprising, acquiescing to the Brezhnev Doctrine, which held that military force was justified to keep the socialist countries of Eastern and Central Europe within the Soviet camp. The threat of credibility loss also served as a driving force in NATO's war against Yugoslavia.

Zbigniew Brzezinski's geopolitical world-view was based on the struggle between Eurasian landpower and seapower. For him, the key to preventing Soviet world dominance lay in US control of 'linchpin' states (1986). These were defined by their geographical locations which enabled them to exert strategic military and/or economic influence. The designated linchpins were Germany, Poland, Iran or Pakistan-Afghanistan, South Korea, and the Philippines. Their dominance by the USA would effectively contain the Russian 'Imperial' power, protecting Europe and Japan and, in the case of South Korea and the Philippines, preventing encirclement by China.

For Brzezinski the US-Soviet conflict was an endless game, and linchpin control was a necessary part of US geostrategic game plan. In this there is little consideration of the geopolitical complexity of the global system and of the multiplicity of forces beyond superpower reach that had become active agents in the system.

4.2 Universalistic Geographical Geopolitics

When geographers re-engaged in geopolitics in the 1960s and 1970s, they developed theories based upon universalistic/holistic views of the world and the dynamic nature of geographical space. Three approaches predominated: a polycentric, international power system; a unitary, economically based world system; and an environmental and socially ordered geopolitics. These fresh geographical theories which challenged bi-polar Cold War geopolitics had little appeal to the Cold Warriors and failed to make their way into 'political' geopolitics.

The polycentric or multinodal/multilevel power approach rejected the Heartland theory of world domination, as had Halford Mackinder in his later

writings. His reinterpretation became a take-off point for those geographers who developed theories of geopolitical polycentrism. In the USA, Saul Cohen proposed a flexible hierarchy of two geostrategic realms, geopolitical regions, shatterbelts, national states, and subnational units within a system that evolved through dynamic equilibrium forces (Cohen 1973). The USA, European Union, and Japan are the cores of the Maritime realm, and Russia and China of the Continental. This developmental theory posited that the structural components of the global system had evolved from a stage of undifferentiation with relatively few parts, to specialization/integration with many parts at different geo-territorial levels (Cohen 1982). Equilibrium is maintained by moving from one stage to another through responses to short-term disturbances.

In England, Chrone presented a geopolitical system of 10 regional groupings that were also hierarchically ordered, and had a historical and cultural basis (Chrone 1969). In Chrone's view, the world power balance was shifting from Europe and the West towards Asia and the Pacific. He predicted that the Pacific Ocean would become the future arena of confrontation for the USSR, the USA, and China.

Peter Taylor, the English geographer, broke away from the 'realistic' school of power-centered geopolitics in applying a world-systems approach based upon global economics (1989). He drew upon the work of Immanuel Wallerstein (1983), who argued that the world economy means a single global society, not competing national economies. Taylor integrated the Wallerstein model with George Modelski's cycles of world power (1978), and presented power and politics within the context of a cyclical world economy in which nation states and localities are fitted.

Taylor, as Wallerstein, viewed global conflict in North/South terms (rich nations vs. poor nations), arguing that capitalist core areas aggrandize themselves at the expense of the periphery. This offered a geopolitical frame for the North-South dichotomy that the 1980 Brandt Report highlighted in its study of Third World issues.

An environmentally and socially oriented geopolitics was promoted by Yves LaCoste in France with the establishment of the *Journal Hérodote*. In moving towards a 'new géopolitique,' he sought to overcome the national chauvinism of the 'old' geopolitics by focusing on the land, not on the state. Hérodote linked geopolitics to ecology and broader environmental issues, as well as such matters as world poverty and resource exhaustion. Much of LaCoste's work was inspired by the French human geographer and political anarchist, Elisée Reclus, who believed it essential to reshape the world's political structure by abolishing states and establishing a cooperative global system. While this French geopolitics did not produce systematic geopolitical theory, it did put the spotlight on applying geopolitics to significant global problems.

5. *Stage Four—The Post-Cold War Era: Competition or Accommodation?*

The end of the Cold War has generated a number of new approaches to geopolitics, in addition to those that continue from the earlier era. For some, it heralds a 'new world order' and the geopolitics of US global hegemony. President George Bush, addressing Congress in 1990 defined the policy behind the war against Iraq as envisaging a new world order led by the USA, 'freer from the threat of terror, stronger in the pursuit of justice, and more secure in the quest for peace ... a world in which nations recognize the shared responsibility for freedom and justice' (Bush 1991).

In contrast, Robert Kaplan's geopolitics of anarchy is based on a world divided into the rich North and the poor South. He concludes that the South, especially Africa, is doomed to anarchy and chaos. His map of the future, dubbed the 'last map' is an 'ever mutating representation of chaos' (1994).

Neither of these predictions has come to pass. In most cases, the overthrow of Communist regimes has not led to stable, free-market economies. The restraints upon the unilateral application of US military, economic, and political power are evident from the failures to gain US objectives in Iraq, Somalia, and Haiti. In southeastern Europe, Russia and China have been called upon to help mediate the NATO–Yugoslav conflict.

Chaos theory gives inadequate attention to the regional and global systemic forces that contain turbulence and absorb its positive aspects into the system. Indochina has been stabilized under the leadership of Vietnam. Nigeria has led the community of West African states in attempts to restore order in countries plagued by civil war. The West cannot ignore instability in the South despite regarding it as 'quarter-sphere of strategic marginality.' This is necessitated by humanitarian considerations, as well as fear of the spread of nuclear weaponry and massive out-migrations.

The main thrust of post-Cold War geopolitics continues to follow the two streams of the previous era—the nation-centered/political and the universalistic/geographical. Political geopoliticians advocate projection of Western power into Central and Eastern Europe to weaken Russia's Heartland position at its western edge. They also advance strategies for penetrating the Caucasus and Central Asia, and for playing China off against Russia.

Brzezinski's prescription for maintaining US global hegemony is to achieve primacy in three parts of the 'Eurasian Chessboard': the West—Europe; the South—Middle East and Central Asia; and the East—China and Japan. He advocates pulling Ukraine and the Black Sea into the Western orbit, strong engagement in Central Asia and the Caucasus, and support of Chinese regional dominance in peninsular southeast Asia and Pakistan. China would be limited

to regional power status by the globally framed US–Japan strategic alliance. On the assumption that Russia's weakness is a permanent condition, this strategy seeks to prevent Russia from reasserting strategic control over 'near abroad' states, or from joining with China and Iran in a Eurasian anti-US coalition (Brzezinski 1997).

Advancing a geopolitics of 'the West against the Rest,' Samuel Huntington argues that world primacy can be maintained by dividing and playing off the other civilizations (1993). His thesis is that the fundamental sources of conflict will not be ideological, but cultural, and the fault lines between civilizations will be the battle lines. In dividing the world into Western, Confucian, Japanese, Islamic, Hindu, Slavic-Orthodox, Latin American, and possibly African civilizations, he makes little allowance for religious, ethnic, economic, or strategic divisions. He also assumes the permanence of the fault lines, despite the massive changes that are brought about by demography, migrations, and modernization.

Geographical geopolitical theory continues to reflect the universalistic approaches advanced during the Cold War. Building on the works of Taylor and LaCoste, 'critical' geopolitics applies social scientific critical thinking to ask how power works and might be challenged. This school views the following as central to geopolitical studies: analysis of discourse—rhetoric, metaphors, symbolism (O'Tuathail and Agnew 1992); feminist approaches to the subject of national security; and the geographies of social movements, particularly in relation to newly radicalized and participative democracy (Painter 1995).

In contrast, reality-based, polycentric geographical geopolitics emphasizes the continuous proliferation and inter-connectedness of the various parts and levels of the world. The current number of 200 national states could increase to 250–275 within the next quarter of a century. As the pace of devolution quickens, some of these new geo-territorial entities will be highly autonomous 'quasi-states.' Regional geopolitical restructuring is also a continuing process. Just as Southeast Asia has disappeared as a shatterbelt, so may the Middle East. The effort to detach Eastern Europe from the Russian Heartland is underway. However, with Russia's revival, the region could serve as a gateway between Europe and Russia.

In addition, the network of global cities—centers of capital flows and financial services, linked ever more closely by cyberspace, tourism and immigrant communities—will emerge as a major new geopolitical level, promoting policies sometimes contradictory to national interests. International social movements, such as environmentalism, are also influential in shaping national and regional policies, including military ones.

Whatever the course of geopolitical restructuring, we are entering an era of power-sharing among a wide variety of states of different sizes and functions.

Geopolitical theory will continue to be a valuable tool for understanding, predicting, and formulating the structure and direction of the world system.

See also: Globalization: Political Aspects; International Justice: Philosophical Aspects; International Law and Treaties; International Relations, History of; Military and Politics; National Security Studies and War Potential of Nations; War: Causes and Patterns

Bibliography

- Blouet B W 1987 *Halford Mackinder*, 1st edn. Texas A & M Press, College Station, TX
- Brzezinski Z 1986 *Game Plan*, 1st edn. Atlantic Monthly Press, Boston
- Brzezinski Z 1997 *The Grand Chessboard*, 1st edn. Basic Books, New York
- Bullitt W 1947 A Report to the American people on China. *Life* 23: 35–7
- Bush G 1991 *The Hard Work of Freedom*, Public Papers of the Presidents of the United States. US Government Printing Office, Washington, DC
- Chrono G R 1969 *Background to Political Geography*. Pittman Publishers, London
- Cohen S B 1973 *Geography and Politics in a World Divided*, 2nd edn. Oxford University Press, New York
- Cohen S B 1982 A new map of global geopolitical equilibrium: A developmental approach. *Political Geography Quarterly* 1: 223–42
- Fairgrieve J 1915 *Geography of World Power*. University of London Press, London
- Gyorgy A 1944 *Geopolitics—The New German Science*. University of California Press, Berkeley and Los Angeles, CA
- Haushofer K, Obst E, Lautensash H, Maull O 1928 *Foundations of Geopolitik*. Kurt Vowinckel Verlag, Berlin
- Huntington S P 1993 The clash of civilizations? *Foreign Affairs* 72: 22–49
- Kaplan R 1994 Coming anarchy. *Atlantic Monthly* 273: 44–6
- Kissinger H 1979 *The White House Years*, 1st edn. Little Brown, Boston
- Kjellén R 1916 *The State as a Form of Life*. Hugo Gebers Forlag, Stockholm
- Mackinder H 1904 The geographical pivot of history. *Geographical Journal* 23: 421–37
- Mackinder H 1919 *Democratic Ideals and Reality*. Constable, London
- Mackinder H 1924 *The Nations of the Modern World*. George Philip, London
- Mackinder H 1943 The round world and the winning of the peace. *Foreign Affairs* 21: 595–605
- Mahan A T 1890 *The Influence of Sea Power Upon History*. Sampson Low Marston, London
- Mahan A T 1900 *The Problem of Asia and its Effect Upon International Policies*. Little Brown, Boston
- Modelski G 1978 The long cycle of global politics and the nation state. *Comparative Studies of Society and History* 20: 214–35
- O'Loughlin J (ed.) 1994 *Dictionary of Geopolitics*. Greenwood Press, Westport, CT
- O'Sullivan P 1982 Antidomino. *Political Geography Quarterly* 1: 57–64
- O'Tuathail G, Agnew J 1992 Geopolitics of discourse. *Political Geography Quarterly* 11: 190–204
- O'Tuathail G 1998 Thinking critically about geopolitics. In: O'Tuathail G, Dalby S, Routledge P (eds.) *The Geopolitics Reader*, Routledge, London, pp. 1–12
- Painter J 1995 *Politics, Geography and 'Political Geography'*. Arnold, London
- Parker G 1998 *Geopolitics—Past, Present and Future*. Pinter, London
- Ratzel F 1882 (pt.1)/1892 (pt.2) *Anthropogeographie*. Engelhorn, Stuttgart, Germany
- Ratzel F 1897 *Politische Geographie*. Oldenbourg Verlag, Munich, Germany
- Taylor P J 1989 *Political Geography*. John Wiley, New York
- Wallerstein I 1983 *Historical Capitalism*. Verso, London
- Walsh E 1944 Geopolitics and international morals. In: Weigert H, Stefansson V (eds.) *Compass of the World*, George Harrap, London, pp. 12–39
- Whittlesey D 1939 *The Earth and the State*. Holt, New York

S. B. Cohen

Geriatric Psychiatry

1. The Field of Geriatric Psychiatry

Geriatric psychiatry deals with disturbances of mental health in older persons. Its practitioners are physicians who have trained first in medicine and general psychiatry, then specialized in the treatment of psychiatric disorders affecting people in later life—usually taken to be from the seventies on. Geriatric psychiatry deals with the following: the diagnosis and accurate assessment of individual patients; the development and use of treatments that are the best available; and the administration of health services for older people, involving family physicians, community health workers, and carers.

The importance of geriatric psychiatry expanded greatly in the latter decades of the twentieth century, mainly because of the unprecedented change in the world's population: many more people are living into old age, and even more are becoming very old. For example, according to predictions published by the United Nations (1998), by 2050, the world's population aged 60 and over will be 74 percent larger than the number of people under 15 years of age. The population aged 80 years and older will rise from 61 million in 1995 to 1,055 million in 2150, a 17-fold increase. And it is in less developed countries where the most dramatic increases will occur. The social, economic, and public health consequences for many countries are profound. This is because the very elderly tend to have poor physical health, a greatly increased risk of dementia, and a degree of personal disablement that makes them dependent on others. The prevention and effective treatment of age-related disease has become a compelling need.

The agenda of geriatric psychiatry falls into three categories: how it is practiced; the main disorders treated; and the research questions or gaps in knowl-

edge that are currently the most important. Jeste et al. (1999) have set out the response that is needed to this forthcoming crisis. It is generally acknowledged that, to be effective, the practice of geriatric psychiatry requires close collaboration with other professionals, particularly nurses, social workers, clinical psychologists, family physicians, and specialists. This is because of the nature of the disorders being encountered, the behavior of the elderly person in poor health, and the simultaneous presence of both medical and psychiatric problems that so commonly occurs. Much of the practice of geriatric psychiatry is in the community and in primary care, not in specialist settings or hospitals.

2. Mental Disorders in Later Life

2.1 The Dementias

The most common group of disorders in geriatric psychiatry is the dementias. In these, changes have taken place in the structure and function of the brain, particularly areas concerned with memory, information processing, and social behavior. 'Dementia' originally meant 'out of one's mind,' from the Latin *de* (out of) and *mens* (the mind). Early in the nineteenth century, Esquirol (1845) gave a succinct definition of dementia as 'a cerebral affection ... characterised by a weakening of the sensibility, understanding, and will' (cited by Caine et al. 1995). In describing with such words how the condition can be recognized, Esquirol drew attention not only to the cognitive features of the disorder, with impairment of memory and thinking in day-to-day life, but also to its other manifestations such as apathy, deterioration in social behavior, occasional aggressiveness, delusional ideas, and hallucinations. These show how widespread the changes are in the brain. Following extensive consultations with experts in some 40 countries, the World Health Organization ([WHO] 1992) has described dementia as:

a syndrome due to disease of the brain, usually of a chronic or progressive nature, in which there is disturbance of multiple higher cortical functions, including memory, thinking, orientation, comprehension, calculation, learning capacity, language and judgement. Consciousness is not clouded. Impairments of cognitive function are commonly accompanied, and occasionally preceded, by deterioration in emotional control, social behaviour, or motivation. This syndrome occurs in Alzheimer's disease, in cerebrovascular disease, and in other conditions primarily or secondarily affecting the brain.

In Alzheimer's disease, an abnormal protein called β -amyloid accumulates in the brain, particularly in the hippocampus. In vascular dementia, areas of the brain become damaged by having their blood supply in-

terrupted, as happens in stroke. Other causes of dementia are Lewy body disease, Pick's disease, Huntington's disease, Creutzfeldt-Jakob disease, and AIDS.

2.1.1 Prevalence. There are published reports on the prevalence of dementia in diverse areas of the world including North America, Europe, Scandinavia, Russia, Japan, China, Singapore, India, Africa, and Australasia. An integrative analysis of 47 surveys across 17 countries has suggested approximate rates under one percent for dementia from any cause in persons aged 60 to 69 years, rising to about 39 percent in persons 90 to 95 years old (Jorm et al. 1987). The prevalence doubles with every five years of age within that range. Throughout the world, many more people than in the past are living to an age when the prevalence of dementia greatly increases, so that the absolute number of persons with dementia is now expanding rapidly. In some countries, the proportion with a dementia is increasing *much faster than the growth of the elderly population*. This is because of the greater increase in the very old, in whom the prevalence of dementia is highest.

2.1.2 Secular change. It is often asked if there is more dementia now than some decades ago. Because there are larger numbers of the elderly now, and more of the very elderly in particular, there has inevitably been a large increase in the number of persons suffering from a dementia. Furthermore, having developed a dementia, an elderly person is now likely to survive longer. This is at least partly attributable to the 'failure of success.' Even with no change in incidence, longer duration of illness leads to an increased prevalence. Studies in Sweden and across Western Europe have found no change in the age-specific incidence of Alzheimer's disease, but it is intuitively plausible that vascular dementia may be decreasing in many countries in parallel with a general decrease in vascular disease.

2.1.3 Age-related cognitive decline. There is no doubt that nearly all people undergo a deterioration in memory and a slowing of mental processes in very late life; and some experience these changes earlier than others for reasons that are not yet well understood. Various terms have emerged to describe such states. The term 'benign senescent forgetfulness' was introduced by Kral to describe one group of such states. Although it was not well defined by him, the term has persisted because it fulfills a need and has no sinister connotations. Other concepts have since

been introduced. A useful review of this important issue has been provided by Ritchie and Touchon (2000).

2.1.4 Risk factors for Alzheimer's disease. As reviewed by Jorm (2000) and Henderson and Jorm (2000), four risk factors can now be regarded as confirmed for Alzheimer's disease: age, a family history of dementia; having the Apolipoprotein E ϵ 4 genotype; and having Down's syndrome. There are then a number of risk factors where the evidence is at present not fully established: a first-degree relative with Down's syndrome, advanced maternal age, a head injury earlier in life, a previous depressive episode, hypothyroidism, and a long-standing history of physical inactivity. Aluminum has been considered, whether ingested in drinking water or the diet, or absorbed from antiperspirants. From all the evidence available, it remains unproven that aluminum from drinking water or any other source plays a role in causing Alzheimer's disease (Doll 1993).

2.1.5 Possible protective factors. Some factors may confer protection against dementia, or even specifically against Alzheimer's disease. For the latter, four have so far been proposed, each having been identified through epidemiological research: education, anti-inflammatory drugs, estrogen replacement therapy, and statins. Each of these points the way to the possibility of prevention. It is well established in medicine that prevention can precede an understanding of etiology.

2.1.6 Education. A recurrent finding in field surveys is that rates for dementia and cognitive impairment are higher in elderly persons who have had little education. One interpretation is that education may delay the point at which a developing dementia becomes clinically manifest. A recent longitudinal study of American nuns (Snowdon et al. 1996) suggests that intelligence or education, which are inextricably linked, may actually protect against the neuropathological processes in Alzheimer's disease. If such a protective effect does apply, the implications for public health could be appreciable because education has become much more widely available to young persons during the twentieth century.

2.1.7 Geographic and ethnic comparisons. Alzheimer's disease and vascular dementia may occur at different rates in different parts of the world and in different ethnic groups. But it is not easy to be cer-

tain what factors are responsible. There may be different patterns of mortality for dementia cases under different living conditions, different genetic vulnerabilities, or different environmental exposures. A finding of potential significance comes from a comparison of prevalence rates for dementia and specifically for Alzheimer's disease in elderly Nigerian Africans and African Americans in Indianapolis by Hendrie et al. (1995). The two groups have much of their heredity in common, but very different lifestyles. The prevalence was only 2.3 percent and 1.4 percent for dementia and Alzheimer's disease respectively in the Nigerians, but 8.2 percent and 6.2 percent in the combined community and nursing home sample of African Americans in Indianapolis. This demonstration of significant differences in rates in two different communities having similar ethnic origins again points to the possibility that some unknown environmental factors promote the neuropathological processes in Alzheimer's disease. The Cree Indians in Canada have also been found to have a lower prevalence of Alzheimer's disease than White Canadians (Hendrie et al. 1993).

In India, Ganguli et al. (1996) and Chandra et al. (1998) undertook a methodologically sophisticated survey of 5,126 individuals aged 55 years and over in Ballabgarh, Northern India. Using the same criteria to define a case, they found a prevalence of only 1.1 percent in persons aged 65 and over for Alzheimer's disease, much lower than in industrialized countries. Neither gender nor literacy were associated with prevalence, but this rose with age for both Alzheimer's disease and all dementias. Their conclusion was that the low prevalence compared to Western estimates could be due to shorter survival of cases, but it is also possible that the incidence is lower, speculatively attributable to differences in the underlying distribution of risk and protective factors.

2.1.8 Anti-inflammatory drugs. An inverse association was noticed between rheumatoid arthritis and Alzheimer's disease, suggesting that persons who have taken aspirin, steroids, or nonsteroidal anti-inflammatory drugs (NSAIDs) over long periods may have a lower risk of dementia, or have slower cognitive decline in late life. The most recent information is that anti-inflammatory drugs probably do prevent or attenuate the symptoms of Alzheimer's disease (Breitner 1996). But it is premature for physicians to prescribe anti-inflammatory drugs for Alzheimer's disease until the size of their effect is established in a randomized controlled trial and the findings are balanced with their risks.

2.1.9 Estrogen replacement therapy. Case-control studies have provided pointers to the possible

protection against Alzheimer's disease afforded to women who take estrogen. But since they also tend to be better educated and to differ in other life-style factors, this finding could be misleading. Again, only randomized controlled trials will establish whether a protective effect exists.

2.1.10 Risk factors for vascular dementia. The assumption is that risk factors are similar to those for stroke. These are age, male gender, a family history of stroke or dementia, hypertension, diabetes mellitus, transient ischemic attacks, elevated cholesterol or lipids, cigarette smoking, heavy alcohol use, oral contraceptives, physical inactivity, obesity, and ethnicity.

2.1.11 Treatment of dementia. The treatment of dementia presently consists of minimizing the deleterious effects of other conditions and of optimizing the person's day-to-day social environment. For Alzheimer's disease, medications have been developed that attempt to correct the deficiency in the neurotransmitter acetylcholine, but none has yet been found to bring about unequivocal and lasting benefit. Intense research is being conducted to find other drugs that enhance cognition. The possibility is emerging that people could be immunized against Alzheimer's disease in earlier adulthood to prevent the deposition of the harmful β -amyloid protein in the brain (St George-Hyslop and Westaway 1999). Meanwhile, remarkable progress has been made by clinical psychologists in developing nonpharmacological methods for treating dementia, including the reduction of severely difficult behavior such as wandering, incontinence, and screaming (Bird in press).

2.2 Delirium

Delirium, a term introduced by Celsus 2,000 years ago, is derived from the Latin *de* (from, away) and *lira* (a track or furrow). It is characterized by incoherence of thought and speech, together with perplexity, in a setting of clouded consciousness that fluctuates in severity. Despite its long-established place in nosology, delirium must number among the most underdiagnosed conditions in medicine. Delirium occurs most commonly at the extremes of life. Its highest incidence is in the elderly, in whom it may indicate the presence of systemic or cerebral disease, or drug toxicity.

2.3 Depression

The elderly might be expected to have higher rates of depressive disorders than younger adults because of

their personal and social circumstances. But here a puzzle emerges. Although a high prevalence is usually found in nursing homes, the consistent finding is that the elderly in the community have *on average* less depression than younger adults (Henderson 1994; Henderson et al. 1998). Some think these findings may be attributable to an error in ascertainment. This could be through inappropriately discounting symptoms that may also be due to physical illness or to age itself, such as tiredness, feeling slowed up, or sleep difficulties. Alternatively, a cohort effect has been proposed, in which people who are now elderly may have been protected against depression; or younger adults may have higher rates through being exposed to factors that older adults have been spared.

There is then a further intriguing possibility that is at first counterintuitive: that aging actually protects against depression, and indeed against a range of other mental disorders. It has been suggested that exposure to adverse experience over a lifetime may induce an increased resistance to depression and anxiety through psychological immunization.

2.3.1 Causes of depression in the elderly. When elderly persons have a depressive disorder, one may reasonably ask if it is a recurrence or if it has developed for the first time. Several studies have noted that severe life events precede the onset of depression in at least half of all patients; conversely, depressive disorder may emerge in elderly persons with relatively untrammelled lives. Risk factors for depression in late life are a past history of such a disorder, physical illness or chronic pain, recent adverse experiences, and the absence of a confiding relationship. Contrary to expectation, elderly persons who live alone have not been found to be at risk for depressive disorder. A consistent finding in community surveys is that the complaint of loneliness is associated more with depression than with actual social isolation.

Where depression develops in elderly persons for the first time in their lives, a recent finding is that the brain often shows selective atrophy on neuroimaging.

2.3.2 Treatment. Depression in older persons is highly responsive to psychotherapy and antidepressant medication that is highly effective, non-addictive, and usually well-tolerated. Often, the main problem lies in having the depression recognized in the first place, instead of its being ascribed by patient, family, and doctor to the inevitable experience of old age.

2.4 Suicide

In most countries, elderly men consistently have higher rates for suicide than other age-groups. Factors

associated with death by suicide in the elderly include depressive disorder, disabling physical illness, chronic pain, and social isolation.

2.5 Anxiety Disorders

Anxiety disorders in late life have been overlooked even more than depression as a significant cause of morbidity. In an integrative review of eight community-based surveys of persons aged 60 years and over, anxiety disorders were found to have a lower prevalence than in younger adults. The explanation is unknown. Co-occurrence of depression with generalized anxiety disorder or phobias was a frequent pattern, the depression usually being inappropriately treated with a benzodiazepine. Anxiety symptoms may also occur in the early stages of dementia and in association with depressive disorders. There is little information on the prevalence of post-traumatic stress disorder in the elderly, risk factors for it or its natural history.

2.6 Schizophrenia

It is recognized that schizophrenia can begin in later life. The term *paraphrenia* has been used for this condition. A number of distinguishing features may be present, as reviewed by Henderson and Kay (1997): a higher ratio of women, a better occupational history, and a higher frequency of marriage than in younger patients, with the occurrence of sensory loss being more common in those with onset at a more advanced age. Psychotic symptoms are appreciably more common than formal diagnostic categories. Impaired hearing and vision have both been found to be associated with psychotic symptoms. Social isolation is a further associated factor, but it is unknown if this is a premorbid characteristic, a causal factor, or a consequence. There is consistent evidence that paranoid or delusional symptoms are more frequent in persons who also have cognitive impairment.

3. Services for the Elderly

The extent to which the elderly with mental disorders receive services is itself an issue amenable to systematic examination. The notion of filters proposed by Goldberg and Huxley (1992) is very useful. The first filter determines which persons with mental disorders reach a physician, whether they initiate the consultation themselves or whether it is initiated by their families or the physician. The second filter determines whether the physician recognizes that significant mental disturbance is present. This may happen in only about half the consultations. The third filter determines referral to specialist services. Although family

physicians recognize early dementia quite efficiently, they may often need further training in the diagnosis of depressive disorders in elderly persons. This is because symptoms of tiredness, poor appetite, insomnia, and apathy may be mistakenly ascribed to old age. The benefits of correct diagnosis for both the patient and the community are self-evident.

4. Disability

From a public health perspective, information about the elderly and the services they require has to include an assessment of disability. This is because it is impairment of functioning in daily life rather than symptoms or diagnoses that brings about the need for services. Dementia is second only to arthritis as a cause of disability in the elderly. Less is known about the impairment of functioning caused by depressive disorders and depressive symptoms but, since they often occur with physical disease, their interactive effect is considerable. One community-based study conducted by an experienced geriatrician found that neuropsychiatric disorders of cognition, behavior, and mood, together with Parkinson-related disorders of gait, are major predictors of disability in late life.

5. Research in Geriatric Psychiatry

There has been an enormous expansion of research activity in recent decades. Much of it is directed at the following questions:

- (a) What determines the course of cognitive decline in old age?
- (b) Are there factors in the environment that contribute to dementia?
- (c) Can dementia be prevented?
- (d) Can drug treatments for dementia be found?
- (e) What are the applications of the new neuro-imaging techniques to diagnosis and treatment?
- (f) What improvements are needed in primary health care for older persons; and how can these be brought about?
- (g) What interventions will promote quality of life in the elderly?

The needs of tomorrow's communities, together with the rapid advances taking place in neuroscience, will ensure an exciting future for geriatric psychiatry. Its practitioners will need, as never before, to be highly trained but interpersonally skilled physicians.

Bibliography

- Bird M in press A psychosocial approach to challenging behaviour in dementia. In: Mulligan R, Van der Linden M, Juillerat A C (eds.) *Clinical Management of Early Alzheimer's Disease*. Erlbaum, Mahwah, NJ

- Breitner J C S 1996 Inflammatory processes and anti-inflammatory drugs in Alzheimer's disease: A current appraisal. *Neurobiology of Aging* **17**: 789–94
- Caine E D, Grossman H, Lyness J M 1995 Delirium, dementia, and amnesic and other cognitive disorders and mental disorders due to a general medical condition. In: Kaplan H I, Sadock B J (eds.) *Comprehensive Textbook of Psychiatry VI*. Williams and Wilkins, Baltimore, pp. 705–54
- Chandra V, Ganguli M, Pandav R, Johnston J, Belle S, DeKosky S T 1998 Prevalence of Alzheimer's and other dementias in rural India: The Indo-US study. *Neurology* **51**: 1000–8
- Doll R 1993 Alzheimer's disease and environmental aluminium. *Age and Ageing* **22**: 138–53
- Esquirol J E D 1845 *Mental Maladies: A Treatise on Insanity*, trans. Hunt E K. Lea and Blanchard, Philadelphia
- Ganguli M, Chandra V, Gilby J E, Ratcliff G, Sharma S D, Pandav R, Seaberg E C, Belle S 1996 Cognitive test performance in a community-based non-demented elderly sample in rural India: The Indo-US Cross-National Dementia Epidemiology Study. *International Psychogeriatrics* **8**: 507–24
- Goldberg D, Huxley P 1992 *Common Mental Disorders*. Tavistock/Routledge, London
- Henderson A S 1994 Does ageing protect against depression? *Social Psychiatry and Psychiatric Epidemiology* **29**: 107–9
- Henderson A S, Jorm A F 2000 Dementia: Disease definition, natural history and epidemiology. In: Maj M, Sartorius N (eds.) *Evidence and Experience in Psychiatry*. John Wiley, Chichester, UK
- Henderson A S, Jorm A F, Korten A E, Jacomb P, Christensen H, Rodgers B 1998 Symptoms of depression and anxiety during adult life: Evidence for a decline in prevalence with age. *Psychological Medicine* **28**: 1321–8
- Henderson A S, Kay D W K 1997 The epidemiology of functional psychoses of late onset. *European Archives of Psychiatry and Clinical Neuroscience* **247**: 176–89
- Hendrie H C, Hall K S, Pillay N, Rodgers D, Prince C, Norton J, Brittain H, Nath A, Blue A, Kaufert J, Shelton P, Postl B, Osuntokun B 1993 Alzheimer's disease is rare in Cree Indians. *International Psychogeriatrics* **5**: 5
- Hendrie H C, O'Osuntokun B O, Hall K S, Ogunniyi A O, Hui S L, Unverzagt F W, Gureje O, Rodenberg C A, Baiyewu O, Musick B S, Adeyinka A, Farlow M R, Oluwole S O, Class C A, Komolafe O, Brashear A, Burdine V 1995 Prevalence of Alzheimer's disease and dementia in two communities: Nigerian Africans and African Americans. *American Journal of Psychiatry* **152**: 1485–92
- Jeste D V, Alexopoulos G S, Bartels S J, Cummings J L, Gallo J J, Gottlieb G L, Halpain M C, Palmer B W, Patterson T L, Reynolds C F, Lebowitz B D 1999 Consensus statement on the upcoming crisis in geriatric mental health. *Archives of General Psychiatry* **56**: 848–53
- Jorm A F 2000 Risk factors for Alzheimer's disease. In: O'Brien J, Ames D, Burns A (eds.) *Dementia*. Arnold, London, pp. 383–90
- Jorm A F, Korten A, Henderson A S 1987 The prevalence of dementia: A quantitative integration of the literature. *Acta Psychiatrica Scandinavica* **76**: 465–79
- Ritchie K, Touchon J 2000 Mild cognitive impairment: conceptual basis and current nosological status. *Lancet* **355**: 225–8
- Snowdon D A, Kemper S J, Mortimer J A, Greiner L H, Wekstein D R, Markesbery W R 1996 Linguistic ability in early life and cognitive function and Alzheimer's disease in late life. *Journal of the American Medical Association* **275**: 528–32

- St George-Hyslop P H, Westaway D A 1999 Antibody clears senile plaques. *Nature* **400**: 116–17
- United Nations, Department of Economic and Social Affairs, Population Division 1998 *World Population Projections to 2150*. United Nations, New York
- World Health Organization 1992 *The ICD-10 Classification of Mental and Behavioural Disorders. Clinical Descriptions and Diagnostic Guidelines*. World Health Organization, Geneva, Switzerland

S. Henderson

Gerontological Clinical Psychology

Clinical geropsychology is the field that specifically focuses on developing empirically supported skills for the psychological assessment and treatment of older adults. Adults who might merit such attention include those with psychopathology or whose quality of life is negatively influenced by psychosocial factors. Given the increasing population of older adults and the complex life experiences and changing demographics of the population, it is important for mental health professionals to be prepared to assess and treat older clients.

1. Proficiency in Clinical Geropsychology

Until recently, outpatient mental health care services have been underutilized by older adults, despite 1975 legislation mandating specialized services for mentally ill older adults within community mental health centers. The majority of the elderly who receive mental health services are seen as inpatients during hospitalization in a psychiatric hospital or a nursing home. Possible explanations for this phenomenon include the idea that there is stigma attached to receiving mental health services within this cohort, but this concern has not been supported in limited research (Rokke and Scogin 1995). Alternatively, mental health professionals have historically displayed 'professional ageism' dating back to Freud, who was pessimistic about psychological change or the benefits of therapy in later life. Finally, Medicare reimbursement for psychological services in the United States is very limited. The majority of the money spent goes toward psychopharmacology as opposed to psychotherapy, reducing the availability of psychological services.

However, patterns of usage of psychological services by the elderly are changing. Successive cohorts of older persons have higher levels of education and a greater acceptance of psychology. Rokke and Scogin (1995), for example, showed that older adults considered cognitive therapy to be more credible and acceptable than drug therapy for depression (see

Cognitive Therapy). It is likely that psychologists could be much more active in reaching out to older adults to provide services, and that they can expect an increasingly positive welcome.

As more older adults seek or accept psychological services, psychologists must be prepared to meet this need. To address issues of responsibility and competency in providing such care, the American Psychological Association (APA) has approved Clinical Geropsychology as an area of special proficiency. Guidelines addressing the needed competencies for proficiency in clinical geropsychology have been drafted and widely promulgated (APA Interdivisional Task Force 1999). In this document, 13 areas of knowledge and skills that define proficiency knowledge in Clinical Geropsychology are specified, as listed below and then reviewed in more detail:

- (a) Research and theory in aging
- (b) Cognitive psychology and change
- (c) Social/psychological aspects of aging
- (d) Biological aspects of aging
- (e) Psychopathology
- (f) Problems in daily living
- (g) Sociocultural and socioeconomic factors
- (h) Assessment
- (i) Treatment
- (j) Prevention and crisis intervention services
- (k) Consultation
- (l) Interface with other disciplines
- (m) Special ethical issues in providing services to the aged.

1.1 Research and Theory in Aging

It is important for psychologists to understand aging in a broad context, as a psychological, biological, and social process. Further, some research methodology issues are especially relevant to aging, in particular understanding the impact of cohort or group differences by age.

1.2 Cognitive Psychology and Change

Changes in cognitive and behavioral function are common with aging; most such changes are relatively minor, such as increased difficulty with short-term memory or recall memory. However, the prevalence rate of more severe cognitive changes, indicating dementia, also increases with age. At age 65 only 1 percent of the population have dementia, compared to 30–35 percent by age 85 and 50 percent by age 90. The concept of dementia has evolved from the nonspecific ‘organic brain syndrome’ to a more precise picture of a complex disorder encompassing several distinctive disease entities, each with its own specific set of treatment implications (see *Dementia: Psychiatric Aspects*). More detail is provided on cognitive changes

with age under proficiencies Sect. 1.5, Psychopathology, Sect. 1.8, Assessment, and Sect. 1.9, Treatment.

1.3 Social/Psychological Aspects of Aging

Aging is a personal process and a social process that occurs in a context of familial, cultural, and political elements. As reviewed above, the demographics of aging reflect a changing social context of aging. This changing context also affects intergenerational relationships, the nature of relationships in later life (e.g., since more adults are living to old age, larger peer networks of older adults exist than in prior generations), and perceptions of aging.

1.4 Biological Aspects of Aging

Important biological aspects of aging include normal biological aging changes, abnormal changes and disease, lifestyle and behavioral factors in health, chronic and terminal illness, and pharmacological issues. Chronic health problems (e.g., arthritis, hypertension, hearing impairment, heart conditions) are often experienced by older adults. Many of these physical conditions are associated with chronic pain, and psychologists can employ strategies to help with pain management. Because of health concerns, many older adults are given medications, often without good explanation of the importance of taking the medication routinely, resulting in erratic usage. Psychologists can work with older adults to encourage consistent use of prescribed medication, to help them understand its importance, and to help identify the health-related belief system of older adults and incorporate those beliefs into the medical regimen. Insomnia, particularly sleep-maintenance insomnia, is an age-related and debilitating condition. Behavioral procedures for treating insomnia in older adults are more effective than relaxation training or imagery training. In addition, they offer the ability to avoid or discontinue sleep-inducing medications that can cause iatrogenic problems (Morin et al. 1999).

1.5 Psychopathology and Aging

The best estimate of the rate of psychopathology in the older population in the United States is approximately 22 percent (Gatz and Smyer 1992), but there are huge differences among different groups of older adults and across different types of psychopathology. Specifying the prevalence and incidence of psychopathology in older adults is complicated by problems with methodology and design. Most studies on older adults do not take racial and or cultural distinctions into account in the manifestation of disease. Finally, studies do not

always agree on diagnostic criteria and terminology, thus yielding imprecise estimates. For all disorders, the accepted criteria for older adults are exactly the same as those for younger adults using the *Diagnostic and Statistical Manual-IV (DSM-IV)* diagnostic system. However, interpreting some of the criteria for each disorder may be complicated by other aspects of the aging process. In the following subsections, some of the most common mental disorders will be examined and, when appropriate, issues that make diagnosis more complex in older adults will be addressed.

Older adults have a lower rate of major depressive disorder than middle-aged adults; however, there is a higher prevalence of subclinical symptoms in older adults (Gatz et al. 1996). Data from a number of studies indicate that across the adult life span, the highest depression scores are found among younger adults and persons 75 years and older. As with younger adults, women show higher rates of depression than men, although studies suggest that by age 80, men and women have an equal rate of depressive symptomatology. A subset of depressed older adults that warrant particular concern is white men aged 60–85, particularly those with medical illness or living alone; this group has the overall highest rate of suicide.

Life events, such as deaths of loved ones and change in residence, have shown an inconsistent relationship with depression. Although these negative life events occur more commonly in older adults, some scholars argue that older adults have experienced a number of stressors earlier in life, and/or that older adults are better able to cope and regulate emotion more effectively (e.g., Carstensen and Charles 1998), causing the relationship between negative life events and depression to be less strong among older adults. Physical illness, although a statistically major risk factor for depression, probably is not the causal variable in this relationship. Instead, loss of physical function, such that one cannot take care of basic needs (e.g., dressing, bathing) or must give up valued activities, is the significant vulnerability factor (Zeiss et al. 1996; see also *Depression, Clinical Psychology of*).

Anxiety disorders in older adults occur more often than depression, with estimates of approximately 5.5 percent prevalence (e.g., Beck and Stanley 1997). However, in comparing younger and older adults, the rate of anxiety is lower in older adults than in any other age group. Older women are more likely to be diagnosed with anxiety than older men. Although more prevalent than depression, anxiety disorders have been less studied in older adults. Panic disorders, phobias, generalized anxiety disorder, and post-traumatic stress disorder in older veterans have received the greatest attention.

Dementia is characterized by loss of cognitive function sufficient to impair performance of daily activities. Most dementias are due to irreversible changes in the brain; the most common are Alzheimer's disease and vascular dementias. Revers-

ible causes occur in 10–20 percent of dementia patients, as a response to a medication or physical illness, such as thyroid dysfunction, vitamin B12 deficiency, or a treatable brain lesion. Usually, when reversible causes of dementia are treated, good recovery ensues with return to prior level of cognitive functioning. Since most forms of dementia are progressive, it is important to identify the disease early. However, assessment of dementia in older adults is difficult; in particular, patients differentiating depression and dementing illness can be difficult to differentiate (Kasznik and Scogin 1995). Alzheimer's disease affects not only the person with the disease, but also the family (Thompson and Gallagher-Thompson 1996). Family members experience a range of emotions that can include guilt, anger, and depression. There are also positive components of providing care for a person with Alzheimer's, and many family members report caregiving as a rewarding experience.

1.6 Problems in Daily Living

Clinical geropsychologists need to be aware of many problems that can interfere with the ability to carry out normal daily activities for older adults. These include deficits in social and daily living skills, decision-making capacity, and elder abuse and neglect. Of these issues, elder abuse is one of the most difficult to handle; law in all states mandates reporting of elder abuse; the primary goal of reporting is to stop the abuse.

1.7 Sociocultural and Socioeconomic Factors

A variety of sociocultural issues are important in understanding the experience of aging, including losses from a variety of sources (deaths, end of employment, limitations in functioning, financial problems, environmental restrictions); cross-cultural and minority aging issues (e.g., interacting with older minorities to convey respect); the role of religion and spirituality; and the roles of intergenerational acquaintances, lifelong friendships, intimate relationships, and other social relations.

1.8 Special Issues in the Assessment of Older Adults

Clinical geropsychologists are asked to evaluate older adults with regard to presence or absence of psychopathology, cognitive capacity, competency to make decisions, substance abuse, etc. As with younger adults, the techniques used in psychological assessments include clinical interviewing, life history record and data review, cognitive and neuropsychological evaluations, behavioral assessments, and situational observations; however, with older adults, psycholo-

gists need to use different norms for tests and to include cognitive testing more frequently. With patients who exhibit disruptive and/or potentially harmful behavior (e.g., wandering, shouting, assaultiveness), behavioral assessment is helpful in determining what type of techniques would be useful for working with the patient (see Lichtenberg 1999 for thorough overview).

1.9 Treatment of Older Adults

Psychotherapy with older adults is as effective as therapy with younger adults. Most research on psychotherapy with older adults has used cognitive-behavioral approaches, and these have been shown to be effective for a wide variety of problems (Scogin and McElreath 1994). There also is a small amount of literature supporting the use of other approaches, such as Interpersonal Therapy and Brief Time-limited Psychodynamic Therapy, especially with depression in older adults (see Zarit and Knight 1996).

There are few major content differences in therapy with older compared to younger adults. Older adults have more health problems resulting in functional impairment, and psychological status is often related to their functional status. In addition, older adults may face obstacles in terms of resources for supporting adequate quality of life, such as limited financial resources or transportation, or the experience of loss of friends or family. The largest content difference is the need to deal more frequently with older adults with therapy issues related to death and dying. Many older adults accept death as part of the life course and engage in active adaptations to death, for example, by making wills, leaving instructions, negotiating interpersonal conflicts, and reintegrating life experiences in a meaningful way, but others have greater emotional difficulty in facing death or dealing with losses of loved ones. Clinical geropsychologists must be sensitive to factors that influence the experience of death and dying, including gender, class, economic variables, and cultural considerations. Recent studies have shown that bereavement practices, once highly structured and culturally determined, are becoming diverse; as a consequence, what practices will be therapeutic is becoming more individualized rather than shared within a common culture (Riley 1992).

With regard to therapy process, chronological age is not a good marker for what adaptation of therapeutic strategies originally developed with younger adults may be needed. Therapy adaptations to enhance utility with older adults might be helpful or even necessary, but assessment of each client determines which adaptations are appropriate (see Zarit and Knight 1996 for a thorough summary of suggestions regarding therapy adaptations). Primary areas of adaptation include changes to compensate for memory and information processing deficits with older adults. As a result of cognitive changes, the pace of therapy may be slower

than with younger adults. More repetition of material may be necessary, and processing of new ideas may be slower. Memory aids, such as audiotapes of therapy sessions or written summaries of each session, may be helpful. It may help to present material in multiple ways, both because of potential sensory loss and because repetitions provide multiple routes to memory storage. A key phrase for therapists working with older adults is 'Say it, Show it, Do it': state it clearly, write it down, and help the client use the idea in a specific way.

On the positive side, some changes in therapy process reflect relative strengths of older adults. One such strength can be thought of as wisdom (Baltes and Staudinger 1993), which is an age-related, but not inevitable, consequence of life experience. Older adults have faced many difficult life experiences, and most can abstract helpful information from those experiences and describe personal skills that have helped them handle adversity. Showing respect for the client's accumulated experience can enhance therapy.

1.10 Prevention and Crisis Intervention Services with Older Adults

Outreach, referral, and early intervention are important for making assessment and therapy available to older adults. The organization of many health care systems presents significant obstacles to good care for older adults who may need care provided in their homes, need short, easy drives to reach services, need care for the patient and the caregiving system, etc. In addition, providing health promotion resources to older adults is useful yet often neglected. For example, exercise is effective in improving physical and mental health for older adults, but may not be recommended or supported by health care providers to the extent it is with younger adults.

1.11 Consultation

Since most health care providers are not well trained in providing services to older adults, yet the demand is great and growing, clinical geropsychologists frequently need to provide consultation. Consultation to other professionals is a major component of this work and includes providing specialized information to health care institutions, agencies, and community organizations; providing staff training; and consulting on program development. In addition, consultation to families, lay caregivers, and self-help and support groups is needed.

1.12 Interface with Other Disciplines

Since typical problems older adults face are multimodal, a comprehensive interprofessional approach

to assessment and treatment that encompasses diverse health care providers (e.g., geriatricians, nurses, social workers) as well as psychologists is usually needed. The *biopsychosocial* model combines clinical and scientific approaches to treatment by defining problems in relation to the interactions among biological, social, and psychological systems. This model captures the complexity inherent in addressing the problems of older adults and highlights the need for an inter-professional approach (Zeiss and Steffen 1998).

1.13 Special Ethical Issues in Providing Services to Older Adults

Special ethical issues with older adults include obtaining informed consent for treatment with cognitively impaired elders, balancing patient autonomy and self-determination with apparent needs for care, balancing competing interests between older adults and family members, role conflicts in nursing homes, and confidentiality issues in working with families and teams.

2. Summary and Future Directions

Older adults are a heterogeneous population in terms of values, motives, social and psychological status, and behavior, as well as a quickly growing one. The need for clinical psychologists to be more informed about work with this diverse population of older adults have been summarized in the definition of proficiency in clinical geropsychology. That definition emphasizes broad-based knowledge of aging and adaptations of assessment and treatment for older adults. It also emphasizes the need for psychologists to work as part of interprofessional teams to meet the broad biological, social, and psychological needs of older adults.

However, as helpful as this summary is, much of what we currently know is based on research with the current cohort of older adults, particularly those who are Caucasian Americans. That research does not fully represent the heterogeneity within older adults now and will be less representative of future generations of older adults. In particular, there is considerable speculation that the large number of children born after World War II (now nearing older adult status) may age differently; two competing issues have been raised. First, it is often suggested that this cohort of elders may be more open to psychological interventions, escalating needs for research on assessment and therapy with older adults and for enhanced training of all clinical psychologists regarding work with older adults. On the other hand, coming generations of older adults may be healthier and more functional (Waidmann and Liu 2000), reducing depression and other psychological sequelae of disability, thus reducing the need for psychological

interventions. The increasing proportion of minority elders also will create challenges; as current generations of minority Americans age they will carry with them salient features such as a continuing imbalanced sex ratio, segregated distribution, and proportion in poverty. Clearly, researchers and clinicians in clinical geropsychology need to be guided by research to date but open to dramatic shifts in the field.

See also: Geriatric Psychiatry

Bibliography

- American Psychological Association Interdivisional Task Force on Qualifications for Practice in Clinical and Applied Geropsychology 1999 Qualifications for practice in clinical geropsychology, Unpublished Working Draft No. 7. Division 12, Section II and Division 20, American Psychological Association, Bethesda, MD
- Baltes P B, Staudinger U M 1993 The search for a psychology of wisdom. *Current Directions in Psychological Science* **2**: 75–80
- Beck J G, Stanley M A 1997 Anxiety disorders in the elderly: The emerging role of behavior therapy. *Behavior Therapy* **28**: 83–100
- Carstensen L L, Charles S T 1998 Emotion in the second half of life. *Current Directions in Psychological Science* **7**: 144–9
- Gatz M, Kasl-Godley J E, Karel M J 1996 Aging and mental disorders. In: Birren J E, Schaie K W (eds.) *Handbook of the Psychology of Aging*, 4th edn. Academic Press, San Diego, CA, pp. 365–82
- Gatz M, Smyer M A 1992 The mental health system and older adults in the 1990s. *American Psychologist* **47**: 741–51
- Kasznik A W, Scogin F R 1995 Assessment of dementia and depression in older adults. *The Clinical Psychologist* **48**: 17–23
- Lichtenberg P A (ed.) 1999 *Handbook of Assessment in Clinical Gerontology*. John Wiley and Sons, New York
- Morin C M, Mimeault V, Gagne A 1999 Nonpharmacological treatment of late-life insomnia. *Journal of Psychosomatic Research* **46**: 103–16
- Riley J W 1992 Death and dying. In: (eds.) *Encyclopedia of Sociology*. Macmillan, New York
- Rokke P D, Scogin F 1995 Depression treatment preferences in younger and older adults. *Journal of Clinical Geropsychology* **1**: 243–57
- Scogin F, McElreath L 1994 Efficacy of psychosocial treatments for geriatric depression: A quantitative review. *Journal of Consulting and Clinical Psychology* **62**: 69–74
- Thompson L, Gallagher-Thompson D 1996 Practical issues related to maintenance of mental health and positive well-being in family caregivers. In: Carstensen L L, Edelstein B A, Dornbrand L (eds.) *The Handbook of Clinical Gerontology*. Sage Publishing, Thousand Oaks, CA, pp. 129–50
- Waidmann T A, Liu K 2000 Disability trends among elderly persons and implications for the future. *Journal of Gerontology: Social Sciences* **55B**: S298–307
- Zarit S H, Knight B (eds.) 1996 *A Guide to Psychotherapy and Aging: Effective Clinical Interventions in a Life-stage Context*. American Psychological Association Press, Washington, DC
- Zeiss A M, Lewinsohn P M, Rhode P, Seeley J R 1996 Relationship of physical disease and functional impairment to depression in older people. *Psychology and Aging* **11**: 572–82

Zeiss A M, Steffen A 1998 Interdisciplinary health care teams in geriatrics: An international model. In: Bellack A S, Hersen M (eds.) *Comprehensive Clinical Psychology*. Pergamon, Oxford, UK, Vol. 7

A. M. Zeiss

Gesture in Linguistics

When we communicate, we not only produce a string of words, but also orchestrate various potentially interpretable body movements such as gesture, gaze shift, holding of a certain posture, interpersonal spacing, self-touch, and bodily symptoms of emotional and physiological states. From the totality of these activities in the communicative situation, this article takes up gesture and examines its relationship to language. In the first section of the article the definition of gesture is discussed and the following sections review its role in conversation, their relation to linguistic semantics and pragmatics, and mental processes underlying speech production.

1. What is Gesture?

Though there is controversy as to how to classify nonverbal behaviors (Efron 1941, Ekman and Friesen 1969, McNeill 1992), the central cases of nonverbal behaviors that have been designated by the term 'gesture' can be captured by the following definition. Gesture is nonpractical, nonlinguistic discrete body movement that arises from the intention to express something, and that encodes some aspects of what is intended to be expressed and/or how the accompanying speech should be interpreted.

Gestures are discrete spurts of body movement, which occur in the background of a period of nongesturing. This is similar to verbal utterances occurring in the background of silence. The sporadic nature of gestures contrasts with the inherently continuous nature of other nonverbal semiotic codes such as posture, interpersonal spacing, and gaze.

Gesture is nonpractical in the sense that gestural body movement is not involved in achieving non-expressive practical goals. For example, the way one holds a cigarette may be used intentionally to express a certain attitude or social affiliation, but it is not considered to be gesture because the body movement serves a practical purpose for smoking.

Gestures arise from the intention to express something. Gesture can be the intended code for transmission of the information. Raising a hand when taking a vote in a meeting is such a case. However, gesture can arise in the situation where gesture seems not to be the code intended for the recipient. For example, it is not uncommon to produce direction-

indicating gestures as one gives directions on the telephone. In such cases, the gestural body movement is still caused by the intention to express the instruction for how to get to the destination.

In this respect, gesture contrasts with self-touch, such as rubbing the chin or fiddling with clothing, in which there is no apparent intention to express. Self-touch also appears in certain noncommunicative settings, for example, while solving a crossword puzzle alone. The presence of expressive intention also distinguishes gesture from bodily symptoms for certain emotional and physiological states, such as blushing. However, some of the affect-displays, such as frowning, can be produced under voluntary control. They can be intentionally produced to express certain affect. This type of behavior blurs the boundary between what is a gesture and what is not.

The issue of intentionality has raised both conceptual and methodological concerns in defining gesture. Many authors maintain that intentionality is an essential ingredient of gesture (Ekman and Friesen 1969, Kendon 1981). However, difficult questions arise once one tries to tackle the issue squarely. Does intentionality have binary states (presence vs. absence) or is it a matter of degree? How can one tell what is intended to be expressed? How can one tell whether a given act is intentional, or how can one assess the degree of intentionality? Kendon (1981) suggested that a solution to the problem of identifying intentionality may be to focus on whether communicational partners treat each other's acts as intentional or unintentional. However, this does not solve the problem for gestures that cannot be seen by the communicational partners, as in the case of gesturing during telephone conversation.

Going back to further definitional features of gesture, gesture encodes some aspects of what is intended to be expressed and/or how the accompanying speech should be interpreted. In other words, gestural body movement itself is the carrier of the information to be communicated. In this respect, gesture differs from 'placing,' in which the body moves an object to a meaningful position (Clark in press). For example, putting merchandise on the cashier's counter signals the intention to purchase the merchandise. In placing, the body movement brings about a meaningful spatial configuration, but in gesture, the body movement itself carries the meaning.

2. In What Way is Gesture Nonlinguistic?

One of the defining features of gesture is its non-linguistic nature. Gesture is distinct from sign language, which is language that transmits codes in the visuo-motor modality (see *Sign Language*). It bears different semiotic properties from language (Kendon 1993, McNeill 1992). Gestures vary as to the degree to which they are semiotically different from language.

2.1 Most Language-like Gestures—Emblems

Gestures that are the most language-like are ‘emblems’ (Efron 1941, Ekman and Friesen 1969) such as the Euro-American OK sign with a ring formed by the thumb and the index finger. Emblems share some properties with words in a language. Emblems have standards of form. Their form-meaning relationship shows a relatively high degree of arbitrariness. Standards of form and form-meaning pairing are shared by a community of users (Calbris 1990, Morris et al. 1979).

Despite the above similarities to words, emblems also exhibit important differences. First of all, emblems are extremely limited in their potential for sequential combination, that is, combinatorial rules that operate over multiple emblems, namely, something equivalent to syntax, are impoverished.

Furthermore, Sparhawk (1978) showed that Persian emblems are not built up on the basis of a unified system of contrasting formational features, namely, something equivalent to phonology. It was found that the palm facing up vs. down is a formational feature that contrasts many Persian emblems. However, for the emblems with a certain class of meaning, namely beckoning emblems (e.g., the ‘come here’ emblem), this feature is not contrastive, and thus the palm can be facing either up or down. This is because the feature is considered to be irrelevant for the iconic make-up of the emblems.

The conclusion that emblems are not governed by unified ‘phonology’ has an important consequence for the ‘Kinesics’ approach to the analysis of semiotic body movements (Birdwhistell 1970), which is inspired by structural linguistics. In this approach, the existence of a system of ‘kinemes,’ the gestural equivalent of phonemes, is a fundamental assumption. However, even in the most language-like class of gestures, ‘kinemes’ do not constitute a unified system.

In summary, emblems differ from words in a language in that emblems are not fully unified by formal principles that relate emblems with one another. In other words, formal systems that are equivalent to syntax and phonology organize emblems only to a very limited extent.

2.2 Least Language-like Gestures—Gesticulations

Gestures that are least language-like are called ‘gesticulations’ (Kendon 1993) or simply ‘gestures’ (McNeill 1992), which spontaneously accompany speech, and usually share semantic and/or discourse-pragmatic contents with the concurrent speech.

Gesticulation consists of two subtypes: ‘beats’ and ‘representational gestures.’ Beats are small rhythmic bidirectional movements, in which hands typically move up and down or flip out and back. They occur when the concurrent speech should be interpreted in a context different from the ordinary flow of discourse

(e.g., speech is being repaired, or speech is expressing the beginning of a new discourse unit) (McNeill 1992). Representational gestures express location, direction, action, movement of objects, and shape of object, by means of iconicity (i.e., resemblance between the gestural shape and the referent) and/or spatial indexicality (i.e., the spatial contiguity between gesture and the referent). Representational gestures can be metaphorical in the sense that they can spatialize abstract ideas (Calbris 1990, McNeill 1992).

Gesticulations share fewer semiotic properties with language than emblems. The form of gesticulations is less constrained by conventionalized standards. The iconic representation in a gesticulation is largely idiosyncratic moment-by-moment creation. Thus, gesticulations are flexibly shaped in accordance with the variation in the spatial or actional characteristics of the referent. A sequence of gesticulations can comprise a larger meaning unit only by virtue of iconicity, as a multigesture depiction of a ‘picture’ of the referent(s).

3. In What Ways Does Gesture Have an Essential Connection to Language?

In the previous section, the difference between gesture and language as semiotic systems was highlighted. Gesture, however, has essential connections to interactional, conceptual, and psychological aspects of language.

3.1 Gesture in Face-to-face Conversation

Gestures contribute to face-to-face conversation in various ways. First, gesture can express some aspects of the content to be conveyed, which may or may not be expressed in the concurrent speech, and gesturally expressed information is taken up by the interlocutor (Kendon 1994). The interlocutor’s attention to gesture is also evidenced by the fact that gestures elicit verbal acknowledgement like ‘uh huh,’ and nonverbal acknowledgement like nodding from the interlocutor (Goodwin and Goodwin 1986, Streeck 1994).

This leads to the second way in which gestures contribute to conversation. Gesture is a part of complex cues as to when nonspeaking interlocutors can jump in and take the speaking turn (Schegloff 1984, Goodwin and Goodwin 1986, Streeck and Hartge 1992). One of the relevant features of gestures is that they are initiated slightly prior to the portion of speech that expresses the related content, and their temporal extent typically overlaps with the co-expressive speech (Schegloff 1984, Morrel-Samuels and Krauss 1992, McNeill 1992). The temporal precedence of gesture projects that the speaker is going to deliver speech further, and this can preempt the interlocutor’s attempt to take the turn.

The third way in which gestures contribute to conversation is to mark discourse-pragmatic status of the concurrent speech. Kendon (1995) reports Southern Italian emblems that indicate discourse-pragmatic status of the concurrent speech. More specifically, they mark the topic-comment structure and the illocutionary force (e.g., demanding information, making a plea for understanding) of an utterance.

3.2 *Gesture and Deixis*

Gesture plays a crucial role in spatial deixis (see *Deixis*), in which the attention of the interlocutor is directed to the direction along which the referent (or the physical clue to the referent) can be found.

Gesture also plays a crucial role in 'abstract deixis,' in which a pointing gesture creates a referent in seemingly empty space in front of the speaker (McNeill et al. 1993). The speaker can point repeatedly to a locus in the empty space, as the same entity is referred to in the concurrent speech. The referent that such gestures spatialize can be something that could exist in space (e.g., a person who is not present), or something inherently nonspatial (e.g., time, plot-line of a story). Thus, the role of gesture in deixis is not only a passive one, in which a gesture is oriented toward a pre-existing target, but also an active and creative one, in which the act of gesturing materializes conceptualization of the speaker as a spatial entity in the extralinguistic context.

3.3 *Gesture and Speech Production*

Gesture and speech are produced in close temporal and informational coordination. As noted in Sect. 3.1, a gesture typically is initiated prior to the co-expressive portion of speech, and the gesture and the co-expressive speech temporally overlap. In addition, the production of gesture has a systematic relationship to intonational units. A gesture rarely crosses tone unit boundaries, and typically, there is one gesture per tone unit (Kendon 1980, McClave 1991). These findings suggest that there is a mental process that coordinates speech and gesture production (Kendon 1980, Morrel-Samuels and Krauss 1992, McNeill 1992) (see *Speech Production, Psychology of; Speech Production, Neural Basis of*).

3.3.1 *Gesture and ideophones.* The exact meeting point of gesture and speech production processes is a highly debated issue (see McNeill 2000). One phenomenon that sheds some light on this issue is the fact that linguistic entities that have imagistic and emotive meaning are highly associated with iconic representational gesture. One such entity is the class of

words called 'ideophones' (also referred to as 'mimetics'). Ideophones are sound-symbolic words, for which native speakers report vivid imagery and emotion as their meaning. They include onomatopoeic words such as animal cries, but their referential domains go far beyond sound emitting entities (e.g., color, smell, wetness, density, emotion, and manner of motion).

Japanese is one of the languages in which ideophones comprise a large word class and they are commonly used. It was found that when Japanese speakers used an ideophone in narrative, it was almost always accompanied by a co-expressive iconic gesture (Kita 1997). This suggests that iconic gestures and ideophones are produced from common underlying mental representation that is imagistic in nature. In other words, iconic gestures are produced from imagistic representation that is evoked in the speaker's mind.

3.3.2 *Gesture and psycholinguistics.* Gestures that spontaneously accompany speech can be used as a 'window' into speech production processes, through which one can gain insights into the speaker's on-line conceptualization (McNeill 1992). In such a line of inquiry, representational gestures have been in the focus. Since the shape of representational gestures is not completely determined by convention, there are ample degrees of freedom left for idiosyncratic expression of the speaker's thought.

Representational gestures are created on-line by the speaker, and are typically co-expressive with the concurrent speech in the sense that they express the same object, action, event, or spatial relation as the speech, albeit from a different perspective and with a different mode of expression. As discussed above, the form of a representational gesture and its meaning are mediated by iconicity and spatial indexicality. Furthermore, ideophones, whose semantics is inherently imagistic, are almost always produced with a co-expressive iconic gesture. In short, representational gesture is imagistic, idiosyncratic, and non-linguistic; in contrast, speech is mainly propositional, and based on a conventionalized system of signs.

It has been argued that the underlying mental representation for a gesture-speech complex is produced by a single mental process involved in the verbalization of thought (Kendon 1980, McNeill 1992). McNeill stresses the significance of the semiotic duality of a gesture-speech complex. He argues that when the speaker tries to verbalize thought, he or she is engaged in two kinds of thinking: imagistic thinking and linguistic thinking. When the speaker finds a connection between idiosyncratic imagery and socially-constituted linguistic signs, a mental seed from which a gesture-speech complex develops is formed. In this process, imagery is grounded in a socially shared linguistic system, and words are con-

nected to a specific image in the speaker's mind. According to McNeill, the formation of a unified representation that integrates idiosyncratic and social sign systems is a crucial initial step in the verbalization of thought.

It has been suggested that gestures are not only the reflection of the speaker's thought, but they also facilitate verbalization of thought. There is controversy as to the exact nature of the facilitation (see McNeill 2000). Some argue that representational gestures facilitate retrieval of lexical items from the mental lexicon (Morrel-Samuels and Krauss 1992), but others argue that conceptual planning for speaking is facilitated (Alibali et al. 2000).

4. Conclusion

Gesture is a semiotic system distinct from language. At the same time, it is intimately linked to various aspects of linguistic performance. Gesture is an integral part of the organization of face-to-face conversation. Pointing gesture is an essential part of referential acts by means of spatial deixis. The production of intonation and ideophones are tightly coupled with the production of a gesture that is co-expressive. The speaker's imagistic thinking during verbalization of thought is revealed by gestures that spontaneously accompany speech. Because of the interactional, referential, and psychological unity of gesture and language, the understanding of language is not complete unless gesture is taken into account.

See also: Communication and Social Psychology; Communication: Philosophical Aspects; Motor Control; Nonverbal Communication, Social Psychology of; Verbal Communication, Psychology of

Bibliography

- Alibali M W, Kita S, Young A 2000 Gesture and the process of speech production: We think therefore we gesture. *Language and Cognitive Processes* **15**: 593–613
- Birdwhistell R L 1970 *Kinesics and Context: Essays on Body Motion Communication*. University of Pennsylvania Press, Philadelphia, PA
- Calbris G 1990 *The Semiotics of French Gestures*. Indiana University Press, Bloomington, IN
- Clark H in press Pointing and placing. In: Kita S (ed.) *Pointing: Where Language, Cognition, and Culture Meet*. L. Erlbaum Associates, Mahwah, NJ
- Efron D 1941 *Gesture and Environment*. King's Crown Press, New York
- Ekman P, Friesen W V 1969 The repertoire of nonverbal behavioral categories: Origins, usage and coding. *Semiotica* **1**: 49–98
- Goodwin M H, Goodwin C 1986 Gesture and coparticipation in the activity of searching for a word. *Semiotica* **62**(1/2): 51–75

- Kendon A 1980 Gesticulation and speech: Two aspects of the process of utterance. In: Key M R (ed.) *The Relation between Verbal and Nonverbal Communication*. Mouton, The Hague
- Kendon A 1981 Introduction: Current issues in the study of 'nonverbal communication.' In: Kendon A, Sebeek T A, Umiker-Sebeok J (eds.) *Nonverbal Communication, Interaction, and Gesture: Selections from Semiotica*. Mouton, The Hague
- Kendon A 1993 Human gesture. In: Gibson K R, Ingold T (eds.) *Tools, Language and Cognition in Human Evolution*. Cambridge University Press, Cambridge, UK
- Kendon A 1994 Do gestures communicate? A review. *Research in Language and Social Interaction* **27**: 175–200
- Kendon A 1995 Gestures as illocutionary and discourse structure markers in Southern Italian conversation. *Journal of Pragmatics* **23**: 247–79
- Kita S 1997 Two-dimensional semantic analysis of Japanese mimetics. *Linguistics* **35**: 379–415
- McClave E 1991 *Intonation and Gesture*. Unpublished dissertation, Georgetown University
- McNeill D 1992 *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press, Chicago
- McNeill D (ed.) 2000 *Language and Gesture*. Cambridge University Press, Cambridge, UK
- McNeill D, Cassell J, Levy E T 1993 Abstract deixis. *Semiotica* **95**(1/2): 5–9
- Morrel-Samuels P, Krauss R M 1992 Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **18**: 615–22
- Morris D, Collett P, Marsh P, O'Shaughnessy M 1979 *Gestures, Their Origins and Distribution*. Stein and Day, New York
- Schegloff E A 1984 On some gestures' relation to talk. In: Atkinson J M, Heritage J (eds.) *Structures of Social Action: Studies in Conversation Analysis*. Cambridge University Press, Cambridge, UK
- Sparhawk C M 1978 Contrastive-identificational features of Persian gesture. *Semiotica* **24**: 49–86
- Streeck J 1994 Gesture as communication II: The audience as co-author. *Research on Language and Social Interaction* **27**: 239–67
- Streeck J, Hartge U 1992 Previews: Gestures at the transition place. In: Auer P, di Luzio A (eds.) *The Contextualization of Language*. J. Benjamins, Amsterdam

S. Kita

Gifted and Talented Individuals: Developmental and Educational Overview

Some individuals stand out among their peers because of their remarkable abilities; they are called gifted or talented. What do these two terms mean? How many can be labeled gifted or talented? Are some youngsters much more gifted/talented than others? Are they more at risk developmentally than their average ability peers? What are their major educational needs, and how do school systems answer these needs? These are the major themes covered by this brief overview of a field traditionally named Gifted Education.

1. The Nature of Giftedness and Talent

Most published definitions of giftedness and/or talent use these two terms as synonyms (Sternberg and Davidson 1986). Yet, most scholars recognize at least implicitly two distinct manifestations of that phenomenon: (a) a natural, nondeveloped form manifested as precocious early development and presented as a 'potential', and (b) an adult well-developed form characterized by outstanding achievements in a field of human activity. Gagné (1993) proposed a developmental model that uses this distinction to create differentiated definitions for the two key concepts of giftedness and talent.

1.1 Gagné's Differentiated Model of Giftedness and Talent (DMGT)

In the DMGT, *giftedness* designates the possession and use of high natural abilities (called aptitudes) in at least one of four ability domains (see Fig. 1), so that the level of performance places the person among the top 10 percent of same age peers. *Talent* is the

demonstration of systematically developed and trained abilities in any field of human activity at a level such that the individual is among the top 10 percent of peers having had equivalent training. The six components of the DMGT bring together in a dynamic way all the recognized determinants of talent emergence in any field of human activity.

Natural abilities, which have a partial genetic origin, can be observed more directly in young children because environmental influences and systematic learning have exerted their moderating influence in a more limited way. However, they still manifest themselves in older children, even in adults, through the *ease and speed* with which individuals acquire new skills. The easier or faster the learning process, the greater the gifts. Many laypersons call these gifts 'natural talent.' Talents progressively emerge from the transformation of high aptitudes into the well-trained and systematically developed skills typical of a particular field of human activity. Thus, natural abilities act as the raw materials or *constituent elements* of talents: one cannot become talented without being gifted. However, the converse is not true: it is possible for outstanding natural abilities not to be translated

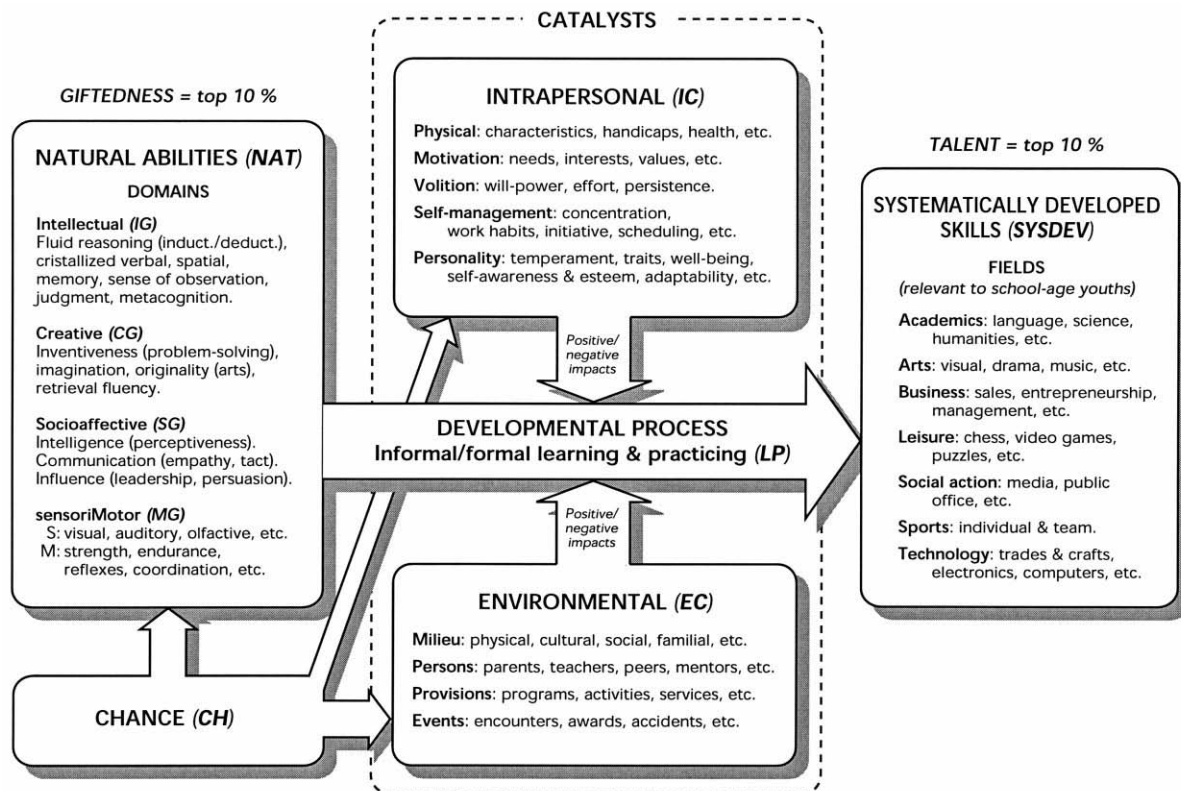


Figure 1
Gagné's differentiated model of giftedness and talent

into talents, for instance when intellectually gifted children achieve poorly in school.

The process of talent development manifests itself when the individual engages in systematic *learning and practicing* (courses, music conservatory, competitive team, etc.); the higher the level of talent sought, the more intensive these activities will be. This process can be facilitated or hindered by two types of *catalysts*: *intrapersonal* and *environmental*. Among the intrapersonal catalysts, motivation and volition (or persistence, see Corno and Kanfer 1993) play a crucial role in initiating the process of talent development, guiding it, and sustaining it through obstacles, boredom, and occasional failure. Genetic predispositions to behave in certain ways (temperament), as well as acquired personality traits and attitudes, also contribute significantly to the talent development process. The environment manifests its significant impact through four agents, either physical or social. *Chance* appears in the model as a fifth causal factor; it affects the others in different ways (e.g., the randomness of the genetic endowment for abilities or temperament, the chance of having good parents, accidents, etc.). In summary, talent emerges progressively from natural gifts thanks to a complex choreography between numerous causal influences.

1.2 How Many Are There?

Giftedness and talent are *normative* concepts. Such concepts define populations that are different from the norm (e.g., obesity, poverty). Good definitions of normative concepts must specify to what extent the target population differs from the norm and what it means in terms of their prevalence. In the DMGT (Gagné 1998), the threshold for both the giftedness and talent concepts is placed at the 90th percentile. That threshold applies to *each* giftedness domain or talent field; thus, the *total* percentage of gifted and talented persons probably exceeds 30 percent. This generous minimum threshold is counterbalanced by a system of levels of giftedness and talent. Within the top 10 percent of 'mildly' gifted or talented persons, the DMGT recognizes four progressively more selective subgroups, respectively labeled 'moderately' (top 1 percent), 'highly' (top 1:1,000), 'exceptionally' (top 1:10,000), and 'extremely' (top 1:100,000).

1.3 Focus

The field of gifted education traditionally has focused on one form of giftedness (intellectual) and one form of talent (academic). There is also some theoretical interest for creative giftedness, although identification procedures rarely include creativity measures. This article will respect that tradition and redefine the gifted and talented as *intellectually gifted and*

academically talented (IGAT) students. Intellectual giftedness usually is measured with individual or group intelligence (IQ) tests, whereas academic talent is assessed through grades or standardized achievement tests.

2. Developmental Overview

Except when noted, statements in this section apply best to mildly and moderately IGAT children, who account for 99 percent of that population.

2.1 Personal and Social Development

Most gifted children manifest some form of precocity (e.g., speaking, reading, counting). The degree of precocity is moderately correlated with the level of giftedness. Within the realm of intrapersonal catalysts, the picture is frequently ambiguous. For instance, gifted students do not appear more intrinsically motivated to learn than their average peers (Shore et al. 1991). But there is a tendency for those who achieve well to manifest slightly more persistence. That slight superiority extends to most measures of personal adaptation, whether global ones, like self-administered personality tests, or more specific indicators, like scales for anxiety, depression, or suicidal ideas. IGAT students tend to manifest fewer behavior problems in school, and the incidence of drop outs and delinquents is lower (Neihart 1999). In the case of self-esteem, study results reveal no clear tendency; differences between studies could be associated with the type of educational environment.

Social adaptation refers to friendships, popularity, leadership, and so forth. Socially, gifted youths fare usually as well, sometimes slightly better, than their average peers. Their achievements will often provoke envy, but, most of the time, it is positive envy (admiration, emulation). Are exceptionally gifted individuals more at risk for social or personal adaptation because of the larger gap with their average peers? The scientific literature shows inconsistent results; some studies indicate an increase in the frequency of maladjusted individuals—who still remain a minority (at worst 20–25 percent)—others do not (Norman et al. 1999). Some of these problems could result from inadequate responses to their educational needs.

With regard to environmental catalysts, statistics show clearly that the parents of IGAT students have on average a higher socioeconomic status. According to Herrnstein's syllogism (Herrnstein and Murray 1994), both the partial genetic inheritance of intelligence and the moderate relationship between IQ and educational attainments could account in large part for that modest over-representation.

In summary, the research base tends to invalidate the 'nerdy' stereotypical image of gifted youths.

Moreover, it proclaims loudly that IGAT students differ much more among themselves than they do when compared, as a group, with average ability peers.

2.2 Educational and Life Achievements

Intellectual giftedness is synonymous with easier and faster school learning, which leads to academic talent for most of them as long as their progress is not impeded by unfavorable intrapersonal or environmental catalysts. The close relationship between IQ scores and academic achievement is probably the best documented causal relationship in education. IGAT students have much better chances of completing high school, and earning a university degree. As the level of talent increases from moderate to exceptional, the DMGT's catalysts and practice component play an increasingly important role, especially will-power (Simonton 1994). Nobel prizes, Olympic medals, or international careers in music require intense and prolonged efforts. But such high achievements are extremely rare. In fact, according to the results of Terman's famous longitudinal study of 1,400 moderately gifted students (Oden 1968), only a minority of IGAT youngsters ever achieve eminence. Other studies with shorter time span confirm that result. This is not surprising since the prevalence of eminent individuals is well below 0.1 percent of the total adult population.

3. Educational Overview

3.1 Educational Needs

Because of their high cognitive abilities, gifted children can easily understand more abstract concepts and more advanced reasoning processes; and they can do that at a much faster pace than average. It is not rare for highly IGAT students to outpace age peers academically by 3 to 5 years. For instance, the US 'talent searches' (Assouline and Lupkowski-Shoplik 1997) identify thousands of mathematically precocious students in grades 7 or 8 by administering them the *Scholastic Aptitude Test*, a test normally used in grade 12 for college admission. Unfortunately, compared with their prevalence in schools, a very small percentage of IGAT students enjoy a school environment that meets their educational needs. The school curriculum targets regular students, and US studies have shown that (a) teachers adjust the normal pace to students close to the 30th percentile, and (b) that the content of school manuals has decreased in density ('dumbing down') since the 1980s (Reis et al. 1993). The short-term consequence is recurring boredom due to lack of challenge. The long-term risks are partial loss of interest for school learning and underdeveloped emotional and strategic skills to overcome obstacles. In summary, IGAT students need a chal-

lenging school experience, both in pace and content. Responses recognized as most appropriate differ from the regular curriculum in terms of both content and format.

3.2 A Specific Curriculum: Enrichment

Curricular adaptations specific to IGAT students are grouped under the general concept of *enrichment*, which can take four major forms—the four Ds. The first, enrichment in *Density*, corresponds to a faster learning pace, allowing for more content per unit of time. It corresponds to Renzulli's 'curriculum compacting' (see Reis et al. 1993). Not only does it make the learning process more enjoyable, but it also frees time for other enrichment activities. Enrichment in *Diversity* allows IGAT students to explore subjects outside the regular curriculum of a given grade level (astronomy, genetics, museum visits, history, etc.). Enrichment in *Difficulty* introduces concepts and explanations well beyond the level planned for regular students. For instance, IGAT students will understand very rapidly the extension of the binary system to any form of non-decimal measurement system. Enrichment in *Depth* allows students to pursue a personal interest through a long-term research project. Most large-scale enrichment programs in schools combine the four Ds in varying proportions.

3.3 Appropriate Formats: Grouping and Acceleration

In theory, thanks to curricular individualization, teachers should be able to respond to the particular educational needs of every student in their heterogeneous classroom. In reality, it is rarely done. An important US survey of classroom practices (Archambault et al. 1993) revealed that little differentiation is made, except for occasional special activities when IGAT students complete their schoolwork early. That demonstrated lack of individualization confirms the imperative need for ability grouping, either part-time or full-time. Numerous studies have shown the academic and socioaffective benefits of full-time grouping (Kulik 1992), subject to one essential condition: offering these students a truly enriched curriculum. Full-time grouping is rarely used below high-school level, except in special schools. The most popular grouping format in elementary schools, called 'pull-out classes,' uses resource rooms where students gather, usually for one or two half-day periods a week. A specially trained teacher supervises enrichment activities, combining the four Ds in various ways to tailor activities to the students' interests.

The concept of school acceleration encompasses various administrative procedures that allow students either to begin earlier than normal or to progress faster through the grade levels. It includes, among others,

grade skipping, covering in one year the curriculum of two grade levels, early entrance to school, credit by examination, and radical acceleration (skipping three years or more). Because it is made possible by intensive enrichment in density, it should be considered a true form of enrichment. Acceleration is undoubtedly the most controversial of all format prototypes. The major fear is the personal and social maladjustment of younger students among older peers due to insufficient maturity. Yet, hundreds of evaluative studies have shown little factual support for that fear (Southern and Jones 1991). When school acceleration is carefully planned and implemented, with appropriate follow-up, the large majority of students adapt very well to their new academic and social environment. But, in spite of compelling scientific evidence for its effectiveness, parents and educators alike remain very reluctant to choose it.

4. Conclusion: The Politics of Gifted Education

Worldwide, IGAT students remain the most underserved of all special school populations. Compared with the millions of dollars spent by ministries of education for all forms of disabling conditions, gifted education receives almost no direct subsidies. Advocates for these children face many objections: fear of elitism, fear of increasing the gap between the highest and lowest achievers, the need to prioritize those at risk, feelings that intellectually gifted students will succeed anyway. The accusation of elitism is based on the already mentioned observation that IGAT students come disproportionately from affluent families; but the disproportion in percentages conceals the fact that, in terms of numbers, many more IGAT students belong to families with average or below average socioeconomic status. Moreover, why should fast learners be denied an appropriate learning environment because of their social origin? As stated in the UN Bill of Rights of children, all children have a right to an education that will foster the maximum development of their abilities and personality.

See also: Creativity Training; Education for Gifted and Talented Children; Educational Institutions and Society; Educational Sociology; Giftedness, Psychology of; Higher Education; Intellectual Functioning, Assessment of

Bibliography

- Archambault Jr F X, Westberg K L, Brown S W, Hallmark B W, Emmons C L, Zhang W 1993 *Regular Classroom Practices with Gifted Students: Results of a National Survey of Classroom Teachers*. The National Research Center of the Gifted and Talented, Storrs, CT
- Assouline S G, Lupkowski-Shoplik A E 1997 Talent searches: A model for the discovery and development of academic talent. In: Colangelo N, Davis G A (eds.) *Handbook of Gifted Education*, 2nd edn. Allyn and Bacon, Boston, pp. 170–9

- Corno L, Kanfer R 1993 The role of volition in learning and performance. In: Darling-Hammond L (ed.) *Review of Research in Education*. American Educational Research Association, Washington, DC, pp. 301–41
- Gagné F 1993 Constructs and models pertaining to exceptional human abilities. In: Heller K A, Mönks F J, Passow A H (eds.) *International Handbook of Research and Development of Giftedness and Talent*. Pergamon Press, Oxford, UK, pp. 69–87
- Gagné F 1998 A proposal for subcategories within the gifted or talented populations. *Gifted Child Quarterly* 42: 87–95
- Herrnstein R J, Murray C 1994 *The Bell Curve: Intelligence and Class Structure in American Life*. Free Press, New York
- Kulik J A 1992 *An Analysis of the Research on Ability Grouping: Historical and Contemporary Perspectives*. The National Research Center of the Gifted and Talented, Storrs, CT
- Neihart M 1999 The impact of giftedness on psychological well-being: What does the empirical literature say? *Roeper Review* 22: 10–17
- Norman A, Ramsay D, Martray C, Roberts J 1999 Relationship between levels of giftedness and psychosocial adjustment. *Roeper Review* 22: 5–9
- Oden M H 1968 The fulfilment of promise: 40-year follow-up of the Terman gifted group. *Genetic Psychology Monographs* 77: 3–93
- Reis S M, Westberg K L, Kulilowich J, Caillard F, Hébert T, Plucker J, Purcell J H, Rogers J B, Smist J M 1993 *Why not Let High Ability Students Start School in January? The Curriculum Compacting Study*. The National Research Center of the Gifted and Talented, Storrs, CT
- Shore B M, Cornell D G, Robinson A, Ward V S 1991 *Recommended Practices in Gifted Education: A Critical Analysis*. Teachers College Press, New York
- Simonton D K 1994 *Greatness: Who Makes History and Why?* Guilford Press, New York
- Southern W T, Jones E D (eds.) 1991 *The Academic Acceleration of Gifted Children*. Teachers College Press, New York
- Sternberg R J, Davidson J E (eds.) 1986 *Conceptions of Giftedness*. Cambridge University Press, New York.

F. Gagné

Giftedness, Psychology of

1. Historical Perspectives

The history of giftedness is over 3000 years old. Interest in giftedness is evident in several cultures, with examples to be found in ancient Greek (Athenian) and Judeo-Christian traditions as well as in the Chinese Ming Dynasty. In the twentieth century there were fluctuations in public as well as scientific interest in this theme (for greater detail see Tannenbaum 2000).

Although the German personality psychologist William Stern promoted the research and diagnosis (identification) of giftedness in a 1916 publication and the American intelligence researcher Lewis M. Terman started the most comprehensive to date ('Genetic Studies of Giftedness' in California in 1922–1923, for which the most recent follow-up research report was presented in 1995), intensified psychological interest in

giftedness only gathered true momentum in the last 30 years of the twentieth century. This was prompted, above all, by two events and their consequences: the Threat of Sputnik in the USA and the Cultural Revolution on the mainland of China. The influential conflict of the 1960s between egalitarianism (or 'democracy') and excellence was, in the 1970s and 1980s, increasingly overshadowed by the economic demand for the activation of human resources through gifted education. Under these conditions, the number of gifted programs grew globally in the last decades of the twentieth century. Accordingly, research questions at the beginning of the twenty-first century are dominated by applied research rather than the acquisition of knowledge in the sense of basic research. Even so, experimental and field studies in the paradigms of cognitive theory, lifespan, social psychology, learning, and instructional psychology have, in recent years, won growing significance (Sternberg and Davidson 1986, Colangelo and Davis 1997, Heller et al. 1993, 2000).

2. Definition Problems and Newer Conceptions of Giftedness

The terms giftedness and talent are used synonymously in everyday language and the scientific literature (Heller et al. 1993, 2000, Colangelo and Davis 1997). Because gifted or talented individuals are a very heterogeneous set of persons, a comprehensive theory is hardly possible. Furthermore, Ziegler and Heller (2000) identify a series of empirical, ontological, and meta-theoretical difficulties with conceptions of giftedness and talent.

'For example, several conceptions of giftedness incorporate genetic influences as a significant component of the basic meaning of talent. In fact, there is a large body of research that proves that genetic influences account for some inter-individual differences. Wagner (1999) indicates ... that these genetic influences have only been proven by untrained persons.' Ericsson drew the plausible conclusion that 'the estimated performance barriers seen as genetically fixed upper boundaries can be overcome through suitable learning processes' (Ziegler and Heller 2000, p. 3).

Eysenck and Barrett (1993) view giftedness as a 'fuzzy concept' that can be defined as synonymous with general intelligence, creativity, and special (e.g., artistic or scientific) ability. Gallagher and Courtwright (1986) point out two distinct meanings based on the psychological context (with a focus on individual differences) and educational practice (with a focus on individual needs and special educational programs). Sternberg and Davidson (1986) analyzed no less than 17 different conceptions, which they divided into explicit and implicit terms. An example of an implicit conception was delivered by Tannenbaum who viewed the interaction of five factors—

general ability, special ability, nonintellective (motivational, etc.), environmental, and chance factors—as resulting in gifted/talented performance. Also Renzulli's giftedness model is counted by Sternberg and Davidson among the implicit theories, while Gagné's or the Munich psychometric model as well as Sternberg's cognitively-based triarchic model of intellectual giftedness are examples of explicit (empirical, experimentally confirmable) theories. A common basis shared by these newer conceptions is that giftedness is conceptualized as a differential construct. For greater detail see Ziegler and Heller (2000).

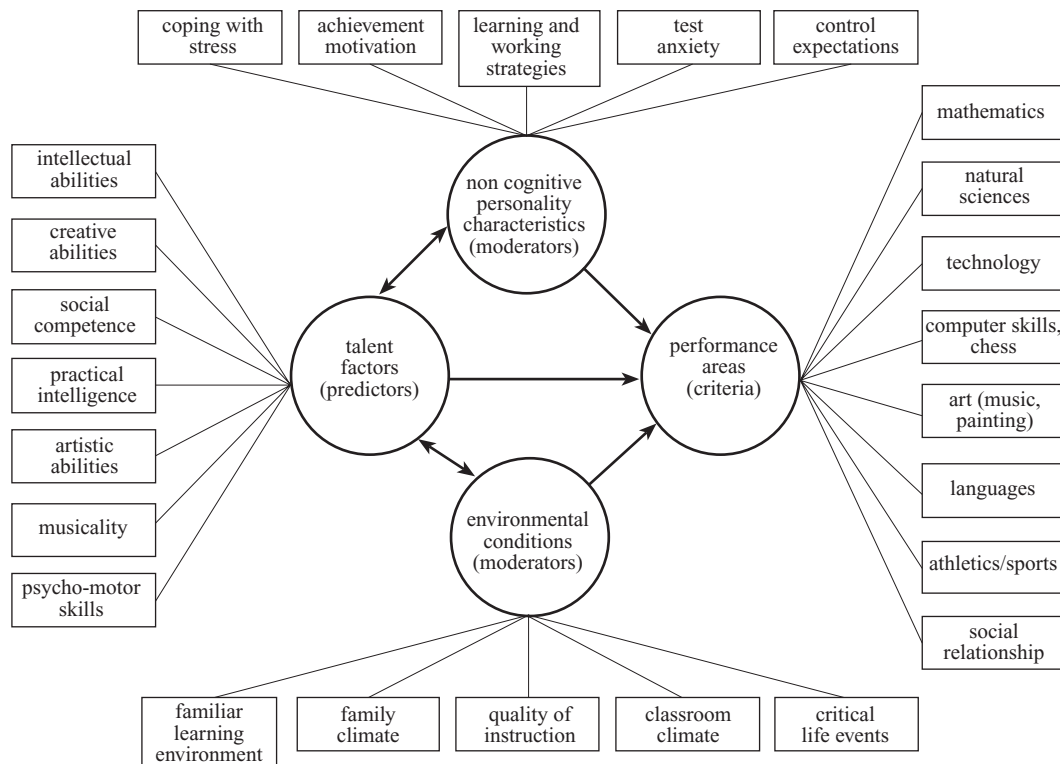
Contrary to the older IQ-based definitions (e.g., Terman defined giftedness by IQ-scores of 135 and higher), models of giftedness developed in the 1980s and 1990s represent multidimensional or typological conceptions. The following is a list of examples:

(a) The *Three Ring Model* by Renzulli (1986). According to this model, giftedness is a result of above average intelligence, creativity, and task commitment. This was expanded by Mönks to include family, school, and peer group, making a dynamic *Six Factor Model*; see Renzulli (1986), Mönks and Mason (2000).

(b) Gardner's *Theory of Multiple Intelligences*. Originally Gardner (1983), Gardner postulated seven distinct human intelligences: linguistic, logical-mathematical, spatial, bodily-kinesthetic, musical, intrapersonal, and interpersonal. Recently he has added 'naturalistic' (the ability to discern patterns in the living world) as an eighth intelligence and two further candidates (8 1/2): 'existential' and 'spiritual' intelligences, which have not yet been confirmed empirically (Ramos-Ford and Gardner 1997). This typological model synthesizes psychometric, cognitive psychological, and neuropsychological approaches to conceptualize *giftedness* (a more accurate label than 'intelligence').

(c) Gagné's *Differentiated Model of Giftedness and Talent*, the newest version of which was presented by the author (2000). Gagné differentiates between three concepts: genetically determined gifts (aptitude domains), talents (fields of exceptional performance), and catalysts (personality vs. environmental supportive factors). The gifts—intellectual, creative, socio-affective, sensorimotor, and other factors—are transformed, with the assistance of catalysts as well as learning and training processes, into domain-specific talents, e.g., in arts, sports, science, and technology.

(d) The *Munich Typological Model of Giftedness* by Heller et al. (see Heller and Feldhusen 1986, Heller 1992/2001). This multidimensional giftedness model consists of seven relatively independent ability factor groups (*predictors*), various performance domains (*criterion variables*), as well as personality (motivational, etc.) and social environmental factors which serve as moderators for the transition of individual potentials into excellent performances in various domains. High ability or giftedness is defined as an individual (intellectual, creative, etc.) potential for



Legend:

talent factors (predictors)

- intelligence (language, mathematical, technical abilities, etc.)
- creativity (linguistic, mathematical, technical, artistic, etc.)
- social competence
- practical intelligence
- artistic abilities
- music ability
- psycho-motor skills

(non-cognitive) personality characteristics (moderators)

- achievement motivation
- hope for success vs. fear of failure
- control expectations
- thirst for knowledge
- ability to deal well with stress (coping with stress)
- self-concept (general, scholastic, of talent, etc.)

environmental conditions (moderators)

- home environmental stimulation ("creative" environment)
- educational style & parental educational level
- demands on scholastic performance made at home
- social reactions to success and failure
- number of siblings and sibling position
- family climate
- quality of instruction at school
- differentiated learning and instruction
- school climate
- critical life events

performance areas (criteria)

- mathematics, computer science, chess, etc.
- natural sciences
- technology, handicraft, etc.
- languages
- music and artistic areas
- social activities, leadership, trade, etc.
- athletics/sports

Figure 1
The Munich typological model of giftedness

outstanding achievements in one—not too narrowly defined—domain (one-sided talent) or two—seldom more—domains (multi-talent). Hence giftedness is conceptualized as a multifactorized ability construct within a network of noncognitive (motivational) and social moderators as well as performance-related variables. For diagnostic purposes (identification), the differentiation between predictor, criterion, and moderator variables (in Gagné's model, so-called 'catalysts') is of particular interest (see Fig. 1). The validity of this model of giftedness has been confirmed in several enlarged studies both in Germany and in international settings (cf. Heller 1992/2001, Perleth and Heller 1994, pp. 77–114). The *Münchener Hochbegabungs-Testbatterie (MHBT)* [Munich High Abilities Test] by Heller and Perleth (eds.) which measures the various model variables (predictors, moderators) has recently been published.

(e) The *Triarchic Theory of Intellectual Giftedness* developed by Sternberg in 1985 (see Sternberg 1986, 1997) represents a process-oriented conception. In cognitive psychological research paradigms, thinking and problem-solving *processes* stand in the limelight. The triarchic theory consists of three subtheories: a) the *componential-subtheory*, whereby *metacognitive* competencies regarding the planning, monitoring, and control of actions, *performance* components in the sense of encoding and the solution of problems, and *knowledge-acquisition* components are of major significance; b) the *experiential- or two-facet-subtheory* with the aspects of *automatization* or *subroutine* vs. insight or creative thinking in the solution of (new kinds of) problems; and c) the *contextual subtheory*, which refers to sociocultural dependence of talent definitions.

Further process-oriented conceptions of giftedness are discussed by Siegler and Kotowsky in their literature survey (1986). Their two levels of giftedness correspond to those delineated by Renzulli through his division into 'schoolhouse giftedness' (intelligence) and 'creative-productive giftedness' (creativity). The diagnostic implications of these have often been confirmed, in that (multifactorial) intelligence tests can predict scholastic success, although they cannot be (exclusively) used to predict academic or career success. On the other hand, creativity tests have proved to be unreliable predictors of scholastic success, although they can often explain significantly larger portions of the variance in career success than intelligence tests. This brings us to our next question regarding the advantages and disadvantages of various research approaches.

3. Research Paradigms in the Field of Giftedness

The field of giftedness research differentiates among several types of paradigms, whereby synthetic approaches have recently been accorded a large amount

of attention. The psychometric paradigm is based on the trait and factor approach (e.g., Terman, Gagné, Gardner, Heller) already proven in the field of psychology. It focuses on the description of the phenomenology and structure of giftedness and talent (nomothetic function) and is indispensable for identification (diagnosis) of giftedness. Investigations based on the cognitive theory paradigm focus on the explanation of differences in the learning and thinking processes of highly gifted versus less gifted individuals (idiographic function). Such information about the mechanisms of cognitive processes and their individual causes provides not only new theoretical recognition but also useful contributions to the practice of gifted education and counseling. Hence both paradigms are needed, and they should be viewed in terms of how they complement one another. Furthermore, the expert-novice paradigm enriches cognitive theory in terms of the motivational effort variable 'deliberate practice' (cf. Ericsson 1996). While cognitive and metacognitive components are the focal point of cognitive componential approaches, in the expert-novice paradigm these are neglected in favor of the quantity and quality of learning and training processes. This approach has particularly proven its worth in the explanation of expertise in the domains of chess, physics, music, and athletics, but the extent to which this approach can be useful for explaining expertise in other domains must still be empirically confirmed. Even more interesting are recent attempts to combine prospective (psychometric) and retrospective (process-oriented) approaches in the form of the expert-novice paradigm. The intention here is to avoid the weaknesses inherent in the trait-oriented psychometric approach (underestimation of the influences of long-term deliberate practice) in so-called 'testing the limits' conditions (see Howe et al. 1998) and of the incremental (implicit) ability theory (according to Dweck) as well as those of the experience-oriented approach in the expert-novice paradigm (underestimation of innate talents in favor of learning processes); for recent contributions to the genetic influence vs. environmentalism controversy see Howe et al. (1998) including the 30 open peer commentaries. In the genetic and brain research-related approach the focus is on a biological explanation of giftedness and talent (cf. Thompson and Plomin, Eysenck and Barrett 2000).

Although the research paradigms sketched out here conceptualize giftedness more or less as individual resources, the following approaches concentrate on social and learning environment-related conditions. The *interactionist* theories try to balance the position between innate capacities and socialization influences with respect to the development of giftedness and excellence in various performance domains. Here lifespan developmental psychology is interested in the description and explanation of the individual's change of behavior throughout the life course (cf. Mönks and Mason 2000). *Psychosocial* and *cross-cultural*

approaches primarily analyze the external context conditions of the development of giftedness and outstanding achievements (cf. Tannenbaum 1983). From cross-national, cross-cultural, and cross-societal research approaches one can expect not only information about various social-cultural influences on the individual developmental and gifted education processes within specific social settings (*-etic* approach), but also explanations regarding the universality of gifted behavior characteristics (universalism hypothesis in the *-emic* approach). The latter function is, above all, indispensable for examining universality assumptions of many theories—here, in the field of giftedness.

4. Identifying and Programming (Gifted Education and Counseling)

Perspectives related to the application of results of research in giftedness include the identification of the gifted and the nurturing and counseling of gifts or talents. Both aim to optimize the development of personality in individual and social developmental conditions. To this end, diagnostic information is necessary, i.e., about characteristics of the individuals and their learning environments. Provided that one can conceptualize the development of giftedness as an interactive product between internal (cognitive and motivational) factors and external socialization conditions, multidimensional tests for the identification of the gifted are necessary. Correspondingly, a comprehensive diagnosis of giftedness cannot be defined by one type (for example within the framework of talent searches for specific promotional courses, scholastic programs, contests, etc.), but rather have to keep individual differences in moderation styles in mind (cf. Fig. 1). It is these moderating variables which frequently offer opportunities to provide effective counseling or intervention, for example underachievers or other at-risk groups. Here, 'at-risk groups' refer to highly gifted and talented girls, gifted handicapped, gifted underachievers, or disadvantaged and culturally different talented, because for these groups a higher risk not to be identified as gifted exists (cf. Yewchuk and Lupart 2000). In order to expand on status diagnostic methods, approaches involving *dynamic assessment* (cf. Kanevsky 2000) are available to assist in the identification of gifted underachievers. Since most norm-oriented ability or intelligence tests, including creativity tests, have proven to suffer from *ceiling effects* in application, they are inadequate for identification, particularly on the basis of IQ-scores. More suitable are multidimensional cognitive abilities tests, such as the Munich High Abilities Test (MHBT) by Heller and Perleth (cf. Heller 1992/2001), developed along the lines of the Munich Giftedness Model (Fig. 1). Here, differential talent predictors with respect to various aspects of intelligence, creativity, musicality, psychomotorics, social and practical intelligence as

well as relevant noncognitive (motivational) and social moderators are measured in order to determine outstanding performance (criterion) in various domains.

In order to reduce the bandwidth fidelity dilemma, or so-called alpha and beta errors as defined by Cronbach and Gleser, a *sequential decision strategy* can be used. Initially a *screening* (e.g., with teachers' checklists) comes up with a rough estimation for the reduction of the beta error ('oversights' of gifts and talents), after which in the second and possibly third steps, domain-specific (multidimensional) abilities tests etc. are probed in order to reduce the alpha error (in order to avoid false identifications). For more detailed explanations see Heller and Feldhusen (1986), Mönks and Heller (1994), Colangelo and Davis (1997, pt. II) or Heller et al. (1993, 2000, pt. III).

Gifted education strives for both the promotion of effective, creative learning environments versus a fit between stimulating family and scholastic settings and the individual needs for learning, e.g., striving for knowledge, exploratory urges, challenges in the sense of testing one's limits, etc. The Aptitude Treatment Interaction (ATI) Model provides the theoretical basis for these constructs. Quick information processing and accelerated cognitive and metacognitive competencies afford these children and adolescents effective self-regulated, discovery learning without a high risk probability for the acquisition of misconcepts (as among low gifted). In order to avoid understimulation and boredom among highly gifted and talented students, multifaceted scholastic differentiation measures are necessary, whereby two prototypes can be delineated: enrichment and acceleration, which can also be effectively combined. For more information regarding various instructional models and programs of nurturing the gifted and talented see Colangelo and Davis (1997, pts. III and IV), Heller et al. (1993, 2000, pts. IV and VI).

Finally, *gifted counseling* is often needed for enhancing the socialization processes in gifted education, especially with respect of the at-risk groups mentioned above. Opportunities for counseling regarding themes relevant to the highly talented differ with respect to quality and quantity, particularly with respect to gender and age of the clientele, but also regarding family and scholastic socialization as well as regional counseling offerings. Furthermore, relationships to socioeconomic family status have been confirmed. For more detailed information see Colangelo and Davis (1997, pt. V), Heller et al. (1993, 2000, pt. V) or Heller and Feldhusen (1986, Chaps. XI and XII). Counseling the gifted and talented should be an integral part of gifted education.

See also: Creativity and Cognition; Education for Gifted and Talented Children; Genetic Factors in Cognition/Intelligence; Gifted and Talented Individuals: Developmental and Educational Overview; Intelligence, Genetics of: Cognitive Abilities; Intel-

ligence, Genetics of: Heritability and Causation; Intelligence: Historical and Conceptual Perspectives; Intelligence, Prior Knowledge, and Learning

Bibliography

- Colangelo N, Davis G A (eds.) 1997 *Handbook of Gifted Education*, 2nd edn. Allyn and Bacon, Boston
- Ericsson K A (ed.) 1996 *The Road to Excellence—the Acquisition of Expert Performance in the Arts and Science, Sports and Games*. Erlbaum, Mahwah, NJ
- Eysenck H, Barrett P T 1993 Brain research related to giftedness. In: Heller K A, Mönks F J, Passow A H (eds.) 1993 *International Handbook of (Research and Development of) Giftedness and Talent*. Pergamon, Oxford, UK, pp. 115–31
- Gagné F 2000 Understanding the complex choreography of talent development through DMGT-based analysis. In: Heller K A, Mönks F J, Sternberg R J, Subotnik R F (eds.) *International Handbook of Giftedness and Talent*. Elsevier, Amsterdam, pp. 67–79
- Gallagher J J, Courtwright R D 1986 The educational definition of giftedness and its policy implications. In: Sternberg R J, Davidson J E (eds.) 1986 *Conceptions of Giftedness*. Cambridge University Press, Cambridge, UK, pp. 93–111
- Gardner H 1983 *Frames of Mind: The Theory of Multiple Intelligences*. Basic Books, New York
- Heller K A (ed.) 1992 *Hochbegabung im Kindes- und Jugendalter [High Ability in Childhood and Adolescence]*. Hogrefe, Göttingen, Germany
- Heller K A, Feldhusen J F (eds.) 1986 *Identifying and Nurturing the Gifted. An International Perspective*. Huber, Toronto, Canada
- Heller K A, Mönks F J, Passow A H (eds.) 1993 *International Handbook of (Research and Development of) Giftedness and Talent*. Pergamon, Oxford, UK
- Heller K A, Mönks F J, Sternberg R J, Subotnik R F (eds.) 2000 *International Handbook of (Research and Development of) Giftedness and Talent*, rev. edn. Elsevier, Amsterdam
- Howe M J A, Davidson J W, Sloboda J A 1998 Innate talents: Reality or myth? *Behavioral Brain Sciences* 21(3): 399–442
- Kanevsky L 2000 Dynamic assessment of gifted students. In: Heller K A, Mönks F J, Sternberg R J, Subotnik R F (eds.) *International Handbook of Giftedness and Talent*. Elsevier, Amsterdam, pp. 283–95
- Mönks F J, Heller K A 1994 Identification and programming of the gifted and talented. In: Husén T, Postlethwaite T N (eds.) *The International Encyclopedia of Education*. Pergamon, Oxford, UK, Vol. 5
- Mönks F J, Mason E J 2000 Developmental psychology and giftedness: theories and research. In: Heller K A, Mönks F J, Sternberg R J, Subotnik R F (eds.) *International Handbook of (Research and Development of) Giftedness and Talent*, rev. edn. Elsevier, Amsterdam, pp. 141–55
- Perleth C H, Heller K A 1994 The Munich longitudinal study of giftedness. In: Subotnik R F, Arnold K D (eds.) 1994 *Beyond Terman: Contemporary Longitudinal Studies of Giftedness and Talent*. Ablex, Norwood, NJ, pp. 77–114
- Ramos-Ford V, Gardner H 1997 Giftedness from a multiple intelligences perspective. In: Colangelo N, Davis G A (eds.) *Handbook of Gifted Education*, 2nd edn. Allyn and Bacon, Boston, pp. 54–66
- Renzulli J S 1986 The three-ring conception of giftedness: a developmental model for creative productivity. In: Sternberg

- R J, Davidson J E (eds.) *Conceptions of Giftedness*. Cambridge University Press, Cambridge, UK, pp. 53–92
- Siegler R S, Kotowsky K 1986 Two levels of giftedness: shall ever the twain meet? In: Sternberg R J, Davidson J E (eds.) *Conceptions of Giftedness*. Cambridge University Press, Cambridge, UK, pp. 417–35
- Sternberg R J 1986 A triarchic theory of intellectual giftedness. In: Sternberg R J, Davidson J E (eds.) *Conceptions of Giftedness*. Cambridge University Press, Cambridge, UK, pp. 223–43
- Sternberg R J 1997 In: Colangelo N, Davis G A (eds.) *Handbook of Gifted Education*, 2nd edn. Allyn and Bacon, Boston, pp. 43–53
- Sternberg R J, Davidson J E (eds.) 1986 *Conceptions of Giftedness*. Cambridge University Press, Cambridge, UK
- Subotnik R F, Arnold K D (eds.) 1994 *Beyond Terman: Contemporary Longitudinal Studies of Giftedness and Talent*. Ablex, Norwood, NJ
- Tannenbaum A J 1983 *Gifted Children: Psychological and Educational Perspectives*. Macmillan, New York
- Tannenbaum A J 2000 A history of giftedness in school and society. In: Heller K A, Mönks F J, Sternberg R J, Subotnik R F (eds.) 2000 *International Handbook of (Research and Development of) Giftedness and Talent*, rev. edn. Elsevier, Amsterdam, pp. 23–53
- Thompson L A, Plomin R 2000 Genetic tools for exploring individual differences in intelligence. In: Heller K A, Mönks F J, Sternberg R J, Subotnik R F (eds.) *International Handbook of Giftedness and Talent*. Elsevier, Amsterdam, pp. 157–64
- Yewchuk C, Lupart J 2000 Inclusive education for gifted students with disabilities. In: Heller K A, Mönks F J, Sternberg R J, Subotnik R F (eds.) *International Handbook of Giftedness and Talent*. Elsevier, Amsterdam, pp. 659–70
- Ziegler A, Heller K A 2000 Conceptions of giftedness from a meta-theoretical perspective. In: Heller K A, Mönks F J, Sternberg R J, Subotnik R F (eds.) *International Handbook of Giftedness and Talent*, rev. edn. Elsevier, Amsterdam, pp. 3–21

K. A. Heller

Global Environmental Change: Human Dimensions

1. Introduction

Global environmental change is as old as planet Earth. Physical, chemical, and biological processes have been shaping and reshaping the earth's environment since its infancy 4.5 billion years ago. In recent time, however, humankind has been one of the major driving forces of change on the planet, including climate change, stratospheric ozone depletion, deforestation, loss of biodiversity, acidification, pollution, and desertification (see, for example, Turner et al. 1990).

Research on the human dimensions of global environmental change is concerned with the human causes of change, the consequences of such changes

for individuals and societal groups, and the ways in which humans respond to the changes.

The causes include processes like industrialization and urbanization, which themselves are driven by human needs and wants, such as the needs for water, food, and shelter. The impacts include water and food shortages, floods and droughts, health risks, and conflicts about resources. Responses to these changes include adaptation, for example, through planting different crops, and even migration from areas most affected by environmental changes and mitigation, for example, by reducing emissions of gases into the atmosphere.

Individuals and teams carry out research on the human dimensions of global environmental change in many universities and research institutes worldwide.

The range of social science disciplines involved in research related to global environmental change is wide, including political scientists studying the effectiveness of environmental agreements, demographers looking at the links between population growth and environmental change, economists considering the possibilities for delinking economic growth and environmental impacts, anthropologists debating human needs and wants and global environmental change, and scholars looking at the psychological dimensions of global environmental change. In addition, the International Human Dimensions Programme on Global Environmental Change (IHDP) coordinates research at the international level (IGBP/IHDP 1999, IHDP 1999a, 1999b, 1999c).

2. The Causes of Global Environmental Change

As pointed out in the introduction, global environmental change has been taking place throughout the earth's history, our focus here is on anthropogenic causes. Changes in the global atmosphere (climate change, stratospheric ozone depletion, and acid rain) are a result of emissions of a number of gases into the atmosphere due to a range of human activities. The increasing greenhouse effect, leading to a warming of the earth's surface and lower atmosphere, is the result of the accumulation of gases including carbon dioxide, methane, nitrous oxide, and chlorofluorocarbons in the atmosphere. The Intergovernmental Panel on Climate Change (IPCC) has documented these changes. Chlorofluorocarbons are also responsible for stratospheric ozone depletion (WMO 1999). Transboundary acidic pollution is attributed to emissions of sulfur dioxide, nitrogen oxide, and other acidifying emissions (e.g., Soroos 1998).

What are the main human activities that lead to these emissions? In the case of climate change the burning of coal, oil, and natural gas, as well as deforestation and various agricultural and industrial practices are altering the composition of the atmosphere and contributing to climate change. It has been

estimated that the use of fossil fuels currently accounts for 80–85 percent of the carbon dioxide being added to the atmosphere (UNEP/WMO 1997). Fossil fuels are burned to produce energy used for transportation, manufacturing, heating, cooling, electricity generation, and other applications. Research shows that patterns and environmental impacts of fossil fuel use vary widely by country and national-level consumption varies with technology, population and other factors, such as the level of industrialization. Land use changes, e.g., clearing land for logging, ranching, and agriculture as well as for urban development, also lead to carbon dioxide emissions (e.g., Turner et al. 1990). Over the past several hundred years, deforestation and other land use changes in many countries contributed substantially to atmospheric carbon dioxide increases. Most of the net carbon dioxide emissions from deforestation are currently occurring in tropical regions. The human activities that lead to methane emissions include rice cultivation, cattle and sheep ranching, and creation of landfills. Methane is also emitted during coal mining and oil drilling and by leaky gas pipelines. It is estimated that human activities have increased the concentration of methane in the atmosphere by about 145 percent above what would be present naturally (UNEP/WMO 1997).

Chlorofluorocarbons (CFCs) have been used in aerosol spray cans, refrigeration, air conditioning, and solvents. However, the production of CFCs is being eliminated as a result of international agreements, in particular the Montreal Protocol signed in 1987 and subsequent amendments agreed to in London (1990) and Copenhagen (1992). Soroos (1998) shows the sharp drop in the production and use of CFCs and other ozone-depleting substances in the 1990s and concludes that the level of stratospheric ozone is likely to reach its lowest level around 2000 and then begin a gradual rise back to its natural level.

The accumulation of gases in the atmosphere that can give rise to climatic change, stratospheric ozone depletion, and acidification is increasingly the result of a wide range of human activities like energy production, agriculture, and transportation. These activities in turn result from the satisfaction of human needs and wants. The topic of 'needs and wants' was tackled in detail from a social science perspective by Douglas et al. (1998).

An ongoing debate with regard to the causes of global environmental change is whether population growth in less developed countries is the main cause of environmental change or whether the main factor is the high level of consumption in industrialized countries. As MacKellar et al. (1998) point out, the population-consumption debate has become the focus of the broader debate about equity and fairness, demonstrated, for example, at the UNCED in Rio in 1992. Some case studies have shown that population growth can be correlated with increasing rates of tropical deforestation, while others have shown that

population growth need not lead to increasing deforestation, if alternatives to land-clearing are provided.

Demographic research does point to some important trends that will play a role in determining future global environmental change: there will be further population growth, a further tilt towards less industrialized countries in the world distribution of population, and further population aging (MacKellar et al. 1998). Another important trend is urbanization. At the beginning of the twenty-first century, more people will live and work in the urban centers of the world than in rural areas for the first time in history (Board on Sustainable Development 1999). The 'rush to the cities' of rural populations in developing countries during the second half of the twentieth century was dramatic. The largest cities of the poorer countries are growing much faster than similar cities in the industrialized world.

In addition to the changes occurring in the global atmosphere, global environmental changes are occurring at the earth's surface. In recent years the question of biodiversity loss has been on scientific and political agendas. As defined in the UN Framework Convention on Biological Diversity, biodiversity is the diversity of species, genetic material and ecosystems or, more generally, the variability among living organisms. Human activities contribute to the loss of biodiversity in many ways. Studies of the human causes of biodiversity loss have to deal with issues such as cultural beliefs, religious teachings, social history, and property rights. In fact, of course, it is not only the issue of biodiversity that encompasses these factors. These forces also influence many other issues related to global environmental change. There is a large body of environmental ethics literature in particular that deals with such issues (see, for example, Gardner and Stern 1996).

A related issue for global environmental change is that of land use and land cover change. In addition to playing a role in biodiversity loss, land use and land cover change can affect the global climate system through emissions of greenhouse gases (carbon dioxide, methane, and nitrous oxide) as well as changes in climatic variables, such as the reflectivity of the earth's surface. Over the course of human history, land use change has contributed as much to increases of the atmospheric carbon dioxide concentration as fossil fuel combustion has (Meyer et al. 1998). The human activities that can change land cover include sedentary agriculture, shifting cultivation, navigation, water supply, forest plantations, settlements, transportation corridors, mining, and habitat and scenery preservation (Meyer et al. 1998). The study of Kasperson et al. (1995) identified world regions where misuse of land and water resources seriously threatens human wealth and well being in the near to medium term. There has also been considerable progress in recent years on questions such as the social causes of

deforestation in areas such as the Amazon River Basin and Southeast Asia and the role of social, political, and economic institutions in land-use decisions (Entwistle et al. 1998).

Recently, advances have been made in combining social science and remote sensing perspectives, techniques, and data (US National Research Council 1998). Studies have shown how remotely sensed and social data can be combined to understand human-environment linkages, especially with regard to land use and land cover changes on the regional level. The combination of social data and remotely sensed data can lead to improved understanding of the consequences of climate variability, including improving famine early warning systems and dealing with serious human health issues.

3. The Human Consequences of Global Environmental Change

Much of the research on the consequences of global environmental changes has focused on the impacts of climatic change, although there have been some studies of the impacts of stratospheric ozone depletion and acid rain.

In recent years assessments of the impacts of climate change have been carried out under the auspices of the UNEP/WMO Intergovernmental Panel on Climate Change, which produced reports in 1990 and 1996 (IPCC 1990, 1996). Given uncertainties in both the expected climatic changes (especially when considering the regional or local scale) in the twenty-first century and the calculations of socioeconomic impacts of such changes, the studies of the impacts of climate change do not produce predictions but generally generate plausible scenarios of future impacts.

The IPCC scientific assessment (UNEP/WMO 1997) estimated that the globally averaged surface temperature will increase by 1–3.5°C by the year 2100, with an associated rise of sea level of 15–95 cm. One of the topics that received wide attention after the IPCC report was the potential impact on human health. The report concluded that the projected increase in the duration and frequency of heat waves would be expected to increase mortality rates as a result of heat stress, especially where air conditioning is not available. McMichael et al. (1996) have discussed the general impacts of climate change on human health.

Many studies have looked at the potential impacts of climate change on agricultural production. The 1996 IPCC assessment concluded that it may be possible for global agricultural production to keep pace with increasing demand over the next 50–100 years if adequate adaptations are made, but that there are likely to be difficulties in some regions. The report concluded that there is likely to be an increased risk of famine, particularly in subtropical and tropical semi-arid and arid locations.

With more than half of the global human population currently living in coastal areas, future sea level rises, alterations in storm patterns, and higher storm surges could have significant effects. About 46 million people are currently at risk by flooding in coastal areas as a result of storm surges. In the absence of measures to adapt, it is estimated that even with current populations a sea level rise of 50cm would increase the number of people whose land would be at risk from serious flooding or permanent inundation to about 92 million, while a 100cm rise would increase this number to 118 million.

As the IPCC report also shows, climate change is likely to affect numerous other human activities and resources valued by humans such as river flow and water supply, transportation systems, energy demand and supply, human settlement patterns, and tourism. These impacts are the topics of research. For example, a range of studies has looked at impacts on energy supply and demand from an economic perspective and these studies are reviewed by the IPCC. Multisectoral regional assessments of the consequences of climate change have also been carried out).

The methods for assessing climate impacts have developed considerably since around 1970. As Parry and Carter (1998) point out, the efforts up until the mid-1970s focused mainly on the (one-way) 'impact' of climate on human activity, while more recently there has been more emphasis on the 'interaction' between climate and human activity by assuming that a climatic event is merely one of many processes (both societal and environmental) that can affect humans and their activities. The interaction approach also introduced adaptation into climate impact assessment, such as changes in crops and irrigation at the farm level and policy responses at the subnational, national, and international levels (Parry and Carter 1998).

Until recently, most scientific assessments of global environmental risks devoted little attention to the possibly endangered entities. Now, questions about the 'vulnerability' of social and ecological systems are emerging as a central focus of policy-driven assessments of global environmental risks in a number of arenas.

4. Human Responses to Global Environmental Change

There are basically three possible responses to global environmental change. First, humans could decide to 'do nothing,' because they don't believe scientific results, because they think that the changes will be gradual and easily accommodated through technology, or simply because they are convinced that it would cost more to do something than the benefits the environmental changes might bring. In doing nothing, they might find that they are forced to undertake

responses, such as migration from flood-prone areas, in the event of global environmental changes (forced adaptation). Second, humans can choose to 'mitigate,' for example, reduce net emissions of greenhouse gases, acidifying substances or ozone-depleting substances into the atmosphere or reduce the rate of deforestation or land-cover change. Third, humans can choose to 'adapt' to global environmental change (anticipatory adaptation).

Humans have shown great capacity to adapt to environmental changes and extreme events in the past and the assessment of the research of past research on adaptation may provide useful insights for the future. However, adaptation in a globalized world with a population of 10 billion people promises to be much more challenging than adaptation in the past.

At the international level, steps to mitigate global environmental change have been taken in recent years. The Convention on Long-Range Transboundary Air Pollution (LRTAP) and its associated protocols have been dealing with emissions reductions of acidifying substances. The Vienna Convention on the Protection of the Stratospheric Ozone Layer and the Montreal Protocol on Substances that Deplete the Ozone Layer, signed in 1987 has led to the phase-out of production of most ozone-depleting substances. At the Rio Conference (UNCED) in 1992 the Framework Convention on Climate Change and the Convention on Biodiversity were signed and have subsequently gone into force with subsequent negotiations on protocols. Clearly, it is not enough to negotiate and sign international agreements to deal with the issues of global environmental change (Underdal 1998, Weiss and Jacobson 1998). Research on policy instruments at the national and international levels of relevance to the implementation of international environmental agreements shows the strengths and limitations of strategies such as regulation, financial incentives and penalties, inducements for technical development, etc. (US National Research Council 1999). Research at the end of the twentieth century and the beginning of the twenty-first century is also showing how social institutions have succeeded and failed in the management of global environmental risks (Ostrom 1990, Young 1999).

Methods for evaluating adaptive responses are discussed by Parry and Carter (1998), who point out that many of the adaptation options are likely to be subject to legislation, influenced by prevailing social norms related to religion or custom, or constrained physically or biologically. This might encourage, restrict, or totally prohibit their implementation. However, it is clear that an assessment of possible constraints must accompany any evaluation of available options. To date, relatively little analysis has been made of the efficacy of different adaptive strategies, partly because of the very wide range of available options.

On balance, however, it seems that much more attention has been paid to the assessment of mitigation

options than to adaptation options. Serious study is still needed of the true costs of adaptation of both plants and people and the differential ability to undertake it.

A controversial topic related to the question of adaptation is the relation between global environmental change and migration. The IPCC report in 1990 suggested that the greatest effect of climate change may be on human migration, as millions of people could be displaced due to shoreline erosion, coastal flooding, and agricultural disruption. Loneragan (1998) attempted to clarify the numerous issues surrounding the linkage between environmental degradation and population displacement. He concluded that generalizations about the relationship between environmental degradation and population movement mask a great deal of the complexity, which characterizes migration decision-making. Environmental degradation and resource depletion may play a contributing role in affecting population movement, often filtered through contexts of poverty and inequity.

5. Attitudes, Behavior, and Lifestyle

While there is ample evidence that public concern about ecological problems and technological risks has increased in Western industrialized societies, as well as some other countries, since around 1980, at the same time there has been a continuing deterioration of the environment on a global scale with some dramatic regional consequences (Brand 1997). Data on the development of environmental consciousness have been and are collected in many countries, although they are difficult to compare especially because the results depend on the wording of questions and the cultural contexts within which surveys are made. As Brand (1997) documents, taken together the empirical findings point to a high, and globally growing, concern about the environment. However, a high degree of environmental consciousness does not necessarily mean that there is environmentally sound behavior. Unfortunately there are fewer comparative data on environmental behavior, and different national contexts make it difficult to make intercomparisons. Empirical findings show that the effects of environmental knowledge and consciousness on behavior are insignificant and that other factors have to be taken into account, i.e., psychological, economic, institutional, and contextual factors.

6. Integrated Assessment

An important contribution to research on the human dimensions of global environmental change has been provided by 'integrated assessment.' According to Parson (1995):

Assessment consists of gathering, synthesizing, interpreting, and communicating knowledge from various expert domains and disciplines, to help responsible policy actors think about problems or evaluate possible actions.

An 'integrated' assessment means that the information is assembled from a broader set of domains than would usually be provided by good research from a single discipline. Integrated assessment of global environmental change issues therefore requires knowledge from all relevant disciplines, natural and social sciences, and the assessment is building a bridge between the science and policy realms. Summaries of available results, for instance in the reports of the IPCC, have illustrated the strengths and weaknesses of available approaches. At the end of the twentieth century methodologies have been developed for 'participatory integrated assessment,' in which stakeholders and even lay citizens are involved in the assessment process (Kasemir et al. 1999).

7. Conclusion

There is no doubt that global environmental changes have occurred and are occurring. Social sciences are playing an increasingly important role in improving understanding of the human causes and impacts of, and responses to, such changes. In a recent report, the US National Research Council (1999) pointed to the following areas in which major advances in knowledge were achieved through human dimension research:

- (a) The nature and role of socioeconomic uncertainties in impact assessments.
- (b) The determinants of environmentally significant consumption.
- (c) The importance of vulnerability analysis in understanding the consequences of global environmental change.
- (d) The role of institutions in determining human-environment interactions.
- (e) Estimation of the costs of global environmental change and of policy response options.

The challenges remain large, especially in the area of obtaining, archiving, and using socioeconomic data and in conducting research on a range of unresolved and difficult questions that can only be tackled by international, interdisciplinary approaches.

See also: Air Pollution; Climate Impacts; Consumption, Environmentally Significant; Ecological Economics; Environmental Cognition, Perception, and Attitudes; Environmental Policy; Protection and Regulation; Environmental Security; Environmentalism: Preservation and Conservation; Industrial Ecology; International and Transboundary Accords, Environmental; Land Use and Cover Change; Water Pollution: Marine; Water Pollution: Terrestrial and Groundwater

Bibliography

- Board on Sustainable Development 1999 *Our Common Journey: A Transition Toward Sustainability*. National Academy Press, Washington, DC
- Brand K W 1997 Environmental consciousness and behaviour: The greening of lifestyles. In: Redclift M, Woodgate G (eds.) *The International Handbook of Environmental Sociology*. Edward Elgar, Cheltenham, UK
- Douglas M, Gasper D, Ney S, Thompson M 1998 Human needs and wants. In: Rayner S, Malone E L (eds.) *Human Choice and Climate Change*. Battelle Press, Columbus, OH
- Entwistle B, Rindfuss R R, Walsh S J 1998 Land use/land cover and population dynamics, Nang Rong, Thailand. In: Liverman D M, Moran E F, Ruidfuss R R, Stern P C (eds.) *People and Pixels: Linking Remote Sensing and Social Science*. National Academy Press, Washington, DC
- Gardner G T, Stern P C 1996 *Environmental Problems and Human Behaviour*. Allyn and Bacon, Boston
- IGBP/IHDP 1999 *Land-Use and Land-Cover Change: Implementation Strategy*. IHDP Report No. 10. International Human Dimensions Programme on Global Environmental Change, Bonn, Germany
- IHDP 1999a *Institutional Dimensions of Global Environmental Change: Science Plan*. IHDP Report No. 9. International Human Dimensions Programme on Global Environmental Change, Bonn, Germany
- IHDP 1999b *Global Environmental Change and Human Security: Science Plan*. IHDP Report No. 11. International Human Dimensions Programme on Global Environmental Change, Bonn, Germany
- IHDP 1999c *Industrial Transformation: Science Plan*. IHDP Report No. 12. International Human Dimensions Programme on Global Environmental Change, Bonn, Germany
- Intergovernmental Panel on Climate Change (IPCC) 1990 *Climate Change: The IPCC Scientific Assessment*. Cambridge University Press, Cambridge, UK
- IPCC 1996 *Second Assessment Report of the IPCC*. Cambridge University Press, Cambridge, UK, 3 Vols.
- Kasemir B, Schibli D, Stoll S, Jaeger C C 1999 Involving the public in climate and energy decisions. *Environment* 42(3): 32–42
- Kasperson J X, Kasperson R E, Turner II B L (eds.) 1995 *Regions at Risk: International Comparisons of Threatened Environments*. UNU Press, Tokyo
- Loneragan S 1998 *The Role of Environmental Degradation in Population Displacement*. Global Environmental Change and Human Security Project Research Report 1. IHDP, Bonn, Germany
- MacKellar F L, Lutz W, McMichael A J, Suhrke A 1998 Population and climate change. In: Rayner S, Malone E L (eds.) *Human Choice and Climate Change*. Battelle Press, Columbus, OH
- McMichael A J, Haines A, Sloof R, Kovats S 1996 *Climate Change and Human Health*. World Health Organisation, Geneva, Switzerland
- Meyer W B, Adger W N, Brown K, Graetz D, Gleick P, Richards J F, Maghalaes A 1998 Land and water use. In: Rayner S, Malone E L (eds.) *Human Choice and Climate Change*. Battelle Press, Columbus, OH
- Ostrom E 1990 *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, New York
- Parry M, Carter T 1998 *Climate Impact and Adaptation Assessment: A Guide to the IPCC Approach*. Earthscan, London, UK
- Parson E A 1995 Integrated assessment and environmental policy making: In pursuit of usefulness. *Energy Policy* 23(4/5): 463–75
- Soroos M S 1998 The thin blue line: Preserving the atmosphere as a global commons. *Environment* 40(2): 6–35
- Turner B L, Clark W C, Kates R W, Richards J F, Mathews J T, Meyer W B 1990 *The Earth as Transformed by Human Action: Global and Regional Changes in the Biosphere over the past 300 years*. Cambridge University Press, Cambridge, UK
- Underdal A (ed.) 1998 *The Politics of International Environmental Management*. Kluwer Dordrecht, The Netherlands
- UNEP/WMO 1997 *Common Questions About Climate Change*. United Nations Environment Programme, Nairobi, Kenya
- US National Research Council 1998 Liverman D, Moran E F, Rindfuss R R, Stern P C (eds.) *People and Pixels: Linking Remote Sensing and Social Science*. National Academy Press, Washington, DC
- US National Research Council 1999 *Global Environmental Change: Research Pathways for the Next Decade*. National Academy Press, Washington, DC
- Weiss E B, Jacobson H (eds.) 1998 *Engaging Countries: Strengthening Compliance with International Environmental Accords*. MIT Press, Cambridge, MA
- WMO 1999 *Scientific Assessment of Ozone Depletion: 1998*. WMO Global Ozone Research and Monitoring Project, Report No. 44. World Meteorological Organization, Geneva, Switzerland
- Young O 1999 *The Effectiveness of International Environmental Regimes: Causal Connections and Behavioural Mechanisms*. MIT Press, Cambridge, MA

J. Jäger

Global Governance

Global governance is the combination of intentional and patterned human interactions that regulate action worldwide for the common good. It is a purposeful order that emerges from institutions, processes, norms, formal agreements, and informal mechanisms that provide a field of action for human activities. Governance is based on shared expectations, as well as on intentionally designed institutions and mechanisms (Rosenau and Czempiel 1992).

In the context of national government, the authority of public institutions lies ultimately in the government's monopoly of the legitimate use of force. In the global context, no such authority exists; there is no world governmental body that enjoys a monopoly of the legitimate use of force. Yet commerce among countries and firms expands, information is communicated across national boundaries in an orderly fashion, and expectations about rule implementation are shared through transnational networks. Cooperative action is based on rights and rules that are enforced through a combination of financial and

moral incentives as well as through the use of threat or force. How this happens is the domain of global governance.

Global governance is distinguished from international governance. International refers to relations among states. Global encompasses activity at the international, transnational, and regional levels. For some, the term also includes a recognition of the effects of local activity on the global, such as individual auto emissions of carbon dioxide on the atmosphere, and global effects on the local, such as transnational capital flows on the health of subnational economies. For most, global refers to those activities in the public and private sectors that transcend national boundaries.

1. Governance in the Context of Rapid Globalization

Heightened attention to concepts of global governance at the end of the twentieth and beginning of the twenty-first century comes as a result of increasing consciousness of the interconnectedness of human activity on the planet. Basic discoveries about the interdependence of species and habitat for survival, about the effects of human activity on the biosphere, about the volatility associated with rapid capital flows across national borders, and new inventions that permit people to talk to one another around the globe contribute to a sense of the globalization of human society. It is in this context, and in the growing knowledge of the harm that such global processes may bring about, that the idea of global governance takes on greater importance. The problems of shaping societies so that their activities do not produce irreversible harm to the biosphere and to the sustainability of human species prompt increased attention to the possibilities of improved global governance for the public good.

2. Perspectives on Global Governance

Ordered and patterned interactions on a global scale are the result of international agreements, of harmonization of laws between countries, of transnational cooperation among functional agencies, of self-enforcing contracts between private firms, and of standard operating procedures encoding daily activities of governmental and nongovernmental organizations. Taken altogether, this array of formal legal arrangements, informal codes of conduct, and habitual practices based on shared expectations about legitimate action, constitute global governance.

So, for example, the United Nations is a formal international arrangement based on legal agreements among states and is considered to be part of global governance. It is one of many formal institutions,

however. Others would include the World Trade Organization, the North Atlantic Treaty Organization, the Organization of American States, as well as other regionally based treaty organizations. In addition, transnational cooperative actions among national law enforcement agencies also are part of global governance. Combating illegal criminal activity, for example, is accomplished through cooperation among agencies in apprehending offenders. In civil law, courts set precedents and decide cases based on case law from other national judicial systems. And in commercial law, private mediation is used by business firms from different countries to settle disputes (Slaughter 1997).

Many of these informal institutional arrangements are supported by networks among individuals and agencies. Cross-national professional relationships provide opportunities to learn from the experience of judges, policy officials, or executives from other countries. The annual World Economic Forum at Davos, primarily for business people, is just one rather visible example of informal forums. Through these and other meetings, professionals and executives begin to develop shared goals and values that can contribute to common standards and rules for conducting business, protecting human rights, conserving the environment, or implementing humanitarian missions.

The development of networks has accelerated with the introduction of high-speed electronic communications technology. Even before the advances in this area, however, increased travel and telephone communication contributed to a worldwide human rights movement, a global environmental movement, an international women's movement, the development of transnational corporations, and of international professional societies. Such networks have played a distinctive role in helping to establish common standards in a number of fields. In the case of human rights, for example, the diffusion of norms about the rights of prisoners came about through the use of international media to publicize cases of abuse. Such public 'shaming' of governments is based on a shared norm about the value of individuals and the need to protect their rights. Some of these values have been codified in formal national and international laws, while others are observed because they enhance the public reputations and legitimacy of governments or of businesses.

In sum, the concept of global governance builds on behavioral understandings of law and society. These perspectives suggest that norms of appropriate behavior emerge from daily interactions and the development of community consensus about what is fair and just, and is then codified in law. Law is a dynamic and changing reflection of shared standards. Implementation and compliance are understood as rather more voluntary than enforced, although the threat of sanctions—either criminal or civil—is recognized as an underlying motivating force for compliant action.

The simple example of stop signs illustrates the dual character of rules. Individuals obey stop signs because they may be afraid of getting caught breaking the law; but perhaps more important for some, they obey the stop sign because they wish to avoid being hit by another car. Both the threat of sanctions and an interest in avoiding harm are thought to be at work. In another example—of the private adjudication of transnational business disputes—firms wish to preserve flexibility and avoid undue government interference, which they may view as harmful. So they are willing to comply with the judgement of an agreed-upon private mediator rather than insist on taking the issue to a public institution for settlement.

3. Elements of Global Governance

It is possible to identify fundamental elements that constitute the building blocks of contemporary global governance. The most prominent of these are standards, norms, and states.

3.1 Standards

One means of coordination among private actors is the setting of standards. For example, the standardization of time through the use of Greenwich Mean Time permits coordination of communications, airline schedules, television broadcasting, and a myriad of daily transactions that we now take for granted. Rules of international navigation and airline procedures ensure the safety of air travel across national borders. A standard railroad gauge eases transborder crossings. Product standardization rationalizes the multinational manufacture of automobiles, televisions, refrigerators, and so forth. We begin to see how useful these standards are when we do not have them. The lack of standardization in electrical plug and socket connections across continents is an example of a lack of common standards.

In general, standards are accepted because they are thought to be voluntary and are arrived at through reason, that is, through the exercise of professional expertise, in which fundamentally equal actors reach decisions through rational deliberations that are open to interested parties. In practice, the process of standard setting by bodies such as ISO 9000 is completed by those with expertise in a field (in accounting or engineering, for example), representing users and pledged to serving the public interest. Efficiency in operations is essential to the acceptance of standards. Furthermore, the legitimacy of standards is based on the desirable consequences of the rule, which allows no exceptions or variations. To the extent that efficiency is a value in itself, and to the extent that professions embody and transmit a set of values, standard setting is not entirely value-neutral.

Yet, in comparison with hierarchy, markets, or norm-based communities, standardization is a minimal form of coordination.

3.2 Norms

Norms are internalized rules based on commonly held values. In the international security area, non-use of nuclear weapons is cited as a shared norm, as well as a major achievement. The norm is based in part on a shared revulsion at the devastation to human life and society that is inflicted indiscriminately in a nuclear explosion. The success of the human rights movement is based on publicizing human rights abuses perpetrated by national governments. To the extent that ‘shaming’ governments is a successful strategy it is because a norm is widely shared about the rules for appropriate treatment of individuals by their governments.

Communication of common values through informal and professional networks is critical to the development of shared norms. It is through informal networks, those that include representatives from international institutions, national governments, and nongovernmental voluntary associations, and in some cases from private business, that common understandings and values are established. These commonly held values in turn, contribute to the management of environmental issues, human rights abuses, humanitarian relief, and security relations. Some analysts view these networks as an essential component of global governance, providing a way of governing without government (Reinicke 1998).

3.3 States

The organization of nation-states in the twentieth century has included a dependence on and growth of bureaucracy, or hierarchy, as a means of governing. The growth of the modern state has included regulation of markets, of educational institutions, of environmental conditions, of criminal activity—in fact of nearly every aspect of daily life as a means of coordinating action for the common good. While some of this regulation has come through standard setting and norms development, the characteristic mode of hierarchical government is through directives. These directives are enforced through the use and threat of punitive sanctions, including imprisonment and fines.

As noted earlier, the enforcement power of the state is based on its monopoly over the legitimate use of force. This monopoly, in turn, is thought to be based on a notion of a social contract, whereby citizens give up some rights to a sovereign government, in return for protection and through the establishment and preservation of social order. The central problem for

theorists and practitioners of global governance is that no world government exists, yet transnational activity generates needs for coordination as well as problems of collective action. We turn next to a discussion of methods of coordination, problem-solving, and conflict management currently in use—methods that constitute global governance in practice.

4. *Methods of Global Governance*

4.1 *Harmonization of Laws*

The implementation of international agreements is accomplished through the adoption of national laws to bring local and national practices into compliance with international rules. For example, carrying out the provisions of the Montreal Protocol to protect the atmospheric ozone layer is done through legislative action in each country to bring manufacturing practices into conformity with the Protocol's agreed directives to reduce production and release of chlorofluorocarbons (CFCs).

Legislative action is one path to harmonization of laws. Another is through the courts and judicial systems of countries. A body of transnational law is developing informally in which judges in one country are basing their decisions on precedents and case law from other countries (Slaughter 1997). This is taking place in areas of civil and commercial law, where standard rules help to smooth business and legal transactions. In the area of human rights law courts are also citing cases in other national systems, in addition to the United Nation's Universal Declaration of Human Rights, to lend credibility to decisions about the protection of civil and political rights in their own countries.

4.2 *International Regimes*

International regimes are defined as a 'set of implicit or explicit principles, norms, rules, and decision-making procedures around which actors' expectations converge' (Krasner 1983). The effectiveness of a regime rests on the operations of institutions, organizations, governments, and international bodies that share a set of principles, rules, and norms in a particular area of international action. Although regimes include formal treaties and national law, they also rely on informal norms and networks to develop and enforce standard behavior in an area of global policy.

For example, regimes to control the development and spread of nuclear weapons had been relatively successful in the context of the Cold War. They were critical to controlling the US–Soviet arms race, to preventing nuclear war, and to limiting the spread of nuclear weapons technology and weapons. These regimes had their basis in a number of formal treaties

and international agreements, such as the Anti-ballistic Missile Treaty of 1972 and the Nuclear Non-proliferation Treaty of 1968, as well as some enforcement powers through the United Nation's International Atomic Energy Agency. For compliance, however, regimes relied on a number of formal and informal surveillance techniques and the trust that emerged from shared norms about the dangers of nuclear weapons, their development, production, and proliferation (de Nevers 1999). A decade after the end of the Cold War, we are beginning to see how much trust the USA and the Soviet Union had in one another, and in the shared norm of the non-use of nuclear weapons, as the arms control regime is being tested by the actions of several parties to these regimes.

4.3 *Global Policy Issue Networks*

Development of global policy and implementation of common laws also depends on informal networks of organizations and individuals. These networks are composed of individuals who share an interest in providing a public good that may not be provided by any single nation-state. Through communication networks, sometimes built around professional associations and scientific bodies, these transnational action networks (Keck and Sikkink 1998) or epistemic communities (Haas et al. 1993) provide information, documentation, and alternative policy solutions to global policy problems.

Some of the most visible networks have formed since the mid-1970s on human rights, environmental protection, and women's issues. In each of these areas, lawyers, scientists, and political and cultural leaders have found counterparts in other countries to form worldwide movements that have publicized the harm to people and the environment brought about by current national policies and practices. In the case of women's issues, the role of nongovernmental organizations at the official United Nations conferences on women and on population has been dramatic. The adoption of women's empowerment as a major path to fertility reduction, rather than an emphasis on birth control, at the population conference in Cairo in 1995 demonstrated the influence and effectiveness of issue networks in shaping global policy.

4.4 *Hybrid Institutions*

As transnational activity has increased, as non-governmental actors have become more prominent, and as national governments have devolved authority to private sector bodies, new institutional arrangements have developed to manage global relations. These arrangements are referred to as hybrid institutions or public–private partnerships because they mix

private sector organizations and a variety of incentive systems with governmental agencies to achieve public goods. They will often combine standard rules, private market arrangements, nongovernmental norm-based organizations, and government agencies—both national and international.

Some of the most inventive have emerged in the area of environmental protection. The certification of sustainably harvested forest products is just one example of a hybrid global governance process. Presented as an alternative to governmental regulation of the timber industry, environmental nongovernmental organizations have worked with major private enterprises in the forest products industry to convince them of the economic benefits of sustainable forest practices. They have demonstrated that consumers will pay premium prices for furniture and other lumber products that are sustainably harvested. They have shown that profits can be achieved over a longer period of time, and that reducing dependence on clear-cutting provides livelihoods for those people, mostly in developing countries, dependent on forests for their survival.

The major method of ensuring compliance with sustainable practices is through certification of wood products. Private firms participate in a council that includes producers, scientists, and other independent interests. This council, in turn, inspects production operations and certifies that manufactured products are derived from sustainably harvested lumber. The manufacturer can use that certification to advertise products and, in most cases, to charge higher prices to consumers. This process takes place without government intervention; it includes market incentives and relies on professional expertise and standardization to enforce rules that will contribute to the protection of natural resources—to the conservation of a public good.

Another example of hybrid institutions can be found in the area of prosecuting transnational criminal activity. Local and national law enforcement agencies, federal bureaus, banking regulation bodies, international police forces, and banking and financial institutions are developing transnational relationships and standard operating procedures to root out and prosecute international criminals. Organized crime has become relatively border-free, especially as liberalization of economic relations has proceeded; yet law enforcement remains within fixed territorial legal frameworks. To deal with this challenge, national law enforcement agencies have developed cooperative relationships with their counterparts in other countries to share information and to apprehend criminals. For example, US and Russian law enforcement agencies—the Federal Bureau of Investigation and the FSB (Federal Security Service, former KGB)—have set up cooperative relations to investigate and help prosecute corruption in banking and industry in their two countries.

To investigate and report money-laundering activity, the US Treasury Department has set up a Financial Crimes Enforcement Network. This network includes US Treasury Department bureaucrats, law enforcement officials, banking regulators, and private banking and financial institutions around the world. Through a series of agreements, government agencies and private sector organizations have established shared authority in identifying suspicious financial transactions. Banks have opened up their accounts to greater scrutiny and, in exchange, government authorities allow banking operations themselves to determine when activity is suspicious, and thus susceptible to further investigation by law enforcement agencies. Standards have been set about what constitutes suspicious behavior, and data about transactions is shared electronically between the private and public sectors (Reinicke 1998, pp. 135–72).

5. *Issues in Global Governance*

While new modes of coordinating transborder activity are emerging that hold the promise of effective global governance, a number of theoretical and normative challenges have been identified. Two are addressed here: the legitimacy of authority and democratic values.

5.1 *Legitimacy and Authority*

Successful governance requires the consent and compliance of those whose activity is being governed, that is, practices and institutions that constrain actions must be seen as legitimate for them to be effective. In contemporary conceptions of governance, this aspect remains problematic. In contrast to theories of governance at local and national levels, the idea of a social contract between citizens and institutions of global governance has not been developed sufficiently to be credible. Such social contracts have depended on theories of participation and representation by citizens in decision-making processes, few of which figure centrally in global governance. Other sources of legitimacy, such as claims of monarchs to ‘divine right,’ or to hereditary rights, are not recognized in twenty-first century conceptions of global governance. Finally, however, legitimacy claims may also be based on what works well, what is efficient, or what is reasonable; contemporary justifications for global governance mechanisms are often based on efficiency warrants.

Issues of authority and legitimacy become even more vexed with the increased participation of nongovernmental organizations in issue networks and hybrid institutions. While the authority of governments and the legitimacy of private firms are based, in some measure, on implicit contracts with citizens or consumers, respectively, the authority and legitimacy

of nongovernmental organizations is based on their claims of acting in the public good—as they define it. Whether by protecting individuals from the human rights abuses of their own governments, or conserving biodiversity in the face of extractive private industry, or saving women from customary rituals that cause physical harm, nongovernmental actors believe they are upholding standards that are in the long-term interests of humanity. These organizations, however, are not assembled through an election process or any other democratic procedure. They are self-appointed and claim to be acting on the basis of values that serve a global public interest. Without an authoritative institution that provides a process for establishing common values and agreement on what constitutes the public good, issues of authority and legitimacy in global governance will most likely remain unresolved.

5.2 Democratic Values and Global Governance

While some definitions of global governance emphasize the neutral management of human affairs, there is a strong normative component to any discussion of the term, in addition to an underlying assumption about who governs and upon whose standards. Some of the normative implications are acknowledged, but others need to be surfaced and examined before agreement can be reached on a standard definition.

In its current conception, global governance implies democratic governance. Most authors are careful to make the distinction between the term ‘world government’—that is, a set of procedures and mechanisms based on a single authority with power to enforce laws, by force, if necessary—and the notion of governance. This conception also recognizes that a greater number and range of organizations, individuals, institutions, and governments participate in decisions that have consequences beyond national borders (Mathews 1997). The informal and pluralistic nature of global governance both recognizes and requires the contribution of many of these nongovernmental actors to policymaking. In its own terms, then, global governance incorporates democratic values including those of equitable participation and the use of representative institutions for resolving conflicts. Yet the reliance on scientific and professional bodies to set standards, rules, and procedures, on bureaucratic agencies to implement policies, and on voluntary organizations to monitor compliance, raises questions about the compatibility of democratic values and the concept of global governance.

See also: Area and International Studies in the United States; Intellectual Trends; Global History: Universal and World; Globality; Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity; International and Transboundary Accords, Environmental; International Justice: Philosophical Aspects; International Organization

Bibliography

- Allott P 1990 *Eunomia: New Order for a New World*. Oxford University Press, Oxford, UK
- Brunsson N, Jacobsson B (eds.) 1995 *Standardising*. Nerenius and Santerus, Stockholm
- Commission on Global Governance 1995 *Our Global Neighborhood*. Oxford University Press, Oxford, UK
- de Nevers R 1999 *Regimes as Mechanisms for Global Governance*. Rockefeller Brothers Fund, New York
- Graham C 1998 *Private Markets for Public Goods*. Brookings Institution Press, Washington, DC
- Haas P, Keohane R O, Levy M A (eds.) 1993 *Institutions for the Earth, Sources of Effective International Environmental Protection*. MIT Press, Cambridge, MA
- Held D 1995 *Democracy and the Global Order: From the Modern State to Cosmopolitan Governance*. Stanford University Press, Stanford, CA
- Kaul I, Grunberg I, Stern M (eds.) 1999 *Global Public Works: International Cooperation in the 21st Century*. Oxford University Press, New York
- Keck M, Sikkink K 1998 *Activists Beyond Borders: Advocacy Networks in International Politics*. Cornell University Press, Ithaca, NY
- Krasner S 1983 *International Regimes*. Cornell University Press, Ithaca, NY
- Mathews J 1997 Power shift. *Foreign Affairs* 76(1): 50–66
- Reinicke W 1998 *Global Public Policy: Governing Without Government?* Brookings Institution Press, Washington, DC
- Rosenau J, Czempel E (eds.) 1992 *Governance Without Government: Order and Change in World Politics*. Cambridge University Press, Cambridge, UK
- Slaughter A M 1997 The real new world order. *Foreign Affairs* 76(5): 183–97
- Young O 1989 *International Cooperation: Building Regimes for Natural Resources and the Environment*. Cornell University Press, Ithaca, NY

K. Benedict

Global History: Universal and World

1. Historiographical Traditions

Global histories have been written since Herodotus (495–425 BCE), whose cosmopolitan concerns were commended by Cicero but deplored by Thucydides—who opined that Greeks had no business investigating alien mythologies, religions, customs, and traditions. History writing, Thucydides suggested, should be modeled upon his own study of the Peloponnesian Wars—sharper in focus, shorter in time horizons, based upon verified facts and prescriptive in purpose. Fortunately Herodotus ignored his ‘Eurocentric’ advice and ranged widely beyond the Hellenic world to include Egypt, India, Babylonia, Arabia, and Persia in his histories in order ‘to preserve the memory of the past by placing on record the astonishing achievements of both our own and of the Asiatic peoples.’ He used

oral testimony, archeological remains as well as written sources and imposed some chronology and order on streams of events that had occurred on several continents and over long spans of time. His (still popular) book ends in 'European triumphalism.' Its story turns into celebration of the victories of the Greek *polis* (especially Athens) over the Persian Empire, which Herodotus represented as a conflict between occident and orient, despotism and freedom.

Global historians applaud Herodotus for his reflexive interest in barbarian virtues and regret the long hiatus that occurred in the writing of secular world histories until Voltaire and his generation carried the project forward again during those brief decades of Enlightenment, before the outbreak of the French Revolution. They are puzzled, however, to observe that with the exception of a few Stoics (Diodorus, Polybius, and Dionysius) and a handful of geographers and ethnographers (Strabo, Ptolemy, and Pliny), histories of the world did not attract more scholars writing during Graeco-Roman times. Although they lived in empires that maintained contact with an array of cultures (African, Arabian, Persian, Indian, and Chinese) historians of the classical era remained overwhelmingly concerned with Europe. Their histories (e.g., Tacitus) deal mainly with the politics and scandals of Rome, or (e.g., Thucydides) with wars among Greeks. Their equally provincial oriental contemporaries, residing in that celestial empire on the other side of the world, asserted with quintessentially Chinese superiority that they remained unimpressed with anything the Romans had to impart, or to trade for, China's silks, ceramics, medicines, perfumes, and spices.

In Europe, steps in an ecumenical direction occurred when Christian and Jewish historians started their narratives with the Creation and set out to ensure that Greeks and Romans should not be left (in the words of Josephus) 'as the arrogant possessors of antiquity.' That was why Orosius (a student of St. Augustine) proclaimed that 'The Roman Empire arose in the west but was nourished in the east.' Narratives in Christian historiography is dominated by chronicles about the evolution of mankind over stages of history, no longer bounded by the frontiers or cultures of the Greco-Roman world. In 1158 Bishop Otto of Freising recognized that 'all human power and learning had its origins in the East.' Medieval historians included Oriental and African civilizations right down to the Reformation when Catholic and Protestant intellectuals began to look for a usable past to help princes to integrate populations and secure the frontiers required for the formation of nation states. Even then both parties to Europe's conflict over religion and national identities remained aware of places, peoples and cultures not merely outside the borders of their enclosed and ideologically encompassing polities, but beyond Christendom itself, and that is hardly surprising.

Residing on an underdeveloped promontory on the western edge of the great Eurasian landmass, with ice to the north, uncharted oceans and territory to the west, and the hostile power of Islamic communities to the south and east, Europeans had not secured their frontiers, religions, cultures, and technological superiority over other civilizations, long before the eighteenth century. It took centuries to clear the Iberian peninsular of Arab power. Moslem fleets threatened European shipping in the Mediterranean for another two centuries after the surrender of Grenada. Turkish armies crossed into the Balkans in 1350, took Constantinople in 1453 and menaced Vienna as late as 1683. For nearly a thousand years after the fall of Rome, Christendom braced its defences, imported useful knowledge, shaped its collective identity and composed ideologically charged histories in the context of conflicts (and encounters) with the powerful civilization of Islam on all its frontiers. Memoirs, travelogs, diplomatic correspondence and investigations into Arabic medicine and astronomy, together with commercial intelligence on societies under Ottoman dominion, flowed from east to west. Much of this knowledge came through Genoa, Florence and Venice, along with spices, herbs, sugar, botanical drugs, jewels, chinaware, silks, cottons, and metal goods, imported through the Middle East, from India and China. The expense of procuring and transporting luxuries moving slowly, over long distances and subject to extortionate charges for protection along the way, prompted Europeans to seek ways to surmount the barriers to commerce with South Asia and the Far East by their Islamic enemies.

Shortly after the voyages of discovery Europeans broke out of their political, commercial and mental encirclement by Islam. Over the sixteenth century flows of information into Europe widened to include the Americas, and increased exponentially as a result of the establishment of regular and direct trade and contacts by sea with India, China, South East Asia and Africa. Yet nearly two and a half centuries elapsed before Europeans produced a secular and enlightened school of world historians. Its godfather was Voltaire, who composed a history concerned with 'only that which is worth knowing; the spirit, the customs, the practices of the principal nations based on facts which one cannot ignore.' Voltaire and his followers in France, the Netherlands, Scotland, Naples, Germany and other parts of Europe, deliberately constructed histories of their expanding world in ways which they conceived as a departure from the providential narratives of their clerical predecessors and the 'nationalism' of historians writing for princes and their obedient subjects. That aspiration was propitious. As a result of regular contacts and commerce the trickle of information about peoples, geographies, technologies, politics, warfare, customs, manners, and the import of all sorts of commodities and artefacts turned into a river of knowledge about places, populations, societies, and

economies beyond Europe. For radical intellectuals China became an alternative model, favorably contrasted with Europe's political systems, social arrangements, and religious beliefs. European histories became not only wider but deeper as scholarship took over from propaganda, sycophancy, fantasy, religion, and national ideologies.

For something like half a century before the French Revolution and the takeover of India and South East Asia, European historians resituated Europe's and Christendom's place in world history by including more ancient as well as classical Mediterranean civilizations and by widening their gaze to take China, India and Japan as well as the newly discovered Americas into account. Europe's new school of world historians dealt with the natural histories, religions and ethnographies of most parts of an increasingly interrelated universe. Favorable interpretations of oriental polities, societies, cultures, and religions by philosophes and physiocrats did not go unchallenged. Bernier and Leibniz, Montesquieu, Defoe, Winckelmann, and Hume remained resolutely unimpressed with the Chinese empire. Nevertheless historians of that era engaged in debates about cultures, geographies, and spans of time that their modern counterparts (confronted with libraries of knowledge and professional interpretations), can only envy.

Alas, the ironic reflexion that their narratives of Enlightenment brought to the task of repositioning Europe in a secular history of the world was overshadowed by a tradition of Western triumphalism which emerged in the aftermath of the Revolutionary and Napoleonic Wars, 1789–1815; when Europeans acquired massive additions to the populations, territories and natural resources of other continents under direct or informal control. During an 'imperial meridian', European naval and military superiority over the states and societies of other continents emerged as irresistible. Furthermore, by 1825, European settlers throughout the Americas and the South Pacific had escaped from all metropolitan and aristocratic restraints upon their economic activities and their relations with the indigenous peoples of colonized worlds. White settlers became free to exploit 'their' heritage under rules of their own making, tempered only by religion. As its technological and economic lead over other continents widened, Europe moved clearly onto a path which led, as time went on, to clear divergences in levels of productivity and living standards between East and West and North and South.

Against this background, the writing of history by Europeans departed from its moment of cosmopolitan enlightenment, moved in a Hegelian direction and returned to the introspection and cultural arrogance associated with imperial Rome. Here and there aspirations to write universal histories survived. Isolated scholars eschewed the condescensions embodied in nineteenth century science, technology, liberalism, and

colonialism, and a scatter of historians continued to work outside the encompassing national frameworks, and dominant teleologies associated with the rise, rationality, and hegemony of Europe. Nevertheless, it took the destruction and cultural shock of two world wars in the twentieth century to encourage a return to the rewriting of histories of mankind. That program, conducted by a small group of erudite scholars, began in the climate of pessimism and breakdown in the liberal consensus that followed the Great War, when Spengler, Toynbee, Sorokin, Wells, Mumford, and Dawson wrote universal histories of humanity to help Europeans comprehend the decline and barbarity of the West. In history they attempted to find some substitute for Nietzsche's proclaimed demise of Christianity and rationality. Their scholarship is inspiring, but their explicit search for spiritual meanings in the histories of the rather amorphous entities called civilizations did become a compelling framework for scholars and students to work within. The attempt by UNESCO, after the World War II, to fund the writing of a multivolume collaborative history of the world (designed to promote peace across nation states) failed to provide a paradigm for the reconstruction of secular histories on an ecumenical scale. Long after the war, historical scholarship continued to be located where its founders had situated it from the time of the Renaissance and Reformation onwards—namely, within the confines, chronologies, and parameters established for comprehending the origins and evolution of European (including North American and Australasian) societies and nation states.

2. Globalization and the Revival of Global History

Fortunately pockets of global history survived—notably at the universities of Chicago, Northwestern and Leipzig. Complementary branches of scholarship continued to flourish in the forms of: survey courses in Western civilization; area studies, histories of trade and imperialism, investigations into the impact Western technology and culture upon other peoples and above all, as flows of encyclopedias, chronicles and textbooks for schools and popular education concerned with the world as a whole. Thus, 'tribes' of professional historians with interests in fields that transcended nation states and Europe possessed the credentials and remained ready to refashion their intellectual identities, when demands for global history reappeared from the late seventies onwards. That revival could be retraced to canonical publications by M. Hodgson, W. McNeill and L. Stavrianos, but research and teaching has only matured into university programmes for the sustained study of world history over the past quarter of a century of 'globalization.' Although there is nothing particularly new about the current phase of economic integration and inter-

dependence which are the core features covered by that label. Long distance trade goes back millennia to prehistory, and the exchange of commodities across frontiers and overseas had always involved arrangements for counterpart flows of funds, conversions of currencies and the operation of transnational, commercial and financial organizations of a more or less sophisticated kind within and across Europe, Africa and Asia.

Nevertheless real discontinuities had occurred. For example, by the late nineteenth century, supported by steam-powered ships and trains, the volume of commodities and services traded across national borders, continental boundaries and oceans had increased dramatically compared to 1800. Commerce was promoted and sustained by movements of capital, migrations of labour and transfers of technology and information around the world on an unprecedented scale and at ever increasing speeds. Political impediments to international flows of exports, imports, money, credit, capital, labor, technology, and information diminished sharply during the liberal international order that prevailed between 1846 and 1914.

State controls revived again during nearly four decades of neomercantilism that succeeded World War I; when governments everywhere sought to insulate their domestic economies from inflows and outflows of commodities, from the emigration of their nation's capital and skilled labor, and from fluctuations associated with unregulated rates of exchange. In the interests of narrowly defined national competitive advantages and political stability, sovereign states proceeded to de-link economies from participation in an evolving global economy. Between 1914 and 1948 overall levels of world trade fluctuated in volume, but ceased to grow. Warfare and political regulations depressed movements of capital and labor across frontiers. Controls on the convertibility of currencies hampered international exchanges of every kind.

Normalcy returned after World War II when a discernible shift towards liberalism in the political and legal framework for commerce among nations was also promoted by even more radical improvements in new technologies for the transportation of goods and people across space and time; and by vastly more efficient communications systems for the widespread diffusion of commercial intelligence, messages and cultural information between distant locations and separated populations. Revolutions in transport, television and communications widened markets and created opportunities for the organization of trans-continental and multinational business across an expanding range of production, distribution, finance and commerce.

Unsurprisingly, the implications of 'globalization' for hitherto separated communities, bounded spaces, distinctive cultures, partially connected local eco-

nomies and nominally sovereign states have been analyzed by social scientists. Intensified demands have appeared for historians to supply perspective on trends towards integration—that often appear to economists, political scientists, sociologists, and philosophers (as well as natural scientists) to represent profound discontinuities in the ways that ubiquitous technological, economic, and cultural forces are clearly shaping our interconnected world. This demand is supported by Western governments, uneasily aware that their powers to control societies, nominally under their rule, are transforming their claims to effective sovereignty over the populations, territories, and assets within their borders into something more tenuous than used to be the case during the long era of competing nation states. Their anxieties (eloquently fostered by 'official' intellectuals) have been compounded by recent military, technological, and economic achievements of several Asian societies, especially Japan and, the 'Tigers' but latterly (and more significantly) by the reawakening of China. The rest are perceived to be catching up and modern mercantilists (who continue to conceive of the world in terms of winners and losers or clashes of civilizations), worry about the relative decline of Western hegemony. Politicians and publics turn to historians for explanations of the past and more reassuring predictions for the future.

National frameworks for political action and academic enquiry are now widely recognized as unsatisfactory. That comprehension is patently functional for natural sciences investigating environmental issues, for biological sciences concerned with the health of human populations, and for social sciences analyzing international relations, crime, migration, communications, and whole ranges of economic activity, which can no longer be regulated within the boundaries of nation states. It is also clear that national and local cultures are being subverted by advertising, by fashion, and by the popular arts, especially music. The means and media of modern transportation, travel and communications (cheap and available to the masses), are opening up cosmopolitan discourses (usually in English) that contribute to the reshaping of personal identities around the world and transform frameworks for behavior, especially among younger generations.

3. Traditional, Postmodern, and Global Histories

Pressures to make, write and teach histories of the world that might become more contemporary and relevant has encouraged the discipline to become less intellectually, politically, spatially and chronologically confined. They are intensifying, seem irresistible and are being welcomed as a return to the ecumenical programmes of Herodotus, Vico, Voltaire, Leibniz and von Ranke who opined 'there is no history but universal history as it really was.'

Intellectual arguments for resituating national and European histories into frameworks that might better address the problems of the new century have been persuasively rehearsed by proselytisers for global history in recent years—as indeed have their anti-theses—recommendations to concentrate attention upon microhistories of difference, diversity and a stance of humility towards the complex histories of ‘unknowable others.’ That postmodern exhortation is understandably suffused with a concern that the revival of global history will implicitly celebrate ‘triumphs of the West’ in science, technology, military power, and economic development. Unfortunately that has already occurred in several recent publications by journalists and even by historians whose pretensions to offer overviews of world history seem truncated in time, confined in space and selective in areas of social endeavour emphasized for recognition as universal achievements; and whose understanding of environmental, biological, and other natural constraints on human action look unscientific.

As humanistic scholarship, global history aims to mature into a renaissance and not a return to the arrogance of Rome or to Victorian triumphalism. Modern history has matured into a capaciously catholic subject, willing to take many kinds of hitherto myopic exclusions into its narratives, to reposition knowledge of the past into more relevant spaces, models and chronologies and to plunder the natural and social sciences (as well as archeological remains) for insights into the evolving human condition. In finding voices and viewpoints to construct histories on a global scale, historians (alive to the second cosmopolitan enlightenment that emerged during the closing years of an otherwise dark twentieth century) have not found it difficult to detach their narratives from another politically correct but romanticized appeasement of the underprivileged on the one hand, or from the implicit denigration of non-Western people and cultures, that flows from Eurocentric convergence and modernization theory on the other.

In an ongoing search for strategies to teach and to write compelling histories on a global scale, they will respect the preferences of their colleagues, whose postmodern anxieties about their capacities to comprehend ‘others’ or to offer plausible conjectures and metanarratives about major trends and significant events, leads them towards dense description, microhistories, and biography. Such flowers should certainly bloom, but the intellectual effort required to landscape the garden also demands attention and imagination from historians.

Global historians confront more challenging intellectual problems in trying to persuade the majority of their academic colleagues (writing bounded histories and bunkered in the security of national archives), that their attempts to write on a global scale could meet the standards of a discipline deeply concerned with detail and documentation that dis-

courages generalization beyond strictly confined geographies and periods. Having rejected Toynbee’s quest for spiritual guidance as a tenable mission global historians tend to make a general case behind arresting metaphors extolling the need for histories constructed as: ‘maps on the largest scale,’ as ‘photographs taken through the lenses of wide-angled cameras,’ as ‘architectural frameworks’ and ‘as views; contemplated from space.’ Many contend that the boundaries and chronologies of parish, let alone national or continental histories, seem equally problematical to define. Nevertheless, coming as most do, as recognized specialists in but one or two fields of history, global historians remain conscious of their credentials and are entirely sensitive to their epistemological vulnerability. It continues to be much easier to conduct research, claim territory and acquire status as national or local historians.

4. Current Paradigms: Connections and Comparisons

Nevertheless, historiographical and methodological discussion as well as the volume of publications (off and on line) has increased exponentially. Most of the bibliography is concerned with connections and comparisons across national frontiers and cultural boundaries. Most of it displays a scientifically informed appreciation of evolutionary environmental and biological constraints on all forms of human activity, and represents a welcome reunification of natural and human history. Celebrated and influential books display a talent for the construction of synoptic histories about matters that have been important for humanity as a whole. As more historians risk writing on a grand scale, the field looks set to produce plausible, but competing metanarratives into which the overwhelming flow of parish, regional and national histories could be reconnected.

Meanwhile two distinct approaches have already developed. First (and central) is the model, persuasively exemplified and elaborated in several books by McNeill. He has inspired a programme of historical investigation into connections across continents, countries and geographies as well as through very long spans of time. McNeill’s assumption is that such connections, interactions, encounters, contacts with outsiders can be represented as the origins and engine of economic, social, political, military, cultural, religious, technological, and all other conceivable types of change studied by historians; and that connection can be pursued with scholarly detachment and a proper appreciation of the long chronologies required to construct histories that aim to select and analyse *major* forces at work in the evolution of humanity, McNeill’s recommendations aim to avoid the condescension of cultures, the restrictions of time and the arrogance of nations built into currently dominant

styles of history. His (and Crosby's) books show that 'interactions' were often malign: the spread of plagues, diseases, and parasites; those waves of destructive invasions by nomads; wars of conquest, plunder, and imperial expansion; forcible conversions to alien systems of religious belief; and the subversion of indigenous traditions, cultures, societies, and communities by strangers, have been part of the universal experience of suffering.

Well-specified types of connection can be investigated over significant periods of contact and transformation. Its taxonomical divisions include: trade, investment, warfare, religion, migrations, transfers of useful knowledge, technological diffusion, botanical exchanges, and the spread of diseases. Proper attention is also accorded to the means of transportation and modes of communication that secured, facilitated and cheapened contacts by sea, land and latterly by air. Once a relevant range of major connections are traced, understood and their significance evaluated global history will broaden our education and provide the spatial and chronological perspectives required for the appreciation of the histories of nations, peoples, localities, and groups embodied within particular histories that the majority of their colleagues study in greater depth and sophistication, but too often in isolation.

Put simply, the second major approach extends geographical catchment areas for comparative histories of topics that could be the subject of illuminating study across national boundaries, continents, and separable cultures. Although the linguistic qualifications for such research are extremely difficult to acquire, and the scholarship needed to make sense of contrasts discovered is formidable, the methodological problems are similar to those encountered for exercises conducted within European, national, or indeed regional frameworks.

Comparative history helps scholars to surmount the complexity and tyranny of local detail and seeks to offer coherent but sensitive answers to a great variety of questions that historians select for investigation. For global historians, the illumination derived from the method tends to be realized when they can concentrate upon well-defined artefacts, institutions, organizations, social practices, attitudes, and beliefs that are found in dispersed cultures, and which have already been studied in scholarly depth, in the context of particular locations and which exhibit comparable, but, more importantly, dissimilar geographical economic, political, social and other features. Then, as Marc Bloch anticipated, the 'comparative method can elicit from the chaotic multiplicity of circumstances those contrasts which were generally effective.'

In order to communicate salient differences global historians will be tempted to aggregate and average contrasts across wider spaces and larger populations (continents, oceans and civilizations) than their colleagues who have the sources to engage in exercises in

comparative history within circumscribed boundaries. For purposes of research, designed as building blocks for the production of metanarratives that are truly universal, the comparative method looks set to dominate the field for years to come. It has, for example, already generated a bibliography of global histories concerned with the family, youth, marriage, diet, housing, health, military organization, government, human rights, parliaments, nationalism, religions, fundamentalism, and revolutions, and there are serious debates on two seminal topics: gender and long-run economic growth.

5. Comparative Histories of Gender and Material Progress

Gender is perhaps the quintessential 'analytical category' for purposes of universal social history and it interacts with other categories including class, ethnicity, race, nation citizenship, and colonialism for purposes of constructing world histories that 'keep humanity in view.' Yet no surprise will be occasioned by the appearance in recent decades of histories seeking to encompass the world economy—going way back in time and written to expound and to explain the disparate levels of material progress achieved by countries, societies, and communities located on all continents. Most people, in most places for most of history, have been preoccupied with obtaining food, shelter, clothing, and other manufactured artefacts that they required to sustain either a basic, a comfortable, and, only latterly, an agreeable standard of living.

A modern generation of historians is resituating the scientific, technological, and economic achievements of Western societies in global context. They call upon libraries of books and articles dealing with African and Asian agricultures, industries, towns, commercial networks, communications, trade, science, technologies, cultures, business organization, taxation, state systems, Government policies etc., etc., covering the last millennium and written in large part by historians from universities not long emancipated from imperial rule. This enormous, but still far from comprehensive, volume of historical research has, moreover, been communicated to the West by specialists in area studies from North American, European, Australian, and Japanese universities. Not long after World War II and during an era of decolonization, historians were offered the opportunity—provided by the accumulation of a large body of archival knowledge—long available in Europe and North America, but emerging in Asia, the Middle East and Africa and Latin America—to reposition their hitherto disconnected histories of wealth and poverty one against another in order to construct global histories of material progress that might have satisfied the aspirations of Montesquieu, Voltaire,

Smith, and their 'enlightened' followers, and pleased Max Weber.

Recent attempts to survey this monographic literature has been dominated by scholars from Western universities. Syntheses from China, Japan, India, Africa and the Near East would be a salutary correction to what is a plural but nevertheless European perspective. Meanwhile, scholars have carried forward a Weberian program for teaching and writing a truly global history of material progress. They have found that the quantitative evidence for such purposes would certainly not be regarded as sufficient to underpin a serious programme of historical research designed to account for long-term economic success and retardation among the national economies of Western Europe, North America, and Australasia. In the absence of hard data or even of an acceptable range of contemporary observations (which could be included in a narrative designed to track chronologies of divergence in productivities and living standards between Europe and Asia), global histories of material progress continued to be marred—by three powerful Marxist and Weberian assumptions. First, that Europe's technological and economic superiority over Asia (which became first visible in the eighteenth century and then unmistakable during the nineteenth and twentieth centuries) had evolved gradually, but cumulatively, over several centuries before 1815. Over that time Europe's political, legal, cultural, and institutional heritage for the conduct of economic activity continued to be exceptionally favorable for the promotion, first, of Smithian growth, and, following on from that, for technological progress. Thirdly (and in contrast to Europe), other populations lived and worked in more hostile environments and this big ecological fact restrained and often arrested long-term economic development in Asia and Africa—although our knowledge of pre-colonial African economies remains entirely thin.

Those who follow Weber, Marx, and a tradition of writing that can be tracked back to Montesquieu, insist that Europe's competing economies had been on a potentially more promising trajectory for long-run growth for centuries. They emphasize persistent contrasts between European and Asian institutional arrangements, political systems, legal frameworks, cultural conditioning, and religious beliefs, within which economic activities continued to be embedded for at least four centuries before 1815. Cruder Weberians juxtapose the preconditions for long-run Smithian growth that they found operating efficiently in Europe (or rather in some societies in north western Europe) in sharp contrast to Asia, where political, social, familial, cultural and other negative institutions and beliefs restrained the emergence of competitive markets for commodities, capital, land and labor.

Their critics, who seem more deeply and widely read in the now massive bibliography of economic and social history dealing with South, East and South East

Asia, and the Ottoman Empire, have produced samples of data (related largely to population growth, levels of urbanization, balances of commodity trade, yields per hectare and real wages) for selected regions of Asia and the Middle East that at the very least qualify any impression that labor productivity levels and standards of living afforded by European economies to their populations were discernibly (let alone significantly) superior long before the outbreak of the French Revolution. Of course at this level of 'continental' aggregation, the selection of samples of data for comparative history remains problematical because the Asian economy includes much larger shares of the world's cultivable land, resources, income, and population than early modern Europe.

Nevertheless, economic historians are a long way from offering statistically secure comparisons for average levels of productivity and living standards. Meanwhile, the accumulating volume of research now available on Asian agricultures, industries, trades, credit systems, mercantile enterprise, transportation networks, markets for commodities, and factors of production degrades simplistic Weberian perceptions that Europe alone had evolved the political, institutional, legal, cultural, and religious frameworks required for an effective process of Smithian growth long before other continents into dubious propositions. As Marshal Hodgson observed decades ago: 'All attempts to invoke premodern seminal traits in the occident can be shown to fail under close historical analysis.' Braudel, Chaudhuri, Frank, Pomeranz, Goldstone, McNeil, Sugihara, Washbrook and other historians of long-run material change concur. Jones would probably (as he did in *Growth Recurring*, 1988) continue to revise and reformulate some of the positions he adopted in the first edition of the *European Miracle* in 1981. From his comparisons of levels and types of development achieved by Europe and Asia for the early modern period, Braudel inferred that 'the populated regions of the world faced with demands of numbers seem to us to be quite close to each other.' But there is he continued 'a historiographical inequality between Europe and the rest of the world. Europe invented historians and made good use of them. Her own history is well lit and can be called as evidence or used as claim. The history of non-Europe is still being written. And until the balance of knowledge and interpretation has been restored the historian will be reluctant to cut the Gordian knot of world history.'

That acute observation applies across the entire field because histories of gender and capitalism are but two current areas of controversy that display the unrealized potential of universal history. In truth, it seems difficult to conceive of any grand theme that attracts attention from scholars engaged in writing national and local histories that could not be repositioned within wider geographies and longer temporal periods. For purposes of research, teaching and

debate, comparisons and connections are now the two dominant styles of global history. Pursued with sensitivity they should deepen our understanding of difference and diversity. In time they will meet demands for perspective from those trying to comprehend accelerated trends towards interdependence and integration on a world scale and allow for a less ethnocentric appreciation of the manifold achievements of more peoples, communities and cultures over long spans of human history.

See also: Civilization, Concept and History of; Comparative History; Gender History; Globalization and World Culture; Historiography and Historical Thought: Classical Period (Especially Greece and Rome); Historiography and Historical Thought: Current Trends; Historiography and Historical Thought: East Asia; Historiography and Historical Thought: Indigenous Cultures in the Americas; Historiography and Historical Thought: Jewish Tradition; Historiography and Historical Thought: Modern History (Since the Eighteenth Century); Historiography and Historical Thought: South Asia; Historiography and Historical Thought: Southeast Asia; Historiography and Historical Thought: Sub-Saharan Africa; History of Science; History of Technology; History: Overview; Modernization and Modernity in History; Time, Chronology, and Periodization in History

Bibliography

- Albrow M 1996 *The Global Age: State and Society Beyond Maturity*. Polity Press, Cambridge
- Anderson B S, Zinsser J P 1988 *A History of their Own: Women in Europe from Prehistory to the Present*. Harper & Row, New York
- Barnes H E 1962 *A History of Historical Writing*. Dover Publications, New York
- Bentley J H 1993 *Old World Encounters*. Oxford University Press, Oxford, UK
- Brook T, Blue G (eds.) 1999 *China and Modern Capitalism. Genealogies of Sinological Knowledge*. Cambridge University Press, Cambridge, UK
- Chaudhuri K N 1990 *Asia Before Europe, Economy and Civilisation of the Indian Ocean from the Rise of Islam to 1750*. Cambridge University Press, Cambridge, UK
- Costello P 1993 *World Historians and their Goals. Twentieth Century Answers to Modernism*. Northern Illinois University Press, De Kalb, IL
- Crosby A W 1986 *Ecological Imperialism: The Biographical Expansion of Europe, 900–1900*. Cambridge University Press, Cambridge, UK
- Curtin P D 1984 *Cross Cultural Trade in World History*. Cambridge University Press, Cambridge, UK
- Diamond J M 1998 *Guns, Germs and Steel: The Fates of Human Societies*. Vintage Random House, London
- Frank A G 1998 *ReOrient: Global Economy in the Asian Age*. University of California Press, Berkeley, CA
- Goody J 1996 *The East in the West*. Cambridge University Press, Cambridge, UK

- Heiser C B Jr. 1990 *Seed to Civilization: The Story of Food*. Cambridge University Press, Cambridge, UK
- Hirst P, Thompson P 1996 *Globalization in Question: The International Economy and the Possibilities of Governance*. Routledge, London
- Hodgson M G S 1993 *Rethinking World History*. Cambridge University Press, Cambridge, UK
- Jones E L 1998 *Growth Recurring: Economic Change in World History*. Oxford University Press, Oxford, UK
- Kishlansky M, Geary P, O'Brien P, Wong R B 1995 *Societies and Cultures in World History*. Harper Collins, New York
- Landes D S 1998 *The Wealth and Poverty of Nations: Why Some are so Rich and Some so Poor*. Little Brown, London
- Lewis B 1995 *The Middle East. 2000 years of History from the Rise of Christianity to the Present Day*. Weidenfeld and Nicholson, London
- Livi Bacci M 1992 *A Concise History of World Population*. Blackwell, Oxford, UK
- McNeill W 1991 *The Rise of the West: A History of the Human Community*. University of Chicago Press, Chicago, IL
- Mazlish B, Buultjens R (eds.) 1993 *Conceptualizing Global History*. Westview Press, Boulder, CO
- Mokyr J 1990 *The Lever of Riches: Technological Creativity and Economic Progress*. Oxford University Press, Oxford, UK
- O'Brien P K (ed.) 1999 *Philips Atlas of World History*. George Philip Limited, London
- Offen K, Roach R R, Rendall J (eds.) 1991 *Writing Women's History: International Perspectives*. Macmillan, London
- Pomeranz K 2000 *The Great Divergence: Europe, China and the Making of the Modern World Economy*. Princeton University Press, Princeton, NJ
- Sanderson S K 1995 *Social Transformations: A General Theory of Historical Development*. Blackwell, Oxford, UK
- Stavrianos L S 1989 *Lifelines from our Past: A New World History*. Pantheon Books, New York
- Smil V 1994 *Energy in World History*. Westview Press, Boulder, CO
- Weber M 1961 *General Economic History*. Collier, New York
- Wong R B 1997 *China Transformed: Historical Change and the Limits of European Experience*. Cornell University Press, Ithaca, NY
- Worster D (ed.) 1998 *Perspectives on Modern Environmental History*. Harvard University Press, Cambridge, MA

P. K. O'Brien

Global Population Trends

During their evolution of tens of thousands of years, humans have progressively, although not without setbacks, learned how to better feed themselves, make clothing, provide shelter, cure and prevent diseases, improve their health, and modify their fertility. People have learned how to interact in societies which were at first based on gathering, hunting, and fishing. Many societies then experienced the agricultural revolution combined with the domestication of animals. The evolving increases in labor productivity and scientific advancements generated the industrial revolution and urbanization, the foundations of contemporary soci-

eties. In the process, populations have grown from very small numbers to the six billion global population reached in 1999. Demographically speaking, for many centuries and millennia changes were erratic and slow until several centuries ago mortality started to decline, life expectancy increased, somewhat later fertility commenced its decline, and populations around the globe began to grow at increasing rates. The more pronounced and rapid changes have occurred in the most recent 100–200 years, although the roots of change reach much further into the past.

1. The World's Population

In 1695 Gregory King (1973) was among the first scholars who attempted to estimate the size of the world population. His estimates were within a range of 495–700 million—preferring 630 million as the base for his projections—in line with what scholars engaged in similar exercises since then consider as realistic (Table 1). In 1999, the size of the world population was estimated to have reached six billion inhabitants. 'Estimated' because even at the end of the twentieth century uncertainties about the exact size of the world population persist, as there are numerous countries which lack a recent reliable census, and which have inadequate population registration systems or none at all.

Through most of human history the increase in the world population was slow and punctuated by periods of decline. The technologies of food production, as well as the ways and means of health protection, were relatively primitive and humans were thus quite significantly exposed to the vagaries of nature. A

notorious case was the epidemic of the plague or Black Death which in the fourteenth century spread from Asia to Europe and Africa and is believed to have decimated Europe's population by one-third.

The growth of the world population gained momentum in the seventeenth to nineteenth centuries. However, the critical acceleration occurred after World War II. During the second half of the twentieth century the global growth rate has been close to 2 percent per year. Improved personal health behavior and public health measures as well as the application of curative and preventive medicine which had been attained and practiced in the industrial countries earlier were rapidly spreading in the developing world (see *Mortality: The Great Historical Decline*). It was during this period that the peak and the turning point in global population growth was reached. Since its climax in the 1960s with an annual rate of 2.0 percent, this rate has declined to 1.4 percent in the 1990s. In many parts of the world, in developed and developing countries, fertility has been declining rapidly (see *Fertility Transition: Cultural Explanations and Demographic Transition, Second*). It is worth noting that the absolute annual additions continued to increase through the end of the twentieth century. They had also reached their peak and were commencing their decline. Around the year 1900 the annual absolute global population increase was about eight million. In the year 2000 it was around 80 million.

Through the end of the twentieth century the number of years required to add an additional one billion population has been decreasing. The first billion was reached around 1804, the second in 1927 (123 years later), the third in 1960 (33 years), the fourth in 1974 (14 years), the fifth in 1987 (13 years),

Table 1
Estimates of world population, 7000 BC–2000 AD

Date	Population in millions	Average annual increase in % since preceding date	Absolute annual increase in millions since preceding date
BC			
7000–6000	5–10	n.a.	n.a.
AD			
1	200–400	0.0	0.05
1650	470–545	0.0	0.13
1750	629–961	0.4	2.9
1800	813–1,125	0.4	3.5
1850	1,128–1,402	0.5	5.9
1900	1,550–1,762	0.5	7.8
1950	2,521	0.8	17.3
1960	3,022	1.8	50.1
1970	3,696	2.0	67.4
1980	4,440	1.8	74.4
1990	5,266	1.7	82.6
2000	6,055	1.4	78.9

Source: United Nations 1973, United Nations 1999

Table 2

Composition of the world population by major region, in percent, 1750–2000

Region	Year					
	1750	1800	1850	1900	1950	2000
World, in millions	791	969	1,265	1,656	2,521	6,055
World, in percent	100	100	100	100	100	100
Africa	13	11	9	8	9	13
Asia	63	66	64	57	56	61
<i>China</i>	27	n.a.	n.a.	26	22	21
<i>India/Pakistan/Bangladesh</i>	23	n.a.	n.a.	18	17	21
<i>Remainder</i>	13	n.a.	n.a.	13	16	18
Europe	21	21	22	25	22	12
Latin America and Caribbean	2	2	3	4	7	9
North America	0	1	2	5	7	5
Oceania	0	0	0	0	1	0
More developed countries	25	25	27	33	32	20
Less developed countries	75	75	73	67	68	80

Source: Durand 1974, United Nations 1973, 1999

and the sixth in 1999 (12 years). Indications are that this number of years will start to increase in the twenty-first century (see *Population Forecasts*).

The distribution of the world's population has been changing over time. For several centuries large populations have been living in China and the Indian subcontinent. In the year 2000 each had about 1.3 billion inhabitants constituting 21 percent of the world population (Table 2).

In the course of the nineteenth century it was the population of the industrial countries in Europe and North America which grew more rapidly than in other regions. Its proportion increased from a quarter of the world's population to one-third between the beginning and the end of the nineteenth century. In the twentieth century, population growth in the developed countries slowed down. Subsequently, during the second half of the century, rapidly declining mortality in Third World countries combined with a fertility decline led to rapid population growth so that their proportion of the world's population increased from two-thirds around 1950 to 80 percent around the year 2000.

2. Mortality

As best as can be estimated, the average human lifespan in prehistoric and during most of historic times was rather short. The scarce archeological evidence available suggests that most adult deaths occurred between the ages of 20 and 30. In the Iron Age in France average life expectancy was presumably 10–12 years. A life table for ancient Greece, prepared from burial inscriptions, indicates an average life expectancy of 30 years around 400 BC. Life expectancy in Geneva was estimated at about 21 years in the period 1561–1600 and about 26 years in 1601–1700.

The latter and other estimates based on fragmentary data in various European countries from the thirteenth to the eighteenth centuries indicate that mortality was very unstable and variable with life expectancy between 20 and 40 years (United Nations 1973). A relatively comprehensive estimate covering Northern and Western Europe, Northern America, and Oceania (The 'West') for the middle of the nineteenth century still resulted in an average life expectancy of 40 years. From thereon health conditions improved and life expectancies grew quite steadily in these countries. Elsewhere in Europe and the former Russian empire mortality started to decline later. In Southern Europe, in Italy, for instance, in 1871–80 life expectancy at birth did not exceed 35 years and a steady fall of mortality started after 1875. Japan was another country where mortality was low quite early. Its crude death rate at the end of the nineteenth century was comparable with many European countries.

Malnutrition, poor hygiene, parasitic infections, and disease were, and in many parts of the world still are, the principal causes for high mortality. It took many centuries of gradual progress to have a visible effect on survival. Changes in diet, personal and public hygiene, reinforced by advances in science and medicine, generated dramatic increases in life expectancy. During the past 100–150 years, life expectancy has essentially doubled from 30–40 years to 70–80 years in those societies where most people at the beginning of the twenty-first century enjoy a decent standard of living.

In the Middle Ages most people lived in primitive living quarters with their domesticated animals and their excrements, with an open fire fueled by wood and manure in the middle of the floor and a hole in the roof to let out smoke and cinders, without any toilets and bathrooms, without soap, windows, furniture, dishes,

Table 3
Mortality measures, world and selected countries, 1950–2000

Period	Life expectancy at birth in years							Infant deaths in the first year of life per 1,000 live births						
	World	Sweden	Japan	Hungary	Brazil	India	Uganda	World	Sweden	Japan	Hungary	Brazil	India	Uganda
1950–5	47	72	64	64	51	39	40	155	20	51	71	135	190	160
1960–5	52	74	69	69	56	46	44	117	15	25	44	109	157	125
1970–5	58	75	73	69	60	50	47	93	10	12	34	91	132	116
1980–5	61	76	77	69	63	55	46	78	7	7	20	64	106	118
1990–5	64	78	80	69	66	60	37	62	6	4	13	47	78	118
2000 estimate	66	79	80	72	67	63	42	55	5	4	10	40	69	101

Source: United Nations 1999

and cutlery. Individual societies around the world were better equipped in some respects, for instance, the Chinese employed coal as a fuel, and soap was commonly used in the Islamic world. Frequent failures of harvests and major epidemics were commonplace everywhere. It took many centuries for standards of living and hygiene to improve and thus provide favorable conditions for better health. Residential chimneys in Europe became common around 1200 and it was Marco Polo who discovered in 1306 that the Chinese used coal to heat their homes. Soap was manufactured for the first time in London presumably in 1259. Frequent bathing, however, was considered unhealthy and as late as in the sixteenth century Queen Elizabeth I was considered eccentric for her frequent bathing—once a month. It was not until the Industrial Revolution of the eighteenth and nineteenth centuries that the way people lived changed considerably and irrevocably. Significant increases in productivity, including agriculture, provided for the possibility of improved nutrition, shelter, and living conditions in general. In combination with the gradually increasing application of personal hygiene, the installation of indoor plumbing and introduction of sewage systems, the accumulation of scientific and medical knowledge as well as the spread of health services networks, this all led to the gradual improvements in health and the decline of mortality.

The profound long-term changes in patterns of death, disease, and disability have been labeled the *epidemiological transition* (see Omran 1971 and *Mortality, Epidemiological, and Health Transitions*). It has three courses of change: a shift from infectious and parasitic to chronic and degenerative diseases, a shift in the burden of death and disease from the younger to the older population, and a shift from a health picture dominated by mortality to one dominated by morbidity. In general, these directions of change occurred in the developed countries before the middle of the twentieth century. In developing countries, during the latter part of the twentieth century, often the increase in chronic and degenerative diseases coexisted with

relatively high levels of infectious diseases (Bobadilla et al. 1993). Most of the morbidity of contemporary advanced societies is preventable because in large part it is due to ills and imperfections of social and individual behavior. It is due to undesirable side effects of industrialization, urbanization, and mass consumption such as atmospheric pollution, environmental degradation, accidents, mental tension and distress, consumption of tobacco, alcohol and drugs, unhealthy diets and obesity (Jozan 1998). Efforts to control and reduce this morbidity are underway, but results are mixed. The complexities are numerous. Consequently, there are large differences regarding success and failure in lowering mortality and morbidity between countries, within countries among regions and social strata.

The progress and problems of mortality and morbidity change in the past 50 years are illustrated by data in Table 3. In most countries around the world major improvements were achieved during the second half of the twentieth century. Nonetheless, there were two types of exceptions. Hungary exemplifies the stagnation in mortality decline of the formerly socialist countries of central and eastern Europe and Uganda represents the countries of sub-Saharan Africa where mortality declines were reversed mainly as a result of the HIV/AIDS epidemic. At the same time, significant improvements in the health conditions of infants have been achieved around the globe, but major differentials persisted through the end of the twentieth century (United Nations 1999).

3. Fertility

In the long run, additions of births have to at least match attrition by death if a society is to survive and perpetuate. That is what has been happening throughout humanity's history and before. Fertility could not decline from traditional levels prior to any significant decline of mortality without causing the extinction of human populations.

As long as mortality was high and a large proportion of infants, children, and adolescents died before entering their reproductive ages, women would bear on average 5–8 children. Such high fertility is persisting at present in a number of countries with high mortality, especially in sub-Saharan Africa. There are, however, only few such countries remaining. For the most part, fertility has either already declined to very low levels or countries are on their way between high and low fertility.

The biological maximum a woman can bear during her childbearing period is between 13 and 17 children. As a rule, this does not happen as there are numerous deliberate/conscious as well as unintended/unconscious mechanisms limiting births. Until the nineteenth century in the advanced countries, and until only several decades ago elsewhere, breast-feeding—occasionally combined with taboos in intercourse during lactation—was the decisive factor bringing down the number of children born to between 5 and 8. In premodern times almost exclusive breast-feeding tended to be the rule for infants and for many children at times until they reached their second or even third birthday. For most women this practice prevented a further pregnancy. An extreme example are the San of the Kalahari Desert who were not subject to any taboo and yet they achieved four-year birth intervals through the contraceptive effect of frequent (four times an hour) breast-feeds. The physiological process taking place is such that when the baby suckles on the nipple, nervous impulses pass from the breast to the brain to inhibit the release of hormones responsible for ovulation from the pituitary gland (see Potts and Short 1999 and *Fertility Control: Prevalence and Consequences of Breastfeeding*).

Women breastfeed as it is the most natural and convenient way to feed infants, at the same time the conception preventing effect was apparently known for millennia and in some societies widely realized. Aristotle stated in the *Generation of Animals* that ‘while women are suckling children, menstruation does not occur according to nature, nor do they conceive.’ In 1792, Mary Wollstonecraft wrote in the *Vindication of the Rights of Women*: ‘Nature has so wisely ordered things that did women suckle their children, they would preserve their own health, and there would be such an interval between the birth of each child, that we should seldom see a houseful of babies.’ Lactation has other critical salutary effects. Maternal milk is the most beneficial source of nutrition finely adjusted to the needs of the offspring. It is also a living secretion, rich in antibodies that protect against disease. Therefore, particularly in developing countries, breastfed babies tend to be healthier than their bottle-fed peers.

Other critical factors reducing fertility are delayed cohabitation of partners/marriage, the use of contraceptives, and the prevalence of induced abortions. Prior to the latter part of the nineteenth century, these

were known, however, they were not widely practiced and therefore they were not decisive in modifying fertility. Western European marriage patterns were the exception. Hajnal (1965) observed that since the eighteenth century fertility had been relatively low and at times declining in countries west of a line between St. Petersburg and Trieste precisely because the age of marriage was high. Later, the considerable fertility decline in western countries in the late nineteenth and early twentieth centuries was apparently brought about by the application of relatively unsafe induced abortions as well as the use of traditional methods of contraception, such as coitus interruptus (Santow 1995). There is, however, scarce statistical hard evidence to support this statement.

During the second half of the twentieth century safe induced abortions became increasingly available for a large proportion of the world population and unsafe abortions were widely practiced in Latin America and Africa (see *Infanticide and Induced Abortion*). A ‘contraceptive revolution’ was underway throughout the developed countries and increasingly in the developing ones as well. In most countries the use of contraceptives increased, modern contraceptives such as the IUDs and oral contraceptives found extensive application, and female as well as male contraceptive sterilization was widely adopted, although differences between countries were considerable. The prevalence of contraceptive use rose to between 70 and 80 percent of married women in the developed countries, and in countries such as Taiwan and South Korea, the use of contraceptives increased from 20 to 80 percent within a period of about three decades (see Frejka and Ross 2001 and *Family Planning Programs: Development and Outcomes; Eyewitness Memory: Psychological Aspects; Fertility Transition: Economic Explanations; Fertility Transition: China; Fertility Transition: East Asia; Fertility Transition: Southeast Asia; Fertility Transition: South Asia; Fertility Transition: Middle East and North Africa; Fertility Transition: Latin America and the Caribbean*).

Relatively late marriage, the use of contraception, and induced abortion are the *means* of modifying fertility (see *Fertility: Proximate Determinants*). The motivations and decisions to regulate fertility, subconscious or conscious, have numerous economic, social, cultural, and political roots. The following developments are critical. With economic progress children have evolved from being an asset to their parents to become a costly investment. Growing proportions of young women and men remained in school or found employment outside the household and therefore entered marital or consensual unions later than their parents. The family, particularly the younger generation, used to take care of older members of the household; increasingly older people have to take care of themselves possibly assisted by the state. As recently as in the nineteenth century women everywhere were oppressed and disadvantaged, now-

Table 4
Fertility measures, world and selected countries, 1950–2000

Period	Crude birth rate							Total fertility rate						
	World	Sweden	Japan	Hungary	Brazil	India	Uganda	World	Sweden	Japan	Hungary	Brazil	India	Uganda
1950–5	37.3	15.5	23.7	21.1	44.0	44.1	51.4	4.99	2.21	2.75	2.73	6.15	5.97	6.90
1960–5	35.2	14.7	17.2	13.6	42.1	42.0	48.7	4.95	2.33	2.01	1.82	6.15	5.81	6.90
1970–5	30.9	13.6	19.2	15.7	33.7	38.2	50.3	4.48	1.89	2.07	2.09	4.72	5.43	6.90
1980–5	27.4	11.4	12.7	12.5	29.6	33.8	49.9	3.58	1.64	1.76	1.81	3.63	4.47	7.00
1990–5	23.9	13.5	9.7	11.8	22.2	28.7	50.8	2.93	2.01	1.49	1.73	2.51	3.56	7.10
2000 estimate	21.4	10.0	10.0	9.6	19.8	24.1	50.2	2.64	1.57	1.45	1.33	2.21	2.93	6.92

Source: United Nations 1999

Table 5
Population growth and percentage aged 65 and over, world and selected countries, 1950–2000

Year	Population size, 1950 = 100							Percentage aged 65 and over						
	World	Sweden	Japan	Hungary	Brazil	India	Uganda	World	Sweden	Japan	Hungary	Brazil	India	Uganda
1950	100	100	100	100	100	100	100	5.2	10.3	4.9	7.3	3.0	3.3	3.0
1960	120	107	113	107	135	124	138	5.3	12.0	5.7	9.0	3.3	3.4	2.6
1970	147	115	125	111	178	155	206	5.5	13.7	7.1	11.6	3.7	3.7	2.6
1980	176	118	140	115	225	193	276	5.9	16.3	9.0	13.4	4.1	4.1	2.5
1990	209	122	148	111	274	238	346	6.2	17.8	12.0	13.3	4.3	4.3	2.5
2000 estimate	240	127	152	107	315	283	457	6.9	17.4	17.1	14.7	5.1	5.0	2.1

Source: United Nations 1999

ere did they have the right to vote, for instance. In various parts of the world many women are still battered, sequestered, mutilated, or exploited, but a basic trend supporting their empowerment is in progress. Thus, increases in women's education and employment, their earning potential in the marketplace, and growing equality vis-à-vis laws have enabled women to partake in decisions regarding fertility behavior.

Population policies—deliberate interventions to influence a population development—have frequently been formulated and employed by governments or international institutions (see *Population Policies, Politics of*). With respect to fertility, governments perceiving it to be too low may devise policies which they believe will raise fertility; those regarding high fertility as a burden implement measures to diminish it. Berelson (1977), for instance, demonstrated that there are numerous policies which may conceivably reduce fertility. He clarified how policy makers taking local circumstances into account should and usually do consider whether the policies they wish to employ are *acceptable* to the political elite and to the public, whether they are *feasible* with given resources and administrative infrastructures, and how *effective* they will be.

Particular social, economic, political, cultural, and religious circumstances sometimes in combination with population policies shape people's fertility desires, motivations, and behavior which is expressed in their marriage patterns, breast-feeding practices, contraceptive use, and prevalence of induced abortion and thus result in a certain level and trends of fertility.

The unprecedented changes in fertility which occurred during the second half of the twentieth century are illustrated in Table 4. The total fertility rate, a measure which can be interpreted as the average number of children borne by a woman, on a worldwide scale declined from about five children in the early 1950s to just over two-and-a-half in the year 2000. In the developed countries, represented by Sweden, Japan, and Hungary, fertility declined from an already low level in the early 1950s to significantly below the replacement level of about 2.1 children per woman. The decline was not linear, some fluctuations did occur, however, at the end of the twentieth century not much less than half the world's population lives in countries with below replacement fertility (see Frejka and Ross 2001). In developing countries around 1950, fertility was high, between six and seven children per woman. In a significant number of countries it declined close to the replacement level, such as in Brazil. In

others like in India fertility was considerably reduced. In some countries, however, no discernible change took place. Uganda is typical of sub-Saharan countries with stable high fertility, although there are some where a decline has commenced.

Even though there has been a considerable fertility decline in most developing countries in the past several decades, continued population growth can be expected and will be forthcoming during the first half of the twenty-first century. Past high fertility has produced large cohorts of young people who are presently in the childbearing stage of their life cycle or who will enter their childbearing years. Even if these generations will have low fertility, the total numbers added to the populations of the developing countries are almost sure to be much larger than the number of people dying (see *Population Dynamics: Momentum of Population Growth and Population Forecasts*).

4. Concerns

Table 5 illustrates basic demographic developments which are frequently causes for concern among policy makers, scientists, and the general public. The world population has grown more rapidly during the second half of the twentieth century than ever before. Rapid population growth, even though at a moderating pace, will continue for some decades. Extraordinarily fast population growth has occurred in developing countries. Brazil's and India's populations, for instance, have grown threefold, Uganda's 4.5 times. Such growth creates obvious dilemmas and pressures for governments, the private sector, and international organizations striving to raise educational levels, improve health conditions, broaden employment opportunities, and elevate living standards in general. Population growth coupled with increasing consumption in developed and developing countries increases dangers of specific resource depletions, environmental degradation, selective plant and animal extinction, which ultimately could pose problems for and harm the human species. Concern is also voiced that rapid population growth can be a factor exacerbating poverty and political instability.

Hungarian population change exemplifies a shrinking population. There were more than 10 such nations in Europe in the late 1990s and thus concerns for a diminishing labor force, declining international stature, the potential of dying out as a nation, and others have surfaced.

Rapidly declining and low fertility gives rise to changes in the age structure and together with increasing longevity are the root causes of population aging. In the late 1990s in many developed countries 15 percent or more of populations were of age 65 and over. The numbers and proportions of the elderly have been increasing and there is no doubt that they are going to continue to do so. This is one of the major

concerns of governments in the developed countries because considerable state expenditures go toward social security benefits of the elderly as well as their health care. As fertility is declining rapidly in many developing countries, population aging is becoming an important social and political issue.

Bibliography

- Berelson B 1977 Paths to fertility reduction: The 'policy cube.' *Family Planning Perspectives* 9(5): 214-19
- Bobadilla J L, Frenk J, Lozano R, Frejka T, Stern C 1993 The epidemiologic transition and health priorities. In: Jamison D T, Mosley W H, Measham A R, Bobadilla J L (eds.) *Disease Control Priorities in Developing Countries*. Oxford Medical Publications, Oxford, UK
- Chesnais J-C 1992 *The Demographic Transition: Stages, Patterns, and Economic Implications*. Clarendon Press, Oxford, UK
- Coale A J 1974 The history of human population. *Scientific American* 231(3): 40-51
- Cohen J E 1995 *How Many People Can the Earth Support?* W W Norton, New York
- Durand J D 1974 *Historical Estimates of World Population: An Evaluation*. Population Studies Center, University of Pennsylvania, PA
- Frejka T J, Ross A 2001 Paths to subreplacement fertility: The empirical evidence. In: Bulatao RA, Casterline JB (eds.) *The Global Fertility Transition*, Supplement to *Population and Development Review* 27
- Jozan P 1998 Some features of mortality in the Member States of the ECE. In: United Nations Economic Commission for Europe, *Background Papers of the Regional Population Meeting, December 1998* (unpublished)
- King G 1973 17th century manuscript book of Gregory King. In: Laslett P (ed.) *The Earliest Classics: Graunt and King*. Gregg, Farnborough, UK
- Omran A 1971 The epidemiologic transition: A theory of the epidemiology of population change. *Milbank Memorial Fund Quarterly* 49: 509-38
- Potts M, Short R 1999 *Ever Since Adam and Eve: The Evolution of Human Sexuality*. Cambridge University Press, Cambridge, UK
- Santow G 1995 Coitus interruptus and the control of natural fertility. *Population Studies* 49(1): 19-43
- United Nations 1973 *The Determinants and Consequences of Population Trends*. United Nations, New York, Vol. 1
- United Nations 1999 *World Population Prospects: The 1998 Revision*. United Nations, New York
- Wollstonecraft M 1985 [1792] *Vindication of the Rights of Woman*. Penguin, London

T. Frejka

Global Positioning System

From the beginning of time man has striven to answer three fundamental questions: where am I now, where I am going next, and what is the best route from here to there? The observation of the sun and stars, and then later the use of the compass, the sextant, navigational

maps, and charts of the heavens, have all been essential aids and instruments in this never-ending quest. GPS is the most recent contribution to this search.

1. *Origins and Instances*

Over the years the need to answer these questions has been the driving force behind the development of many civil and military applications. During World War Two, one of the nations in combat realized that, even if they managed to build warships relatively rapidly, it was much more difficult to find suitable commanders for the ships. Such positions required candidates with a high level of experience, especially for mapping the ship's course. From this dilemma LORAN (Long-range Radio Navigation) was born, a system essentially for use at sea, based on the reception and phase difference of radio waves transmitted from two antennae. These two stations transmitted a continual flow of coordinates to navigators equipped with the correct passive antenna, which they then used to pinpoint their exact location on the earth's surface. This facilitated the commander's decision-making enormously, as he knew the ship's position at any time and, by using additional electronic aids or by simply mapping the points already crossed on a nautical chart, he also knew exactly where the ship was headed. The GPS (Global Positioning System) was the natural evolution of LORAN.

At any given moment and point in the area covered by the system, GPS is capable of providing the three-dimensional coordinates (x, y, z, or latitude, longitude and height). Rapid progress in electronic component design, has miniaturized GPS equipment to the extent that even watches and telephones can contain GPS and can provide the wearer with their exact position, the route they followed, and where they have to go.

On certain commercial flights a screen displays a schematized map of the territory below and, using GPS, a pointer indicates the current position of the airplane, the flight path, and, amongst other information, how much time remains before landing. Passengers draw comfort from knowing where they are and when they will arrive. From the cabin they may be able to use the features mapped on the screen to recognize the physical characteristics of the territory below.

Aside from the passive and temporary use of GPS to determine position, route and destination, the system is also important for activities with long-lasting and consequential effects. The most important example is the use of GPS in cartographic or topographic research, and, by extension, to many forms of inquiry in social science. With GPS, the earth's surface becomes an enormous digitizer, and any point can be identified instantly by its three-dimensional coordinates and archived in a suitable information system, potentially increasing the intensity of sampling events and objects in space and time by several degrees

of magnitude. Such measurements affect research and practice in archaeology, economics, geography, planning, and sociology, as well as in the natural sciences. By determining precise coordinates of points on land or on the surface of water, GPS makes it possible to update maps by locating physical and social features from which GPS gathers data instantaneously and/or detecting changes that have occurred in reality but have yet to be recorded (e.g., new construction, physical changes in the environment, transient events, or those in constant transformation such as people, vehicles, coastlines, woods, glaciers and so on).

2. *The General Principles of GPS Functioning*

GPS receivers use radio waves to determine positional coordinates. The radio waves are emitted by a constellation of orbiting satellites, which function as a reference system for GPS. The constellation consists of 24 satellites, which are located in six orbits, inclined at an angle of approximately 50 degrees, at a height of around 20,000km. Known as NAVSTAR, the constellation is the property of the US government and is controlled and managed by the Department of Defense. The orbits and the distance between the satellites are organized in such a way that, ideally, every point on the earth's surface can be seen by at least five satellites at the same time, although this is not always this case in practice. The GPS carries out the triangulation of three satellites visible at the same time above the horizon, measuring the distance that separates it from each of these by calculating the time between each satellite (transmitter) to the GPS itself (receiver) using precise time signals emitted by the satellites. The triangulation allows the GPS to determine the latitude and longitude, though a fourth satellite is also necessary to determine a third coordinate giving the height at which the GPS receiver is located. To measure the time required by the signal to travel the distance from each satellite to the receiver, it is essential to use extremely precise equipment to transmit the signal, to calculate the time lapse, and also to pinpoint with extreme accuracy the position of each satellite in space. To this end each satellite is equipped with an atomic clock and its position is constantly monitored and regulated by a series of earth stations.

It is essential to consider the problem of errors incurred by external agents when dealing with instruments as sophisticated as the GPS. To determine the exact time the signal takes to reach the GPS receiver, it is also necessary to understand all the modifications and distortion the signal may undergo while passing through the ionosphere and the atmosphere. The causes of error and/or the malfunctioning of the system are varied, and can range from atmospheric conditions to the position of the satellite constellation in the sky and their visibility from the receiver. The location of the GPS receiver itself may cause problems.

For example the receiver may be unable to detect a satellite if it is located under a canopy of thick vegetation or in a city canyon, where its horizon may be modified by the height of surrounding buildings.

Those managing the system itself may introduce certain deliberate errors. GPS, for example, has been subject to SA (Selective Availability), a randomized error deliberately incurred in the time signal from the satellites to degrade the positional accuracy of the receiver to the order of 10 to 100 meters. Acting on a declaration made in 1996, on May 1, 2000, the President of the USA announced the end to the use of SA in GPS, with the immediate effect of increasing and encouraging the use of GPS for civil and commercial purposes throughout the world. In fact, the demand for GPS precision was such that, as soon as the system was recognized by civilian users as being of major importance, they also wanted instruments of reduced size and low cost to have the same excellent functionality found in the military. While knowing one's position with reasonable approximation (within tens of meters, or at least under 100m) is satisfactory for most marine or aerial navigation purposes, this level of approximation is not acceptable in other contexts. In an urban situation for example, the user may wish to know what side of the road he is on or which traffic lane to be in, and in the development of technical charts, cadastres, and engineering plans, the precise coordinates of the point on the ground are needed in order to represent it on a topographic map. In the latter case it is necessary to reduce approximation to within a few decimeters or centimeters. In reality, GPS is capable of reaching precision of less than a millimeter, if certain components of the technology are exploited and if the period of observation can be maintained for a sufficient length of time.

The precision of GPS measurements is also influenced by factors such as whether the station is mobile or fixed (at least for the time being) and whether the receiving station functions alone or in connection with other stations. The ability for a station to operate in conjunction with at least one other fixed station provides optimal precision. This technique is called Differential GPS and it is based on the registration of signals transmitted by the satellites and simultaneously picked up by a fixed or base station (Master), of which the precise geographical coordinates are known, and by a mobile station (Rover), located anywhere within a range of around 150km from the Master station. Post processing the data eliminates the errors in reception by the Rover. The precision of the GPS in the Rover station is able to satisfy all the requirements of civil use from cartographic applications, to support systems for the management and control of transport fleets, and personal mobility.

The late 1990s have seen an increase in the demand for differential correction and it is now easy to find

private and public companies that broadcast the data necessary for differential correction in their zone over the radio waves. Certain GPS models are already equipped to receive such radio signals. In the future GPS will continue to improve, in particular after the announcement by the US military controller of the system, with the launching of new satellites, of which 18 are planned in the near future, and with the refinement of certain technical characteristics. The future will also probably see full integration with the Russian-developed GNSS system, which is already integrated with GPS in certain commercial devices.

3. *The Relationship Between GPS and Information Systems*

The most relevant aspect of GPS to the information society in 2001 is the ability to provide the geographical coordinates of any object on the earth's surface and to increase our knowledge of the object's characteristics with ancillary data, collected in real time and directly from the field, at very low cost. Without doubt, the geographical component is an essential part of current databases and is generally deduced from a classification of objects by name (city, road, house number, etc.) or is represented by the cartographic coordinates of the single object gathered from maps (for example the centroid of a manhole cover, telephone or electricity pole, the axis of a water or gas pipe). In each case the geographical component, obtained by traditional means, is approximate, expensive, or unreliable in relation to both time and space. Time, because things can change from one moment to the next (an electricity pole can be moved relatively rapidly in reality but not so rapidly on the charts representing it). Space, because an element located in a certain territory undergoes continual transformation, the shores of a lake or a coastline for example can be modified by environmental factors.

Conferring a geographical address on each item of data allows for the application of the analytical, inquiry, and management functions of a geographic information system (GIS). Real time availability of the geographic address means that the spatial analysis can be carried out instantaneously in the field. Typical spatial analysis functions include: investigating relationships of proximity to, or distance from, the object, superimposition of information layers, the search for the shortest or any route that satisfies specific conditions, identification of spatial environments that group objects with specified characteristics, monitoring transformations of the object, and so on. Typical functions of inquiry analysis are those that allow us to understand the information linked to an object by selecting that object from a spatial database rather than selecting an object from the alphanumeric database and obtaining thematic charts *ad hoc*. Typi-

cal functions of information system management are those that permit us to continually update the constituent elements by refreshing the data and integrating the geographic database whenever necessary.

The continual information flux is one of the major problems in information society and updating coordinates in three-dimensional space is even more challenging. Connecting GPS to other techniques helps solve this problem. For example, we can consider digital cartography, necessary for a decision-making process in a territorial or urban context. This is archived in appropriate databases and territorial information systems use it to undertake analysis and to aid in making the above-mentioned decisions. But what happens when digital cartography no longer reflects reality? A city or region can sometimes change very rapidly and if the digital map is not updated, then we run the risk of the information system elaborating data that is no longer. Invalid information may be merely inconvenient or lead to serious social or legal complications. GPS helps greatly, but it is not only maps that need continual updating but also the geographic address of data that is particularly useful in social science. Among these are census tracts, political districts, administrative units, pollution zones, and epidemiological information, or research concerning traffic and movement. More controversially we may monitor individuals.

4. Examples of the Use of GPS and its Social Implications

The development of GPS and its diffusion is a step towards having geographic address (x, y, z coordinates in space, or longitude, latitude and height) and exact time as a component of any type of information. GPS equipment is already widely used by utilities companies for the identification of the components in their networks (poles, pylons, wells, technical components, etc.), while among the earlier applications we should remember the use of GPS for the positioning and identification of submarine oil wells from offshore oil platforms. Additionally anyone, at any moment in time, can visualize the location or destination of any vehicle transmitting its coordinates to be collected in real time by GPS. For example when mounted on a public bus, a GPS can supply immediate information on the vehicle's location, the journey time and anything else that is necessary to satisfy the needs of the driver, passenger, or manager alike. For some time now, insurance companies have been using this system to reduce the incidence of car thefts. Car manufacturers are now equipping cars across the range with navigators: electronic navigation systems containing a digital map of the area in which the car is travelling and on which the positioning and movement of the vehicle itself is traced, thanks to the continual tracking of its location through GPS. Integrated with other

systems, GPS permits a vehicle to proceed automatically, without the aid of the driver or pilot. Systems already exist in marine and aeronautical applications that are capable of functioning as autonomous autopilots. Exact details of military applications are not public knowledge, but it is easy to guess that the same basic technology is used with more sophisticated developments. An internal GPS in a cellular phone provides the user with their exact coordinates, and can be used to visualize the location of similarly equipped phones. Such devices are in regular use at large conventions to facilitate (or avoid) meetings. The same device alerts emergency services to incidents on lonely stretches of unfamiliar road or, worn as clothing, GPS may guide the sight-impaired or protect children against becoming lost. All fields that use geographic coordinates may derive benefits from the use of GPS.

There are two principal problems linked to the use of GPS: a reduction in the capability of orientation and the limitation of privacy.

The problem of the reduction in the capacity to orientate oneself is due in part to the use of digital coordinates, rather than analog information such as: go to the square, turn right at the newsstand, go straight on to the white building, etc. With GPS a person can supply a visitor with the coordinates of their position by simply pressing a button; the visitor receives the coordinates and transfers them to his autopilot or navigation system, and the vehicle navigator directs the visitor to the location they require. Nothing will remain in the visitor's memory, unless he has paid attention to the features of his route and not simply paid attention to the screen. It is a question of personal choice whether to learn to navigate in one's environment or to rely wholly on digital assistants.

The privacy problem is a crucial aspect of the use of GPS in society. The power of knowing in a given moment the location of someone equipped with a GPS can be a source of enormous security in certain cases; for example for those suffering from an accident or particular illness who may require treatment as quickly as possible. But knowing someone's movements is potentially a great intrusion on personal privacy. For example, with GPS and GIS we could discover the shop, office or facility someone is currently visiting, or recreate their activities during the past 24 hours or longer. The surveillant society is the flip-side of the information society. Separating valuable monitoring from unwarranted intrusion into personal privacy, and deciding who has the power to distinguish between them, poses a major problem for society.

5. Summary

Cartographers, engineers, planners, archaeologists, and geographers have drawn enormous advantages from the active use of GPS and will continue to enjoy

the benefits of a new technology that allows for rapid and much less costly collection and updating of geographic information. The integration of engineering designs realized with CAD, digital mapping, and GPS, introduces the possibility of devices capable of elaborating designs for civil engineering and proceeding through interactive surveys, to the construction of roads, buildings, etc. GPS also integrates well with the analysis of data that has been collected via satellite for the observation of the earth, in that it allows for the exact identification of the coordinates of specific features on the satellite image. Social applications range across all walks of life. Inevitably, as with all technological innovations, its use should be evaluated critically, and the limits of its appropriate use determined.

See also: Applied Geography; Cartography; Geographic Information Systems; Geographic Information Systems: Critical Approaches; Location: Absolute/Relative; Navigation; Remote Sensing and Geographic Information Systems Analysis; Spatial Analysis in Geography; Spatial Data; Time-Space in Geography; Transportation Geography; Transportation Planning; Urban Planning: Methods and Technologies

Bibliography

<http://www.trimble.com/gps/>
<http://giswww.pok.ibm.com/gps/>
<http://www.navtechgps.com/>
<http://www.colorado.edu/geography/gcraft/notes/gps/gps-f.html>

M. Salvemini

Globality

In the English language the word globality is of relatively recent coinage, although the noun 'globe' has been used rather a lot since the sixteenth century along with the adjective 'global.' The concept of globality has come to be employed widely since the early 1980s (Robertson 1984), along with the concept of globalization. Both globalization and globality became significant terms during the 1980s, in the wake of expanding scholarly use of the adjective global, in primary reference to the dual aspects of the intensifying connectivity (Tomlinson 2000) and of the spreading and deepening consciousness of the world-as-a whole (Robertson 1992, 2001a). This development

within intellectual discourse has proceeded in tandem with the ever more conspicuous invocation of the adjectival form in promotional and advertising contexts. Both these trends were largely sparked by the phrase 'global consciousness,' meaning consciousness of the world as a whole, in particular, but not sole, reference to environmental or ecological issues. Note must also be made of McLuhan's extremely influential conception of the global village (McLuhan 1964, p. 93, Gordon 1997). This was centered on the expansion of forms of communication since the invention of movable type and then the printing press in the fifteenth century up to global communication via radio and television, but prior to the World Wide Web, the Internet, and other by now commonplace features of the contemporary period of electronic communication.

Throughout most of human history there have been different civilizational conceptions of the scope, features, and conditions of the world. Scholars have, over thousands of years, speculated and theorized about the place of terrestrial earth in the cosmos. Of particular import in this regard was the period of the European Renaissance, an era of radical restructuring of philosophy, the arts, and the sciences in the fifteenth and sixteenth centuries (Cassirer 1963). A pivotal occurrence of this period was when the Polish scientist-philosopher Copernicus showed in 1512 that the earth was not the center of the universe, but that planets exist in a heliocentric universe. This perspective had truly major historical ramifications, in that the Renaissance as a whole, to put it all too simply, involved within Christendom an increasing valorization of *humanitas*, in contrast to previous worldviews that stressed the preeminence of God. The Renaissance view, well articulated by Cusanus (Cassirer 1963, pp. 7–45), did not, however, elevate humanity to the level of God. For Cassirer, the significance of Cusanus was in the latter's exposition of a move from the pre-Renaissance idea of a metaphysical world of transcendence to the Renaissance view of a scientific world of immanence. This shift represented not a triumph of the latter over the former, but rather a complex form of complementarity (Domandi 1963).

Closely intertwined in the European context with the crystallization of ideas concerning humanity were voyages and travels of 'discovery' to other parts of the world, particularly to Asia and to the Western Hemisphere. The 'discovery' of Africa, from the European standpoint, was not to come until a little later (Wills 2001, pp. 32–44). This expansion of the heliocentric world added a crucial spatial dimension to the more scholastic ruminations of leading figures of the European Renaissance, helped greatly by map-making becoming much more expansive, detailed, and accurate than had been the norm prior to that. It is of great importance to recognize the intimacy of the relationship between the motifs of humanity and globality, which were brought closely together during the Euro-

pean Renaissance and were to come even closer, up to and including, the European Enlightenment of the eighteenth century. This involved, again, an enrichment of the conception of humanity and an even broader and more concrete extension of the ideas concerning global variety. Consciousness of the latter was not often explicitly expressed via the specific concept of globality, this being a latent rather than a manifest theme.

Ever since the Moon landing of 1969, consciousness of what has come to be known as globality has increased. Space travel generally, as well as much of science fiction and the modern cinema, have greatly enhanced our tendency as human beings to gaze upon our global habitat, to become increasingly conscious of and sensitive to our globality. Even though this exact term may not yet have been used much, the ancient Greek notions of *oikoumene* and *okeanos* have remained central to conceptions of what are now considered under the rubric of globality. For example, the recent work of the Swedish anthropologist, Hannerz (1996, pp. 1–13), is centered on the concept of the global *ecumene* (*oikoumene*), a term which he has borrowed from the cultural anthropologist, Alfred Kroeber. Kroeber (1945, p. 9) saw the scope of anthropology to be the global, ecumenical circumstance, suggesting that humanity at large consists of interwoven relationships among a variety of societies and civilizational regions, with particular reference to culture.

In spite of this long etymological history, had it not been for the enormous and still growing degree of attention given to the contested concept of globalization (Robertson 1992, 2001a, 2001b), the notion of globality would not have acquired the contemporary significance that it now has. The discussion of globalization in a considerable number of social sciences and humanities has led to increasing talk about our now living in a global age (e.g., Albrow 1996), even though, as has been stressed, the philosophical, metaphysical, and theological interest in this topic, in a diffuse sense, dates back to ancient speculation about what we would now call the human species and the shape, configuration, and place of the earthly world in the universe.

In fact the phrase, 'the global age,' has become a common, although not unproblematic, characterization of the present condition of humankind. In this regard, there is a controversy in a number of disciplines concerning the relationships between such themes as tradition, modernity, postmodernity, and globality (Robertson 1992, pp. 138–45). Indeed, it is not fully possible to delineate the contours of globality without situating this concept, in contemporary terms, within this array of themes. Likewise, it is not possible to neglect the issue of locality in relation to globality, largely because the two have often, although very simplistically, been viewed as opposites, within the debate about globalization and related issues in recent

decades of so-called modernity. In fact the relationship between globality and modernity is particularly controversial, basically because some have claimed that the process of globalization and the consciousness of the condition of globality have been a consequence of modernity (e.g., Giddens 1990), while others (e.g., Robertson 1992, 2001a) have insisted that globalization was a precondition for the rise and spread of different types of modernity in, or to, different parts of the world.

Globalization as an analytic concept first began to be used rather widely in sociology, religious studies, and, and to a lesser extent, anthropology in the early 1980s, although the theme of world, or global, history had been thematized in different civilizational contexts much earlier in the twentieth century than that, falling, however, into temporary decline from the 1920s until the 1980s. The concept of globalization came to be employed in a truncated and increasingly politicized way in some social-scientific, political and journalistic circles, but particularly in the field of business studies, from the late 1980s onward (Robertson and Khondker 1998, Robertson 2001a). This narrowing of the concept of globalization has had the consequence of pushing the theme of globality into the background. Even so, even were this circumscription to be comprehensively embraced, the theme of globality would not thereby be eliminated. This is so in spite of the fact that many academics, not to speak of the 'informed public,' do not appreciate that the concept of globalization was being debated quite extensively before it was more or less independently taken up in business studies, policy-making, and ideological forms. The theme of globality has very little to do in any significant respect with conceptions of globalization that limit it primarily to economic change and politico-economic policy. But the notion of globality is not necessarily marginalized as a consequence. The great attention that has been given to human rights and related concerns since the late 1970s helps to demonstrate this. For the virtually worldwide interest in and controversy about human rights have become closely bound-up with concrete matters of politicoeconomic policies.

When the term globality is explicitly employed, it usually refers to an historical circumstance of human life, and it is the extent to which this circumstance marks a new phase in the history of humankind as a whole that is particularly important, as well as problematic. More specifically, it may be defined as the intensification of more-or-less worldwide connectivity and increasingly reflexive global consciousness (Beck 2000, Robertson 1992). Overall, the condition of globality is accentuated greatly by the compression of the world. To this definitional statement must be added that, in a simply objective sense, this worldly planet was as global millennia ago as it is now. In this regard the world has always existed in-itself, but over many centuries it has become more and more for-itself, this meaning that as human beings we

have become more and more conscious of and have acquired doctrines and views about the future of the world.

Looking at this from the opposite direction, the notion of globophobia (Burtless et al. 1998) may be invoked. Globophobia, a term coined originally by a group of economists to characterize resistance to the opening of national economies to free trade and globewide economic competition, may be defined, more generally, as a sentiment involving resistance to globewide, more loosely what are perceived to be foreign or alien, phenomena and trends. Such trends are regarded as threatening the integrity and primordality of a particular collectivity. Globality is in this sense a form of complexity, a complexity that is seen by some as going beyond the complexity of what has usually been called modernity—or, more recently, postmodernity. For example, in the USA, so-called Christian fundamentalists have spoken vociferously of the evil of (American) secular humanism. In more recent years, however, the enemy has become secular globalism (Bowen 1984). Even though parallel conflicts are to be found in many societies around the world, that between the self-proclaimed antiglobalists and those who subscribe in one way or another to a much more open society within the USA is, given the paramouncy of the USA in the world, of unavoidable—indeed global—significance (Robertson 2001b).

The perceived threat of globalism as a conspiracy against core American values has intermittently been, but particularly in recent years, a conspicuous feature of political discourse in the USA. Many societies have traditions of belief in conspiracy, but the USA has been one of the most prominent in this respect. This has partly been a response to the opposite trend—to be found in many American educational circles—involving the advocacy (often, but by no means only, for economic reasons) of teaching about other cultures and languages, as well as the condition of the world as a whole. This leads to a much more general—indeed, global—issue concerning the ways in which, to greatly varying degrees, people experience existential difficulty in being confronted by alternative ways of life. This is often encapsulated in the idea of alterity, or otherness. Rigid rejection of ‘the other’ is to be seen most clearly in what has come to be declared by some as fundamentalism (Robertson 1995). Thus, when normatively embraced, globality involves an eager acceptance of the idea of the global-human condition. At the other extreme, fundamentalism entails adamant rejection of the latter, at least insofar as one confines the concept of fundamentalism to specific societies or localities. A complication here, however, is that there is a form of global fundamentalism, one that stresses the priority in everyday life of globality and humanity over such aspects as international relations, features of particular societies or the life-cycles and identities of individuals.

The quotidian problem of alterity arises primarily from processes of relativization. Relativization involves the situational problems arising from the challenges posed to a particular way of life by the problematic realization that there are other, alternative ways of life. In a world marked by its high degree of connectivity, confrontations between different forms of life are ever-increasing. For, as the world becomes more connected and people around the globe become increasingly aware of both their differences and their shared fate, the likelihood of dilemmas and ambivalences concerning ‘others’ greatly intensifies. Indeed, probably the most general feature of long-term globalization, over many centuries, has been the persistence of relativization processes.

Although the notion of globalism is often used, even by mainstream academic contributors to the discourse of globalization, in its broad, comprehensive—as opposed to its narrow economic—sense, it is misleading so to do. Specifically, the sentiment, or even ideology, of globophobia often involves the presentation of ‘the enemy’ as the ideology of globalism or what is also often called, in a pejorative sense, one-worldism. Thus, even though some social scientists use the idea of globalism with the best of intentions, that is, to indicate their or others’ cognitive and moral commitment to a global perspective, they nonetheless unwittingly engage in the rhetoric of resistance to the idea that we now live in a condition of shared globality. So globality is a much more preferable concept in relation to much, but by no means all, of the talk about globalism. The two terms have very different connotations, although they are often used interchangeably, with subsequent confusion. Globality indicates—for the most part—a condition, or circumstance, while globalism is an ambiguous term, ranging in its meaning, on the one hand, from the relatively innocuous expression of adopting a so-called global perspective to a pejorative, condemnatory declaration, on the other hand, of the evils of subscription to the view that there is a general, inclusive conception of humanity and that, moreover, we all, every single human being, live in and share a single, global place (Robertson 1992).

When we read or make such statements as ‘we now live in a global world,’ ‘the world is my home,’ or ‘we must take the feeling of home into exile’ (see Iyer 2000), we are drawing attention to global complexity and at the same time welcoming it. This is indeed a personal and seemingly positive kind of globalism, but at the same time, from the analytic standpoint, such attitudes are merely symptomatic of the condition of reflexive globality. The question necessarily arises as to whether these orientations are merely to be found among certain kinds of intellectual or more privileged people, such as those who flit from one habitat, or habitus, to another. Thus it is not unreasonable to ask the question as to whether it is only certain global elites who dwell in and welcome the circumstance of

globality. However, in many ways it is the world's most deprived peoples, not to mention those, such as students, who move back and forth between 'tradition' and 'modernity,' that are most reflexively conscious of the global condition. Moreover, the everyday life of people all over the world, even though they may be relatively, or even absolutely, immobile geographically is becoming increasingly global, not least through representations of 'other cultures' in movies on cable or satellite TV, through popular culture generally, or everyday interaction, including electronic communication. In other word, vicarious, virtual interactional communication is increasingly part-and-parcel of everyday life. Even though celebrities, business elites, and some academics are increasingly conscious of the condition of globality, they do not in fact have a privileged position in this respect; for in one way or another people all over the planet earth are evidently becoming more and more 'global.' While it has to be conceded that there is much resistance all over the world, at the beginning of the twenty-first century, to strangers, it has been argued persuasively that this, often violent, opposition to polyethnicity and poly-religiosity (in the symbolic form of ethnic cleansing), constitutes the dusk of the norm of homogeneity and the contested return of what has been described as the polyethnic norm (McNeill 1986).

Whether 'the average person' is or is not fully conscious of their 'global membership,' many 'average people' nonetheless participate in increasingly reflexive global consciousness. The development of such has been proceeding for a very long time, at least since the crystallization of the so-called world religions in what has been called the Axial Age (Jaspers 1953), for example, through pilgrimages, world map projections, voyages of 'discovery,' and so on. In much more recent times such a variety of phenomena as two World Wars, the Cold War, the European and other holocausts, the global AIDS pandemic, awareness of global ecological problems (notably global warming), the World Soccer Cup and other world championships, rapidly burgeoning touristic travel, global festivals, and so on and so forth—the list could be vastly extended—amply illustrate this.

While, in its most simplistic formulation, globality may be interpreted as referring to planetary territoriality and the beings (including nonhuman forms of life) that inhabit this earthly place, it is increasingly clear that globality involves and indicates variety and difference with respect, *inter alia*, to gender and sexuality, 'race' and ethnicity, nationality, and so on. Moreover, each of these, and other, attributes of human beings have become increasingly elective. In our time, to varying degrees, one choses one's body, one's 'roots,' one's gender, one's ethnicity, one's nationality, even one's own form of globality. For some (e.g., Beck 1999) we now live in a global risk society, a world in which we are continuously confronted with a multitude of choices, choices that we

make without the firm norms and values of tradition and in the absence of secure agencies of authoritative trust. In short, our planetary circumstance is 'a runaway world' (Beck 1999, 2001).

When the idea of the global-human condition is invoked, the aim is to underline that we would not now be talking about globality were it not for the connection of this theme with that of humanity. The latter certainly has had a very long, although until fairly recently, rather abstract, history. But since the European Renaissance and particularly since the European Enlightenment, the theme of humanity has become increasingly concretized. The intersection of the concept of humanity with that of globality is to be seen with particular clarity in the French philosophical and sociological tradition from the early nineteenth century through the first quarter of the twentieth century. This was a period during which theological ideas were widely rejected, but also one when a new kind of, or surrogate, religion was deemed to be required. For example, a disciple of Saint-Simon wrote that 'there is in fact only one branch of knowledge, that of humanity, which includes everything and epitomizes everything' (Merle 1987). Saint-Simon's followers elaborated his thinking in a new journal, the *Globe*. August Comte, who had been Saint-Simon's secretary, was to advocate a religion of humanity in his later years. In sum, major French writers of the first half of the nineteenth century maintained not merely that globality and humanity were complicitous, but that a new form of social-science (sociology) inevitably arose from this connection. In the late nineteenth century this relationship was further elaborated by major sociologists such as Emile Durkheim and Marcel Mauss.

Since the 1960s there has been a strong predilection to conceive of 'the local' as the opposite of 'the global'—as reflected in the maxim 'think globally, act locally.' However, recent demonstrations in various countries against global institutions—such as the World Bank, the International Monetary Fund, and the World Trade Organization—have clearly involved a severe modification of this maxim. Protesters against global institutions have come to recognize that global targets require global movements (Robertson 2001a, 2001b). This is probably the most obvious way in which the condition of globality has become inexorably central to the life of many people across the globe. For ostensibly, antiglobal movements have, paradoxically, become part of the process of globalization itself.

See also: Global History (Universal History, World History); World Systems Theory; Globalization and World Culture; Globalization Anthropology of; Globalization, Geographical Aspects of; Globalization, Legal Aspects; Globalization, Political Aspects; Globalization Subsuming Pluralization, Transnational Organizations, Diaspora, Postmo-

dermity; International Organization; International Political Economy; International Trade: Economic Organization; Local-Global Links, Environmental; Capitalism: Global

Bibliography

- Albrow M 1996 *The Global Age: State and Society Beyond Modernity*. Stanford University Press, Stanford, CA
- Beck U 1999 *World Risk Society*. Polity Press, Cambridge, MA
- Beck U 2000 *What is Globalization?* Polity Press, Cambridge, MA
- Beck U 2001 Living your own life in a runaway world: Individualisation, globalisation and politics. In: Hutton W, Giddens A (eds.) *On the Edge: Living with Global Capitalism*. Vintage, London, pp. 164–74
- Bowen W M 1984 *Globalism: America's Demise*. Huntington House, Shreveport, LA
- Burtless G, Lawrence R, Litan R, Shapiro R 1998 *Globophobia: Confronting Fears About Open Trade*. Brookings Institution Press, Washington, DC
- Cassirer E 1963 [1927] *The Individual and the Cosmos in Renaissance Philosophy*. University of Pennsylvania Press, Philadelphia, PA
- Domandi M 1963 Translator's introduction. In: Cassirer E (ed.) *The Individual and the Cosmos in Renaissance Philosophy*. University of Pennsylvania Press, Philadelphia, PA, pp. vii–xii
- Giddens A 1990 *The Consequences of Modernity*. Polity Press, Cambridge, MA
- Gordon W T 1997 *McLuhan for Beginners*. Writers and Readers Publishing, New York
- Hannerz U 1996 *Transnational Connections: Culture, People, Places*. Routledge, London
- Iyer P 2000 *The Global Soul*. Bloomsbury, London
- Jaspers K 1953 *The Origin and the Goal of History*. Yale University Press, New Haven, CT
- Kroeber A 1945 The ancient Oikoumene as an historical cultural aggregate. *Journal of the Royal Anthropological Institute* 75: 9–20
- McLuhan M 1964 *Understanding Media*. McGraw-Hill, New York
- McNeill W H 1986 *Polyethnicity and National Unity in World History*. University of Toronto Press, Toronto, Canada
- McNeill W H 1992 *The Global Condition*. Princeton University Press, Princeton, NJ
- Merle M 1987 *The Sociology of International Relations*. Berg, New York
- Robertson R 1984 Interpreting globality. In: *World Realities and International Studies Today*. Pennsylvania Council on International Education, Glenside, PA, pp. 7–19
- Robertson R 1992 *Globalization: Social Theory and Global Culture*. Sage, London
- Robertson R 1995 The search for fundamentals in global perspective. In: van Vucht Tijssen L, Berting J, Lechner F (eds.) *The Search for Fundamentals and Modernization: The Process of Modernization and the Quest for Meaning*. Kluwer, Dordrecht, The Netherlands, pp. 239–62
- Robertson R 2001a Globalization theory 2000+: Major problematics In: Ritzer G, Smart B (eds.) *Handbook of Social Theory*. Sage, London, pp. 458–71
- Robertson R 2001b Anti-global religion? In: Juergensmeyer M (ed.) *Global Religions*. Oxford University Press, Oxford, UK

Robertson R, Khondker H H 1998 Discourse of Globalization: Preliminary Considerations. *International Sociology* 13(1): 25–40

Tomlinson J 2000 *Globalization and Culture*. University of Chicago Press, Chicago

Willis J E 2001 *1688: A Global History*. Granta Books, London

R. Robertson

Globalization and Health

Recent revolutions in communication have resulted in the development of a global village. Isolated areas of the world have 'joined' the technologically sophisticated world and developed, and sometimes, parochial regions have been increasingly faced with the realities of life in less fortunate parts. Growing awareness of the disparities in lifestyles and health around the world are forcing an increasing recognition that global health policies and practices are a need, not for the future but for the present.

Health care in this increasingly disparate community must shift from the high-technology focus, which has gained predominance in technologically sophisticated societies, to incorporate broader social, cultural, and psychological concerns. For example, it is probably of far greater importance in the fight against disease to provide clean water in most parts of the world than it is to even provide doctors. Concerns with malaria, TB, AIDS, and HIV dominate global health while poverty, discrimination, inadequate education and environmental contamination provide fertile growth for even greater social epidemics. Health-care disparities are extreme with most of the world's health-care money, interest, and attention being focused on those with the least need. Ultimately, however, health care in both the developing and the developed world are interrelated and setting priorities as to how these issues can be addressed with mutual benefit is needed.

A number of themes characterize health in the global environment in the past and future decades. These will be dealt with in turn.

1. Technological Development

Health care in the industrialized world has mirrored the dominant ethic of the twentieth century: technology development. The remarkable advances in technology that have resulted in a rapidly changed and changing environment at the close of the millennium have also brought with them increasingly complex challenges.

As technology is developed, so too is the commercial basis for its adaptation and use. Overuse of technology

in health care has been the result, with new advances being applied for the benefit of all rather than only the few for whom the technology is specifically directed.

On the other hand, technology is not readily available in less affluent societies. We need to ask whether it is acceptable to have a two-tiered system of health care with regard to appropriate use of technology for developing and developed countries? How well have we addressed issues of equity regarding access to appropriate medical care? How do financial considerations influence who receives differing types of care? To what extent does remuneration of health care providers and/or health care services influence the quality of care provided or the use of available technology? How well are we monitoring the appropriate use of technology and the medical model worldwide? Are the currently available indicators of health sufficient to assess the influence of this level of functioning or are alternative indicators needed to address this issue? How should we evaluate the appropriate use of technology worldwide?

2. Evidence-based Medicine

Evidence-based medicine is one response to emerge in reaction to the technological explosion. Demanding that technology be applied when, and only when, it is effective, and ensuring that the strictest scientific rigor be applied to judge such effectiveness is leading caregivers to question the overuse of technology (Cochrane Collaboration 2000, Entin et al 1996).

We have developed a computer-based globe, moving increasingly into cyberspace in addition to its more earthly properties. Health and health care have become enmeshed in this web so that medical information and knowledge are moving more readily into the hands of the layperson than ever before. Medical practitioners are having to share their knowledge with their clients, raising issues such as shared decision making, collaborative care, information exchange, confidentiality, and informed choice by recipients of care. Decision making regarding health care, by both professionals and by families, needs to be evidence based: a fair challenge to achieve. Central to these developments are also legal and ethical issues which surround the management of health care and influence health-care decisions. Health-care services, to cope with these demands, need to move from a physician-focused service to a client-centered approach.

3. Family-centered Care

The client too is changing and is being viewed not just as a single entity but as a member of a family unit (whatever that may be), so that care is becoming family-centered care (Chalmers and Levin 2001). The family itself is changing rapidly too, meaning that

health care is directed to anything from nuclear, extended, or single-parent families, or to same-gender couples. The traditional role of the doctor-patient relationship is becoming chameleon-like in its complexities.

4. Psychosocial Health Care

The need to incorporate holistic components of care into health has never been more strongly felt than now, although this is not a new idea. As long ago as 1948, the World Health Organization's Constitution (World Health Organization 1948) defined health as 'a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity.' It is far preferable to care for the client as a whole person rather than as separated fragments. In earlier medical texts, support was often expressed regarding treating the 'patient as a whole person with thoughts, feelings, and fears, and not only as a body progressing through the health care system.' Unfortunately, in technologically sophisticated societies, superspecialization also occurs. Physicians no longer care for all aspects of a client personally: For example, in reproductive health, some specialists are focused on care of the cervix, others on the endometrium, ultrasound diagnosis is delegated to specialized technicians, and even such simple procedures as taking blood are relegated to yet other caregivers. Research has documented that as many as 16 different caregivers in a 6 hour period (Enkin et al. 1996) may look after a woman during a normal birth. A holistic or psychosocial approach to care hardly exists in this structure.

Assessing people's perceptions of their satisfaction with care is one approach to monitoring the integrity of their health care. Feedback regarding educational programs is well accepted in technologically sophisticated societies. No conference ends without a delegate evaluation being completed. Why is this not similarly integrated into health-care services? Is it because the clients' voices are still not acknowledged as an integral part of healing and are weighted as of far less significance than the biological basis of their illness? Becoming partners in care, rather than simply clients, is still the challenge for the future. 'Humanizing' the medical setting has become a clichéd term, yet it is still needed.

5. Cultural Differences in Understanding of Health and Illness

Cultures differ widely in their beliefs and understanding of health, illness, and appropriate health care. Many regard the biological level of functioning as of primary significance. In contrast, significant sections of the global village regard this level of functioning as of least importance with spiritual, social, and psycho-

logical concerns playing a primary role in illness etiology, treatment, and prognosis (Chalmers 1990). It is as unacceptable to the latter to neglect consideration of spiritual rites in diagnosis and treatment of illness as it would be to fail to take a woman's blood pressure during pregnancy in technologically sophisticated societies.

While these differing approaches to care have existed peaceably side-by-side for centuries, recent global population movements have encouraged the intermingling of cultures through migration, refugee relief and more temporary forms of translocation, with resulting conflicting health care paradigms. Compounding the problem are language differences combined with status inequalities which may make it extremely difficult for the foreigners' health care needs to be met or, indeed, understood in their adopted country. As multicultural societies continue to develop rapidly in the coming decades, the challenge of integrating their health care needs will become increasingly evident.

Traditional practices of healing should be considered with sensitivity and tolerance, and encouraged unless evaluated as distinctly harmful. Condemning or disallowing such practices because they are different is insufficient for although they may not contribute to biological health, they might be important for psychological health and should be encouraged. They should be evaluated as biologically beneficial, inconsequential, or harmful with only the latter forms of care being discouraged.

6. Education

Education to prepare for a multicultural, psychosocial, and evidence-based approach to health is a prerequisite in the twenty-first century. A number of medical institutions have introduced teaching psychosocial issues to medical and allied students. For the most part these have failed. Frequently based in Social or Behavioral Science Departments, they tend to be offered in preclinical years and to focus on theory. They are often perceived by students as programs to be endured, or which are undemanding, and not central to their ability as future caregivers. When offered at a graduate level, as the real challenge of caring for people is encountered in clinical practice, their worth is sometimes reconsidered.

Rather than being offered as distinct, independent course modules, the psychosocial and cultural issues involved in clinical practice should be integrated into mainstream medical teaching programs, e.g., when clinical management of a fetal death is discussed the psychological impact of diagnostic and treatment procedures in addition to longer-term family adjustment issues should be considered. Teaching should be shared by the physician and social scientist concurrently and such joint teaching programs, when tried,

have shown much promise of a new and better model to develop for educational success (Chalmers and McIntyre 1993). In addition, providing an integrated model of teaching in the classroom will also establish a model for integrated clinical practice.

It is evident that health science faculties will need to undertake radical curriculum reforms in the twenty-first century to incorporate new, and fundamental, principles of education into their teaching. At the very least, these reform programs need to incorporate evidence-based medicine, family-centered care, multidisciplinary approaches, and psychosocial and multicultural aspects of care. Educational institutes need to be structured in such a way that facilitates the advocacy of both medical and social models of health care. How many medical schools teach social and cultural components of care as an integral component of medicine and how many simply pay lip service to this? Do we even know?

7. Efficiency of Care

A significant contributory factor to unease about care are the lengthy delays often encountered when seeking consultations with specialists, diagnostic assessments and their results, and treatment. Once a problem has developed, anxiety levels are high, particularly when fear of severe illness (such as cancer) or potential surgery becomes a possibility. Yet people frequently have to wait weeks or longer for simple diagnostic procedures such as ultrasound and even longer for more sophisticated tests. Further delays before receiving the outcomes of tests compound the problem. It is not surprising that frustration, fear, and helplessness are aroused with concurrent irritation and suspicion of the health-care services.

We frequently use the term 'consumers' to describe health-care clients. If customers in commercial stores were treated with similar delays in service and impersonal attention they would simply shop elsewhere. Within the health care service, with its more threatening and intimate experiences for clients, and its lack of real alternatives from which to choose in many instances, there is reluctance to apply a similar model. This does not mean that all is well in the state of health care, however. That dissatisfaction exists is evident in the increasing reality of litigation.

8. Conclusion

This article highlights some of the critical issues surfacing in the health-care services of the developed world from a broader global perspective. These kinds of concerns underscore the importance of those long-held values: concern for the whole person and not only the body, respect for people (and their families) and for their perceptions and needs when undergoing

medical care, a sincerely gentle and caring approach to care with sensitivity rather than routine politeness being the order of the day combined with evidence-based practice. It is apparent that while we have long recognized the importance of providing respectful treatment and care and have gone far in introducing systems which encourage such approaches, we are still grappling with the effective application of this fundamental tenant in many parts of the world, both developed and developing.

See also: Ecology and Health; Health Education and Health Promotion; Health in Developing Countries: Cultural Concerns

Bibliography

- Chalmers B 1990 *African Birth: Childbirth in Cultural Transition*. Berv Publications, River Club, South Africa
- Chalmers B, Levin A 2001 *Humane Perinatal Care*. Tea Publishers, Tallinn, Estonia
- Chalmers B E, McIntyre J A 1993 Integrating psychology and obstetrics for medical students: Shared labour ward teaching. *Medical Teaching* **15**: 35–40
- Cochrane Collaboration 2000 The Cochrane Library. Update Software, Oxford, UK
- Enkin M, Kierse M, Renfrew M, Neilson J 1996 *A Guide to Effective Care in Pregnancy and Childbirth*. Oxford University Press, Oxford, UK
- World Health Organization 1948 *Constitution of the World Health Organization*. World Health Organization, Geneva, Switzerland

B. Chalmers

Globalization and World Culture

1. Globalization and the Construction of World Culture

Globalization involves expanding worldwide flows of material objects and symbols, and the proliferation of organizations and institutions of global reach that structure those flows. World culture refers to the cultural complex of foundational assumptions, forms of knowledge, and prescriptions for action that underlie globalized flows, organizations, and institutions. It encompasses webs of significance that span the globe, conceptions of world society and world order, and models and methods of organizing social life that are assumed to have worldwide significance or applicability.

While many types of global flows have been rising cyclically for centuries, it was only in the second half of the nineteenth century that a transnational cultural

complex took a sufficiently organized form to constitute an emerging world culture. The primary locus of this nascent world culture was Europe, particularly the powerful Western European countries whose empires extended to most corners of the globe. Europeans promoted political and philosophical principles, societal and individual goals, modes of organizing, and ways of conceiving and manipulating reality that they deemed universally applicable. Epitomized above all by technical and scientific principles and practical knowledge presumed to be invariant across time and space, emerging world culture also included more historically bound constructs and ideologies, such as nationalism, citizenship, and individualism. This early version of world culture, more properly called ‘transeuropean’ culture, was carried far and wide by missionaries, traders, military expeditions, colonialists, intellectuals, and travelers.

In this same period, transnational organizations and global structures emerged with increasing frequency, eventually to form a structural backbone or framework for world culture. The vast majority of these transnational organizations were products of international or global ‘civil society’—voluntary associations founded and operated by individuals from many countries to pursue specific goals through democratically coordinated action. Typical early examples include the International Charity Association (1855), the International Sugar Union (1864), the Scandinavian Dental Association (1866), the Permanent International Committee of Architects (1867), and the International Meteorological Organization (1873). By the 1890s, such international nongovernmental organizations (INGOs) were appearing at the rate of more than 10 per year, across a wide range of social sectors, drawing participants mainly from Europe and North America but also from Latin American and some Asian countries, particularly India. These bodies defined themselves as global actors and sponsored periodic conferences at which universalistic issues, problems, methods, and solutions were proposed and debated. They came to constitute a formalized global public realm in which world culture was defined, documented, elaborated, and propagated to what the growing number of participants in this public realm were beginning to think of as a single world society.

The calamitous world wars of the twentieth century severely interrupted world-cultural structuration, but after each war the process rebounded quickly. States became increasingly engaged in transnational cooperative relationships through intergovernmental organizations (IGOs, which were rare before 1920). After World War II, the IGO population burgeoned to a total of several hundred organizations, while the number of INGOs soared into the thousands. This expanding complex of global organizations came to center on the United Nations, whose agencies and programs became axes of global governance

regimes in such major institutional areas as education, health, and development.

Until the 1970s, the content of world culture remained primarily Anglo-European, increasingly dominated by the USA. American popular culture and products proliferated throughout the globe—rock music, soap operas, bluejeans, Coca-Cola, fast foods, Hollywood movies. The globalization of tastes in food, dress, and music also promoted a global identity model, that of the freely choosing, pleasure-seeking consumer. At a deeper level, the plethora of global organizations, both nongovernmental and intergovernmental, produced a range of standardized principles, models, and methods for the organization of social life. For example, a basic world model of the modern nation-state crystallized, stipulating that the state was to be responsible not only for internal order and external defense but also for building a modern society and promoting citizen welfare and civic engagement. States were therefore expected to generate national educational systems, health care programs, retirement and unemployment insurance schemes, and cultural promotion programs, among many other duties. In particular sectors, globally legitimated models emerged to guide states in meeting these responsibilities, such as the standardized educational models that took shape under UNESCO's aegis in collaboration with a range of educational INGOs. Despite enormous variation in national circumstances and resources, states have implemented these models in remarkably uniform ways.

By the 1980s, world-cultural structuration had produced largely standardized global models for an enormous range of activities—in science, medicine, health, business, technology, even recreation and leisure (sports, tourism, entertainment). Pushed by the carriers of these models, including INGOs and IGOs but also business consultants, technical advisers, and academic consultants who explain and interpret the models, states, local governments, corporations, social movement organizations, and many other types of social units go to considerable lengths to keep up with the leading edges of world-cultural development. The net result is a tremendous force for the homogenization of the world.

2. Globalization and Cultural Differentiation

World culture is not only a homogenizing force; it also engenders and supports diversity and differentiation. Recognition of this feature of world culture was slow in coming; through the 1970s, most analysts interpreted globalization as essentially equivalent to homogenizing Americanization. Five factors are important for understanding world culture's promotion of heterogeneity; ironically, several of these are also important elements of world culture's homogenizing capacity.

(a) *Success of the nation-state political form.* Rapid decolonization after 1945 produced a world organized almost entirely as independent states. Most new states have eagerly joined global governance organizations, debating and helping to shape agreements expressing world-culture principles and prescriptions. Citizens of the new countries became avid joiners of INGOs, expanding the range of their memberships much faster than citizens of older countries. Thus, the voices and views of the Third World are increasingly prominent in world-cultural development, and debates about proper social organization and action are often less consensual than in the past. Cultural controversy has been especially acute in such domains as women's rights, environmentalism, and human rights.

(b) *Cultural relativism and the ideology of cultural authenticity.* Intellectual movements in the social sciences and humanities, coupled with the ideologies of nationalism and national self-determination, have made the principle of the fundamentally equal value of all human cultures a central assumption of world culture. Ethnocentrism has come to be seen as both trap and injustice; tolerance and, indeed, the championing of difference occupy the moral high ground. Of particular moral virtue in contemporary world culture are the poor, the excluded, the oppressed—marginal peoples whose right to their own cultures has been violated by the onslaught of globalization (some see this as a return of 'noble savage' idealization). This universalistic form of particularism impels peoples to emphasize or invent tradition and distinctiveness in counterpoint to universalistic world-cultural principles that are supposed to operate uniformly in all places. In Western societies, such particularism is stimulated by cosmopolitan connoisseurs seeking authentically exotic cuisine or ethnically distinct 'world' music, as the genre is known. At the same time, politically astute groups understand the power of purported cultural authenticity as a fulcrum for prying rewards from global systems, so distinction and difference have become strategic resources for collective mobilization. In consequence, ethnonationalist movements of many sorts (Basques, Biafrans, Bretons, Croats, Eritreans, Quebecois, Scots, Tamils, etc.) and multiculturalist restructuring by states, churches, schools, and other organizations have become the order of the day.

(c) *Regionalism.* Roughly half of the international organizations founded since 1950 have been regional in scope, activating European, Latin American, Asian, francophone, Islamic, Andean, and many other sub-global identities. Like ethnonationalism, regionalism has flourished in the wake of world-cultural intensification. Global structures, ideologies, principles, and models provide an overarching framework of commonality and shared meaning for disparate social units. With the framework well in place, diverse regional organizations and movements have expanded to implement, modify, and argue about the frame-

work's content and implications. We see the same process at work within international organizations and global corporations. For example, Cable News Network International (CNNI) began as a unitary organization bringing a standardized English-language product to televisions around the world, but it eventually established numerous regional centers producing local-language broadcasting adapted to specific target audiences.

(d) *Consumerism as adaptive interpretation.* While global popular culture contains products and symbols that penetrate most local markets, its accompanying ideology of economic freedom and consumer choice encourages varied uses and interpretations of standardized consumption patterns. For instance, consumers in Hong Kong might appreciate a McDonald's restaurant as much for its clean facilities as for its food. They expect fast, not smiling, service, and reserve the right to eat at their leisure. In a city pressed for space, the restaurant can become a gathering spot for youth or a favored locale for birthday parties. By adapting marketing to local attitudes, McDonald's has lost its foreign aura, blending into Hong Kong's diverse, cosmopolitan culinary scene. In Beijing, meanwhile, McDonald's American origins can be attractive to upwardly mobile Chinese as a symbol of participation in global society, but among many Koreans in Seoul those same origins evoke criticism. While not denying the power of multinational corporations to foster acceptance of their products and logos, recent research suggests that they must take into account the distinctive tastes of potentially fickle consumers. Popular culture increasingly creates global tastes, but it also creates critical consumers who keep this culture from becoming entirely uniform.

(e) *Creolization.* In many domains, the homogenizing force of world culture encounters ingrained local tastes and traditions, which may in turn be the residue of earlier such encounters with neighbors or colonizers. Due to creative interaction processes, common elements take different forms in different places. The spread of world culture therefore produces not homogeneity but new cultural melanges, each 'creolized' in its own way. A case in point is West-African popular music. Popular styles that themselves are blends of variegated traditions (Brazilian, blues, British rock) become the ingredients of new and distinctive styles. Adapting indigenous instruments or rhythms, African musicians participate in a global musical movement. Cultural flows come from many directions, intersect in unpredictable ways, and continue on in new forms to yet other places. How they continue, and the publics they eventually reach, depend in part on the global commercial media system's marketing and genre-promotion efforts, but that system is only one factor shaping local aesthetic orientations. Moreover, commercial interests have much to gain from promoting diversity. Thus, as new musical or other fads make their way across the globe, they produce new forms of

localized diversity and unforeseen local reactions and interpretations. In the expressive realm, the varied forms of local-global interaction are not inexorably yielding to a standardized, stifling, hegemonic world culture.

3. Dimensions of World Culture

Most discussions of world culture focus on its expressive or normative dimensions. Expressive (popular) culture includes media products, consumable goods that assume iconic status as symbols of modernity or avant-gardism, foods and clothing styles originating in particular cultures that become worldwide fads, and so on. Normative culture comprises the values and goals promoted by global corporations and organizations—fetishistic consumerism and capital accumulation, individualism and human rights, democracy and political participation, gender and racial equality. The expressive dimension has received the greatest attention, particularly with respect to the global popularity of American films, television shows, sports figures, and lifestyle elements. Some scholars decry the obliteration of local cultures; others argue in favor of creative interpretation and adaptation, as local cultures integrate new cultural elements while retaining their basic identities (hybridization or creolization). Similarly, analysts decry some aspects of normative cultural globalization and praise others, often sharply disagreeing about the desirability of particular global norms and values and disagreeing as well about the extent to which a global normative consensus is emerging or even possible.

Less attention has been paid to the ontological and cognitive dimensions of world culture. At the ontological level, the reorganization of social life in accordance with dominant world-cultural models engenders increasing individualism, the disenchantment of nature, reliance on rationalized images of society, and an action-oriented model of the individual. Put another way, world-cultural penetration changes prevailing views of 'the nature of things,' promoting the worldview that underlies such major modern institutions as formal education, experimental science, national accounting and statistical systems, and advanced-technology engineering. These and similar institutional complexes bring with them broad bodies of knowledge and distinctive cognitive styles, but the depth of cultural reconstruction they engender remains largely unexamined because few attempts have been made to come to grips with their cultural implications. Science, mathematics, bookkeeping, organizational structure, technology, tools, construction materials, land surveying—all these have achieved the status of almost unassailably universalistic activities, uninteresting because they are mundane, routine, matter-of-fact. Thus, these highly formalized arenas seldom become the focus of ideological debate,

with the notable exceptions of such issues as nuclear power, genetically modified organisms, global warming, and large dams. The most highly institutionalized forms of cultural globalization remain largely invisible.

4. Globalization and Cultural Conflict

World culture engenders conflict in world society through a process that has, as it were, turned the West against itself. By the 1970s, when scores of new states had formed in Africa and Asia, vociferous opposition to continued economic and cultural domination by Western countries (labeled neocolonialism and cultural imperialism, originally formulated in the West) began to penetrate global organizations, especially UN bodies. Various associations of less developed countries called for a restructuring of world society—a New World Economic Order and a New World Information Order that would put restraints on the operations of transnational corporations and shift resources to the poorer countries. Former colonies began assertively invoking world-cultural principles of equality and development as basic human rights. Further tensions arose as the global human rights sector (itself increasingly involving individuals and groups from outside the West) turned its scrutiny on the new states, decrying their selective championing of certain global universals while ignoring or severely violating others, especially those relating to the integrity and political participation of citizens. Thus, both the particularism of world culture (the moral legitimacy attributed to national or ethnic units) and its universalism (the insistence that all units abide by basic principles and values) yield forms of disagreement and conflict that would not emerge in a less globalized world.

As non-Western cultures and regions have become more prominent in the world polity, it makes increasing sense to speak of world cultures (in the plural) rather than a singular world culture. Dominant Western models have penetrated deeply in most places, but they have also evoked resistance as well as efforts to revivify and globalize alternative models. Most notable in this regard is the assertiveness of Islamic cultural carriers, particularly since the 1970s. Many Muslim leaders and organizations promote a societal model that infuses the state with religious precepts and recasts the relationship between state and citizen (codified in 1981 as the Universal Islamic Declaration of Human Rights). African and Asian models of social organization and development have also emerged, and some observers argue that conflict in the twenty-first century will revolve primarily around grand civilizational axes rather than the nation-state clashes that have dominated in recent centuries.

By the same token, more fine-grained analyses of world culture(s) identify multiple models of central

world-cultural constructs. For example, derived from the Western tradition are liberal, socialist, corporate, and welfare models of the state; from Asian sources, quasifamilial and state-led development models. Multiple models of the individual, the business enterprise, and the national polity are further examples. World-cultural complexity has increased rapidly in recent decades, perhaps most sharply since the collapse of Communism and the end of the Cold War, as ever more cultural centers generate more alternative cultural constructions. Poorer and more peripheral societies are less able to bring their cultural models to the world-cultural table, but many participants in the global arena from richer societies have become strong advocates of the poor and peripheral, helping to ensure that world culture becomes yet more complex and incoherent but also more significant for national and local structuration and change.

5. Religion and World Culture

Since the early period of European expansion, many of the symbolic flows, organizing institutions, and foundational assumptions that constitute world culture have been religious in origin. Through conquest, evangelization, and migration, Europeans made much of the world 'safe for Christianity,' giving emerging world culture a distinctively Christian cast even though most of the world did not become Christian. In the current phase of cultural globalization, the overall role of religious institutions may have diminished but various religious actors on the global stage vitally affect world culture in new ways, notably by providing alternative worldviews. Religious globalization has been important in constructing world culture, contributing to its heterogeneity, and producing new forms of conflict, as explained below.

Religion played a distinctive role in the construction of world culture, providing motivation for some globalizers and legitimating the actions of others. It also contributed to the foundations of contemporary world culture. The conception of societal progress that prevails almost everywhere stems in part from the Christian legitimation of rational world mastery; the global view of the individual as agent and citizen similarly has roots in Christian views of the person; the global script for organizing society in a rational-legal manner and legitimating authority voluntaristically derives indirectly from Christian precedents. More concretely, missionaries not only spread their faith, they also provided education and health care in far-flung places, diffusing secular commitments that since have become globally entrenched. Their actions prefigured the efforts of later movements in global civil society referred to above. Today distinctly religious views are less obviously influential in world culture, yet they still shape its evolution. According to some, a vigorous Pentecostal movement, successfully expand-

ing to Latin America, Asia, and Africa, carries with it the 'American' gospel of material success and individual well-being, thus supporting global corporate culture. The World Council of Churches, others argue, contributed centrally to global environmental concerns with its initiative on 'Justice, Peace, and the Integrity of Creation,' launched in 1983.

Religion's promotion of world-cultural heterogeneity reflects in part the increasing prominence of Third World voices in global religious organizations. Pope John Paul II has appointed a record number of non-Western bishops and cardinals. In the Anglican Church, non-Westerners significantly aid opposition to gay rights. Non-Western Pentecostals have not only grown dramatically in numbers, they have also devised their own spiritual practices to suit local circumstances. Cho Yong-Gi, pastor of the large Yoido Full Gospel Church in Seoul, initiated a world mission program that has sent hundreds of missionaries and established seminaries in several developed countries, including the USA.

Religion's relation to cultural relativism, another aspect of heterogeneity, varies greatly by tradition and is therefore more complex. In some cases, such as that of Hindu nationalism, religion becomes the primary vehicle for expressions of national distinctiveness. While few Christian churches accept the equal value of all cultures, many have in fact accommodated major cultural differences within loose doctrinal parameters. In Korea and Brazil, for instance, Pentecostal churches incorporate elements of traditional non-Christian spirituality. Islamic fundamentalism presents yet another variant, claiming global legitimacy for its defense of 'authentic' Islam while rejecting the value of other traditions. Finally, religion helps to promote heterogeneity through the creolization process mentioned earlier. Pentecostalism is once again a case in point. An American-inspired movement with a core evangelical message and a secular commitment to material progress has become a mosaic of styles and practices as local congregations have adapted global models to their own needs without direct outside control. As new global movements make inroads in 'national' religious traditions, they produce new juxtapositions that creolize religious cultures. Of course, this process is not unprecedented. Many successful religious traditions have balanced universal ambition with local creativity to produce hybrid religiosity 'on the ground' in counterpoint to unification at the doctrinal level.

Religion also contributes to world-cultural pluralism and conflict. Religious traditions contain diverse views of the good society, leading to varying interpretations of seemingly shared global values. Some view the dominant form of world culture as a coercive, Christian-biased Western force that threatens the integrity of their own traditions. Islamic fundamentalists oppose what they perceive as a godless consumer culture that undermines people's faith and

serves the political interests of the West. Pope John Paul II has criticized an amoral, neoliberal world order that puts profits ahead of the needs of the poor.

Specific elements of world culture also provoke distinctly religious opposition. In recent decades, for example, the principle that women are citizens—individuals to be granted equal rights and opportunities—has been institutionalized in world culture. Within many traditions, however, equality for women is problematic. The result is severe contestation of women's equality in some places, most notably by the Taliban in Afghanistan. These religiously inspired actors have helped generate a more general backlash against globalization, even while they accept many basic elements of world culture. Religion becomes a vantage point from which to oppose the threat represented by globalization as such. The struggle, as some groups see it, is between globalization and their cultural survival.

In sum, religion is a central aspect of cultural globalization, as traditions spread, transnational networks expand, national cultures become more mixed, and new ways of experiencing the world emerge. In many countries, religion mediates the pluralizing effect of world culture. It plays an important role in the intense contest concerning global values and world order. Yet the construction of world culture has become a mainly secular process; it has no transcendent content in the conventional sense. Cultural heterogeneity and conflict themselves take many forms, only some of them religious. While world religion is intimately connected with globalization and involved in the latter's dynamics, it is by no means a dominant force. Whether it can, or should, take on a greater role in defining the desirable world order is likely to be a central issue in future global cultural contestation.

See also: Capitalism: Global; Globality; Globalization, Anthropology of; Imperialism: Political Aspects; United Nations: Political Aspects; Values, Anthropology of

Bibliography

- Appadurai A 1996 *Modernity at Large: Cultural Dimensions of Globalization*. University of Minnesota Press, Minneapolis, MN
- Boli J, Thomas G M (eds.) 1999 *Constructing World Culture: International Nongovernmental Organizations Since 1875*. Stanford University Press, Stanford, CA
- Brouwer S, Gifford P, Rose S D 1996 *Exporting the American Gospel: Global Christian Fundamentalism*. Routledge, New York
- Featherstone M (ed.) 1990 *Global Culture: Nationalism, Globalization and Modernity*. Sage, London and Newbury Park, CA, in association with *Theory, Culture and Society*
- Friedman J 1994 *Cultural Identity and Global Process*. Sage, London

- Hannerz U 1996 *Transnational Connections: Culture, People, Places*. Routledge, London
- Huntington S P 1996 *The Clash of Civilizations and the Remaking of World Order*. Simon & Schuster, New York
- Jameson F, Miyoshi M (eds.) 1998 *The Cultures of Globalization*. Duke University Press, Durham, NC
- King A D 1997 *Culture, Globalization and the World-System: Contemporary Conditions for the Representation of Identity*. Reprint edn. University of Minnesota Press, Minneapolis, MN
- Lechner F 1992 Global fundamentalism. In: Swatos W H Jr. (ed.) *A Future for Religion? New Paradigms for Social Analysis*. Sage, Newbury Park, CA, pp. 19–36
- McLuhan M 1994 *Understanding Media: The Extensions of Man*, 1st MIT Press edn. MIT Press, Cambridge, MA
- Robertson R 1992 *Globalization: Social Theory and Global Culture*. Sage, London
- Smith A 1980 *The Geopolitics of Information: How Western Culture Dominates the World*. Oxford University Press, New York
- Thomas G M, Meyer J W, Ramirez F O, Boli J 1987 *Institutional Structure: Constituting State, Society, and the Individual*. Sage, Newbury Park, CA
- Tomlinson J 1991 *Cultural Imperialism: A Critical Introduction*. Pinter, London
- Van Vucht Tijssen L, Berting J, Lechner F J (eds.) 1995 *The Search for Fundamentals: The Process of Modernisation and the Quest for Meaning*. Kluwer Academic Publishers, Dordrecht, The Netherlands, and Boston, MA
- Watson J L (ed.) 1997 *Golden Arches East: McDonald's in East Asia*. Stanford University Press, Stanford, CA

J. Boli and F. J. Lechner

Globalization, Anthropology of

1. The Disciplinary Context

Anthropological interest in globalization is difficult to trace precisely but can be recognized as early as Appadurai's work on the global cultural economy (1990, 1996), Hannerz's analysis of cultural complexity and creolization (1987, 1992), and Friedman's work linking global capitalism and cultural processes (1988, 1994), all of which built on earlier anthropological interest in cultural exchange and world systems. From the start, such work has been part of two sets of scholarly exchanges and debates. The first links anthropology with a wider debate about globalization in other social science fields, notably geography, political science, and sociology. The other is internal to anthropology and links it to longstanding concerns with diffusion, comparison, ethnography, and the study of large-scale historical change. A brief look at each of these contexts might help to frame the anthropological perspective on globalization.

As regards the social sciences in general, anthropologists in the late 1980s found themselves, like many

other scholars, trying to understand the processes that led to the breakdown of the Soviet world and to guess at the shape of things to come. They soon found themselves having to engage powerful perspectives on the emergent world order, which came out of the Marxist tradition in geography (exemplified by Harvey 1989), by a renewed interest in global political culture, exemplified by the polemical and much discussed work of Samuel Huntington (1996), and some prescient studies of the new forms being taken by global capitalism in the last decades of the twentieth century (Lash and Urry 1987). In addition, anthropologists found themselves in a rich, sometimes competitive dialogue with scholars in literary and cultural studies, notably those influenced by British cultural Marxism, and most powerfully represented by Stuart Hall (1986). In addition, the publication of Benedict Anderson's landmark study of nationalism (1983) provoked a strikingly wide debate about the links between politics, the imagination, and national identification. These stimuli helped to shape anthropological research on globalization, which in many ways is marked by the ongoing effort to link broad structures and processes in the world economy to the subtleties of communication, interpretation, and translation that govern everyday life in all societies.

Within anthropology, the study of globalization built on several well-established currents of interest. In the United States, the study of globalization fit well with Boasian traditions in the study of diffusion, cultural change, and culture contact and with a longstanding interest in urban settings and complex societies. In Europe, anthropologists were slower to take an interest in globalization, but here too longstanding interests in problems of scale, social cohesion, and structural change made globalization a subject of growing interest after 1990. In other parts of the world, such as Latin America, Africa, and India, the anthropological interest in globalization was more closely linked to problems of ethics, development, and inequality.

In spite of these various currents of research and theorization within which the anthropological interest in globalization has developed, from the beginning there were serious doubts about whether globalization by its nature was a topic suitable to the special strengths of anthropology in the study of intimate social relations and of societies governed by non-market social principles. Many of these anxieties have been translated into methodological concerns, reflected in a significant body of methodological work about how anthropology ought to address the emergent world of globalization (e.g., Fox 1991). There is still a considerable body of opinion among professional anthropologists that globalization may well be a mere trend, an artifact of academic fashion, and that in any case, it is not an ideal subject for anthropological research because of its conceptual and social scale. But this rearguard anxiety, not always easy to distinguish

from fear of disciplinary failure, has not prevented a rich set of theoretical debates and empirical studies from emerging out of anthropological traditions, both in the Atlantic world and beyond. These studies, largely a product of anthropological work in the 1990s, are discussed in the following sections.

2. *Global Cultural Flows*

The image of 'flow' has been a central trope of anthropological work on globalization. In the work of Appadurai (1990, 1996), the idea of 'flow' was used to capture a complex dynamic in which objects, ideas, ideologies, technologies, and images were placed in a single economy of circulation, with an eye to distinguishing different emergent mosaics of cultural form and social design. In Hannerz's use of this concept (1992, 1996), it was fruitfully linked to Kroeber's idea about large-scale civilizational and interactional spheres—ecumenes—as well as to ideas about creolization, cosmopolitanization and the professional networks associated with rapid flows of cultural commodities.

In these early usages of the idea of flow, and in subsequent elaborations of them by others, the idea of flow was sharpened to redress the assumption that cultural flows in the era of globalization were unrestricted in scope and range. A salutary emphasis on boundaries, limitations, selectivities, and exclusions reappeared, thus placing the classic debate between Marxist and cultural approaches once again at the center of attention. This broad interpretive struggle continues, and the choices it poses are reflected in many studies associated with the anthropology of globalization, on such subjects as migration (Basch et al. 1994), nationalism (Chatterjee 1993), colonialism and postcolonialism (Comaroff and Comaroff 2000a), media (Feld 1999), diaspora (van der Veer 1995), and markets (Hart 1999).

The positive contribution of the idea of flows was to open up new questions about the relation between legal and illicit flows of commodities and persons, about the changing nature of borders and boundaries, about the implications of new regimes of financial circulation and cybercommerce, and about the evolution of new forms of diasporic identification and mobile tradition building. The image of flows also opens a potential conversation, yet to be fully exploited, between economists and anthropologists on the question of the changes in the world economy suggested by the idea of globalization, and of the best ways of measuring their scale and significance. In a general way, the engagement with cultural flows, their associated economies and interpretations, has decisively marked a shift, in which anthropologists are no longer confined to the study of simple market societies but are addressing such complex phenomena as

cybercommerce, intellectual property, and new forms of genetic engineering and marketing. Likewise, the anthropological focus on cross-border flows, as we will see shortly, has opened up a useful conversation between anthropologists, political theorists and sociologists on new forms of citizenship and sovereignty.

Above all, the focus on flows, regardless of context, commodity, or sphere, has brought anthropology back to one of its earlier insights, namely that most human societies have always been in interactive relations with others, forming spheres of exchange and circulation. Thus the images of social stability, impermeable boundaries, and natural divisions between ethnic groups, polities, and communities, sometimes implicit in anthropological practice, have been put under sharp scrutiny. In turn, the widespread recognition that societies in the era of globalization are inevitably parts of wider circulatory systems and networks, has posed new challenges for how anthropologists will need to theorize classical models of cohesion, consensus, and order. At the end of the twentieth century, there is some uncertainty about how to reconcile the undoubted significance of large scale cultural differences with the reality of the velocity of cross-border flows and the uneven thickness of the membranes between societies, whether organized as nation states or not. It seems probable that the question of how social order and cultural coherence in lived communities are maintained in the face of complex and uneven cross-border flows will be a major theoretical challenge for twenty-first century anthropology.

Many of these concerns have produced a special interest in the anthropology of nationalism and a series of contested views about the salience, form, and future of the nation state as a form of global political organization. The pendulum appears to swing between those (both inside and outside anthropology) who predict the early demise of the nation state, and those who see it as growing stronger than ever. Sometimes these debates are reflections of actual differences between existing states and sometimes they are the result of blurring the lines between nations and nationalism on the one hand, and states and their powers on the other. The emerging consensus seems to be that while nation states are unlikely to disappear in the near future, they are evolving their own powers in a transformed ecology of sovereignty. This ecology is affected by changes in the velocity of global capital flows, in the scale and speed of cross-border shadow economies, and in the geography of dual, mobile, and diasporic political identifications among citizens. These processes are discussed further below (see Sect. 4).

The interest in various forms of global cultural flow, and in the nonisomorphic relationship between them, has provoked a strong debate about whether, and the extent to which, globalization is a new social regime. Many voices within anthropology argue vigorously that globalization is no more than a semiotic con-

spiracy of the media and of global capital and its allies in the world of development. In this view, globalization is only a clever way to repackage colonialism, imperialism, modernization, or various combinations of these ideologies. It is old wine in new bottles. This view runs into the difficulty of having to accommodate too many new animals into an existing theoretical Ark, and suffers from the further disability that it does not provide any criteria for recognizing genuine systemic change. It ends history in the name of theory. The more widespread band of opinion among anthropologists is that something surely has changed. The challenge is, as always, to establish what has changed and how it affects the things that have not changed. Thus, there is a renewed interest in how identities are negotiated across multiple locations, how existing structures like witchcraft (Geschiere 1997) become vehicles and sites for new anxieties and energies, and of how new legal and economic arrangements (those sometimes lumped together under the rubric of neoliberalism) call forth new perceptions of danger, wealth, and risk (Comaroff and Comaroff 2000b).

Finally, ideas and debates surrounding the idea of global cultural flows have cast fresh light on a classic concern of anthropology, namely the process of identification and the production of identities. These processes have to be reexamined in a world characterized by massive information flow, heightened media images of life possibilities for ordinary people, and new fantasies of wealth and mobility. The rapid growth in mass-market media penetration across national boundaries, the associated growth in the battle between global and national media barons, the projected increase in various kinds of bandwidth, and the emergent possibility of new, hybrid instruments for cheap, high-speed communication, all combine to change the grounds on which stable cultural identities are produced and secured. Likewise, these factors make it impossible to presume the image of the local as an unchanging ground against which the tableau of global change is played out. Indeed, many anthropologists now have come to take as great an interest in 'the production of locality' (Appadurai 1996) as in the dynamics of globalization.

With its special interest in language, everyday life, and small-scale social interaction, anthropology is evolving a special purchase on the vast terrain represented by globalization, by paying particular attention to the vagaries of translation and localization and to the special role of language and semiotic mediation in producing new forms of lived experience. This has resulted in a sustained effort to examine the cultural dimension of objects and technologies that have their primary home in the spheres of law, science, and the market. Therefore, anthropologists at the beginning of the twenty-first century study the organ trade, refugee incarceration, NGOs, musical copyright law, and derivatives, to name just a few examples of social and technical forms whose cultural design is sometimes

overlooked. In this proliferation of topics (some of them addressed by other authors in the Encyclopedia), there are some major crosscutting themes, three of which are briefly discussed below.

3. Heterogeneity or Homogeneity?

Most anthropologists share the widespread presumption that globalization is a new phase in the integration of the world economy, principally powered by new scales and speeds in the circulation of financial capital and the associated explosion in electronic communication technologies. When we recall that the major models of global economic processes have come out of the Marxist tradition, it helps to explain why there is a tendency among some anthropologists to argue that globalization is no more than a new phase in the generally history of capitalist commodification. From this follows the possibility that, from a cultural point of view, the forces of commodification will produce increasing levels of cultural homogenization and social standardization, particularly in the realm of media icons, social styles, and consumption values. The further extrapolation, rarely stated explicitly, is that these processes will in turn erase cultural differences both within and across societies. This is the nightmare of globalization as coca-colonization (Hollywood, McDonald's, and Disney are the favored icons of this line of speculation). Yet the evidence that societies appropriate various forms of commodification in their own terms, that media messages are downloaded idiosyncratically, that weapons of standardization are frequently subverted to produce further brand differentiation, and that the best laid plans of the standardizers are frequently thwarted by local entrepreneurs and customers, is mounting.

A splendid example of this process of differentiation is provided in a collection of anthropological essays on the role and style of McDonald's, as it functions in six distinct East Asian settings (Watson 1997). The ethnographic essays in this collection illustrate that, as a global profit-making organization, McDonald's is certainly eager to create uniformity in its products and in its outlets wherever possible. But the essays also shows that in each of these settings McDonald's encounters a quite different set of needs, anxieties, and possibilities and is drawn into a specific local mosaic of social patterns and cultural orientations, involving age, leisure, work, and freedom. Though McDonald's is a heavily capitalized multinational corporation, with an unerring eye for its local and global profits, its sensitivity to local markets inevitably contributes to the emergence of different desires and practices in its outlets. This does not make it any less a global corporation, but its strategies are regularly bent to fit local wishes and projects. However, these essays also show that this is not a simple matter of Goliath being repackaged to suit David, culturally speaking. In fact,

what McDonald's produces in these settings are new and unintended styles and contexts for consumption, for sociality, and for public interaction, each culturally distinct but not in any obvious or predictable way. Many other studies, especially in the sphere of consumption (involving clothing, music, food, leisure, and housing) also confirm this process.

Heterogenization is not a mechanical product of the sphere of consumption. It is itself produced by local makers of ideas, images, and commodities who give to the culture industries of different societies and nations their own distinctive stamp. Thus, anthropologists have been able to show that the simple distinction that places global producers and uniformity on the one side, and local consumers and difference on the other, is overdrawn. We have a growing series of studies of cultural production worldwide, especially in the areas of music, film, and advertising, which let us look into the sites and institutions through which global commodities are locally interpreted by producers as well as consumers. The study of 'world music' by ethnomusicologists is perhaps the best developed of these subfields. In general, these studies have produced a broad consensus that cultural differentiation tends to outpace homogenization, even in this most interactive of economic epochs. Naturally, this conclusion fits with the methodological and ethical proclivities of anthropology, with its professional investment in cultural difference and its ethical commitment to cultural diversity. Nevertheless, there appears to be substantial empirical basis for this consensus. But anthropological interest has by no means focused solely on the good news of cultural differentiation and has not been uncritically celebratory in its treatment of globalization, as the following section will show.

4. Sovereignty, Citizenship, and Violence

Anthropologists have not failed to notice that the world of global cultural flows is hardly free of suffering, injustice, and dislocation. They have certainly been aware that the incidence and brutality of forced population movements have in some respects increased, producing many pockets of human dislocation and frequently of ethnic violence. They have paid attention to the dilemmas of proletarian labor migrants, who struggle with their own translocal loyalties—frequently to religious formations which are themselves increasingly transnational in scope—against the pressures of the varied ideologies of cultural and legal inclusion which characterize their new homes.

Considered transnationally, poorer migrants face bewildering new combinations of democratic ideology and racist reality. In some cases, such as the oil kingdoms of the Middle East, nonimmigrants monopolize most of the benefits of citizenship. In others, the benefits of citizenship are more widely shared, but

everyday racism is a constant threat. In some cases, all refugees are drawn into multinational ethnic wars (as in Central Africa in recent years), while in other cases, the revitalization of longstanding ethnonational boundaries flies in the face of intermarriage and ethnic mixing as social realities. In Central Europe after the death of Tito, the new national constitutions frequently encouraged political campaigns that exacerbated genocidal identities parading as 'primordial.' Many nation states now practice various forms of cultural fundamentalism and majoritarianism, as they struggle for new forms of legitimation in a global economy. In the process, they increasingly both produce and demonize 'minorities.' In many of these cases, language, blood, and race are overdetermined icons of peoplehood which force anthropologists to reconcile their own knowledge of the fragility and historicity of these constructions against the popular, lived sense that these realities are timeless, natural, and nonnegotiable.

As anthropologists examine these bloody sites of ethnic violence and violent social purification (Appadurai 1998, Das 1995, Feldman 1991, Malkki 1995), they find themselves having to study a variety of antecedent and emergent social realities. These include new forms of 'flexible' citizenship (Ong 1999); new ideologies of citizenship and law (Comaroff and Comaroff 2000b, Sassen 1996); new battles over indigenes and outsiders (Meyer and Geschiere 1999); and new patterns of state-sponsored violence and legality. In engaging with these topics, anthropologists discover that they need to link longstanding professional interests in conflict, bodily ritual, mythologies of origin, and ethnocultural ideology, with new regimes of law, incarceration, borders, and migration.

In particular, anthropologists are now tackling the ways in which the forms of violence and dislocation produced by globalization may be related to new structures of sovereignty, new forms of nationally sponsored globalization, and new niches for the location of the global economy. They thus find themselves in a productive dialogue with scholars in adjacent fields who are making important contributions to these subjects (Castells 1997, Held and McGrew 2000, Sassen 1998). This engagement promises a more sophisticated answer to the question of how new forms of global linkage are also producing new forms of cultural fundamentalism, a question to which somewhat Manichean answers have been provided in the popular media and its favored public intellectuals in many national contexts.

In the United States, and to some extent in Europe, there have been strong alarms raised by economists and development specialists to the effect that globalization is producing both greater wealth and greater inequality. Yet, in popular opinion, especially among the wealthier countries and classes, there seems to be an emergent sense that globalization is a world of the victors in the Cold War and their global allies who

have cracked the codes of the global market, the information economy, and the new international order. The losers are frequently viewed as cultural holdouts at best, cultural terrorists at worst, people who have not understood the new alphabet of the globalized world. Anthropological work on globalization, in this context, has a challenge and an obligation. The challenge is to take its classical strengths in the study of intimate relations, everyday life, and cultural difference and use them to cast light on circulatory processes that yield local and global results simultaneously. The obligation is to provide a way of thinking about why this era of global integration is yielding at least as much genocidal violence as it is democratic voice. That may not seem like a particularly anthropological obligation, but given that anthropology has long tried to create new ways of showing how broad human processes are found in translation, globalization may need anthropology at least as much as vice versa.

5. Looking Ahead

It is not easy to predict what topics will find special favor among anthropologists interested in globalization in the coming decades. It is likely that some of the broad themes discussed in this essay will continue to elicit empirical research and theoretical debate, but it also seems likely that the institutional and geographical range of the anthropological debate is likely to widen. Research institutes, conferences, workshops, publications, and translations concerned with anthropological contributions to the study of globalization are appearing in Brazil, Turkey, Hong Kong, India, and in many other sites outside of the North Atlantic region. In these many sites, anthropologists are building their own critical resources as they engage with specific kinds of national debate, cultural conflict, and research tradition (Appadurai 1999). Whatever the individual outcomes of this process, two results seem likely. Anthropology is unlikely to remain a bystander in any important national debate about globalization. Moreover, these debates are increasingly unlikely to be determined by scholars, projects, and institutions confined to Europe and the United States. What more appropriate topic could there be for international research than globalization, and what opportunity could be more appropriate for anthropology to contribute to a global scholarly debate about the world we are about to build?

Bibliography

- Anderson B 1983 *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London
 Appadurai A 1990 Disjuncture and difference in the global cultural economy. *Public Culture* 2(2): 1–24

- Appadurai A 1996 *Modernity at Large: Cultural Dimensions of Globalization*. University of Minnesota Press, Minneapolis, MN
 Appadurai A 1998 Dead certainty: Ethnic violence in the era of globalization. *Public Culture* 10(2): 224–47
 Appadurai A 1999 Grassroots globalization and the research imagination. *Public Culture* 12(1): 1–19
 Basch L, Glick S N, Szanton B C (eds.) 1994 *Nations Unbound: Transnational Projects, Postcolonial Predicaments, and Deteriorated Nation-states*. Gordon and Breach Scientific Publishers, Langhorne, PA
 Castells M 1997 *The Power of Identity*. Blackwell, Oxford, UK, Vol. II
 Chatterjee P 1993 *The Nation and its Fragments: Colonial and Postcolonial Histories*. Princeton University Press, Princeton, NJ
 Comaroff J, Comaroff J L 2000a Millennial capitalism and the culture of neoliberalism. *Public Culture* 12(2) Special issue
 Comaroff J, Comaroff J L 2000b Millennial capitalism: First thoughts on a second coming. *Public Culture* 12(2): 291–343
 Das V 1995 *Critical Events: An Anthropological Perspective on Contemporary India*. Oxford University Press, Delhi, India
 Feld S 1999 A sweet lullaby for world music. *Public Culture* 12(1): 145–71
 Feldman A 1991 *Formations of Violence: The Narrative of Body and Political Terror in Northern Ireland*. University of Chicago Press, Chicago
 Fox R G (ed.) 1991 *Recapturing Anthropology: Working in the Present*. School of American Research Press, Santa Fe, NM
 Friedman J 1988 Cultural logics of the global system. *Theory, Culture, and Society* 5(2–3): 477–90
 Friedman J 1994 *Cultural Identity and Global Process*. Sage, Thousand Oaks, CA
 Geschiere P 1997 *The Modernity of Witchcraft: Politics and the Occult in Postcolonial Africa*. University of Virginia Press, Charlottesville, VA
 Hall S 1986 Cultural studies: Two paradigms. In: Collins R (ed.) *Media, Culture and Society: A Critical Reader*. Sage, Beverly Hills, CA
 Hannerz U 1987 The world in creolisation. *Africa* 57: 546–559
 Hannerz U 1992 *Cultural Complexity: Studies in the Social Organization of Meaning*. Columbia University Press, New York
 Hannerz U 1996 *Transnational Connections: Culture, People, Place*. Routledge, New York
 Hart K 1999 *The Memory Bank: Money in an Unequal World*. Profile Books, London
 Harvey D 1989 *The Condition of Postmodernity: An Enquiry into the Origins of Cultural Change*. Blackwell, Cambridge, MA
 Held D, McGrew A 2000 *The Global Transformations Reader: An Introduction to the Globalization Debate*. Polity Press, Malden, MA
 Huntington S P 1996 *The Clash of Civilizations and the Remaking of World Order*. Simon & Schuster, New York
 Lash S, Urry J 1987 *The End of Organized Capitalism*. University of Wisconsin Press, Madison, WI
 Malkki L H 1995 *Purity and Exile: Violence, Memory and National Cosmology among Hutu Refugees in Tanzania*. University of Chicago Press, Chicago
 Meyer B, Geschiere P 1999 *Globalization and Identity: Dialectics of Flow and Closure*. Blackwell Publishers, Malden, MA
 Ong A 2000 *Flexible Citizenship: The Cultural Logics of Transnationality*. Duke University Press, Durham, NC
 Sassen S 1996 *Losing Control? Sovereignty in an Age of Globalization*. Columbia University Press, New York

- Sassen S 1998 *Globalization and its Discontents: Essays on the New Mobility of People and Money*. The New Press, New York
- Soares L 1996 *Cultural Pluralism, Identity, and Globalization*. Conjunto Universitario Candido Mendes, Rio de Janeiro, Brazil
- van der Veer P 1995 *Nation and Migration: The Politics of Space in the South Asian Diaspora*. University of Pennsylvania Press, Philadelphia, PA
- Watson J L 1997 *Golden Arches East: McDonald's in East Asia*. Stanford University Press, Stanford, CA

A. Appadurai

Globalization: Geographical Aspects

1. Introduction

Space, place, distance, proximity, maps, boundaries, territoriality, and spatiality. These are some of the keywords of geography; keywords that distinguish the interest of geographers in the phenomenon of globalization, from that of others in the social sciences. The distinction is not clear cut by any means, not least because 'globalization' so centrally evokes geography: the rise of world-scale processes and phenomena, the intensification of linkage between distant places and cultures, and the associated unmaking and remaking of territorial boundaries and identities. Perhaps where the key difference lies, however, is that the other social sciences are not interested primarily in the intrinsic spatiality of globalization.

In economics and business studies, for example, studies of globalization focus on the implications of the borderless economy (driven by international market integration, global flows of money and information, and transnational production and regulation) on national and subnational economic sovereignty, competitiveness, and growth. The optimists celebrate the offer of new market opportunities, the pessimists warn of the real and discursive dangers of neo-liberalism and international dependency, the historically minded reassure us that there is nothing much new (Drache 1999, Hirst and Thompson 1996), while the measured observers of qualitative change reveal the new aspects of combined and uneven development (Dicken 1998, Held et al. 1999).

In politics and international relations, the debate has centered on the erosion/reformulation of the nation-state as a unit of authority, following the rise of nonstate institutions of regulation and governance, and on the challenge to citizenship and democracy posed by the rise of transnational political organisations (e.g., global NGOs) as well as plural, non-national sources of political rights and identification (e.g., EU citizenship). Here too there is no consensus. Some herald a new era of reduced nation state influence, eroded national welfare commitment and crisis of parliamentary democracy (Garrett 1998),

some assert the continuity of national social models (Hay and Marsh 1999, Weiss 1998), others signal the rise of new forms of unsettled jostling between the nation-state and other organizations organized at local, national, and international scales (Jessop 1994, Keil 1998), while others still find new prospects for cosmopolitan government, democracy, and civic mobilization (Held et al. 1999, Scholte 2000).

In social anthropology and media and cultural studies, the attention has fallen on the implications of global consumerism, media communication, and international mobility and cultural mixture for individual and social identities and lifestyles in different local settings. Early warnings of the erosion of local difference under the weight of the same world products and consumption norms and the same global corporations (i.e., the so-called MacDonaldization of society), have given way to sensitivity to the persistence of local difference, but also the nuances of local change resulting from heightened global exposure and virtual and real linkage with different parts of the world (Robertson 1992, Albrow 1996, Hannerz 1996, Morley and Robins 1995).

Geography, to repeat, is everywhere in these studies of globalization, but the spatiality of social relations, or put differently, how place, space, and spatial difference are being reconstituted is not the prime interest. Geographical theory, in contrast, has been concerned with the spatiality of the contemporary world, and is interested in understanding whether places—cities, regions, and nations—are perforating as geographically contained spaces, how the insertion of places into geographically stretched relations matters, and how new geographical scales of organization and influence associated with globalization are challenging old scales of identification and action.

This geographical imagination, to be clear, is not confined to the discipline of geography. In fact, some of the most influential accounts of the spatiality of globalization have come from social theory outside the discipline. For example, the sociologists Giddens and Held have long argued that a key aspect of globalization is spatial connectivity and reach, such that the conceptualization of place/locale as 'in here' happenings and space as 'out there' happenings is no longer tenable. Held (1995, p. 20), building on Giddens's idea of contemporary time-space distancing (Giddens 1990), proposes:

...globalisation can be taken to denote the stretching and deepening of social relations and institutions across space and time such that, on the one hand, day-to-day activities are increasingly influenced by events happening on the other side of the globe and, on the other hand, the practices and decisions of local groups can have significant global reverberations.

The sociologist Therborn (1998, p. 7) too asserts that globalization should force us to think 'seriously

about space, about the spatiality of the social, about territories and their delimitations.' Two examples of spatiality that he identifies are: first, in a world seen as system, the 'flattening of social processes,' associated with the gradual global spread of dominant social practices (e.g., the triumph of capitalism on a global scale, the hegemony of neoliberalism, the Americanization of culture); and second, echoing Giddens and Held, the 'world as an arena where nationally determined actors meet, interact, and influence each other.'

In contrast, the international studies scholar Scholte (2000) interprets the spatiality of contemporary globalization (compared to preceding periods of intense internationalisation) as 'deterritorialization.' He notes that the rise of a 'social space that transcends territorial geography' is new (composed of such phenomena as transnational production, instantaneous worldwide communication, consumption of the same commodities and images, etc.), such that:

'global' relations are social connections in which territorial location, territorial distance and territorial borders do not have a determining influence. In global space 'place' is not territorially fixed, territorial distance is covered in effectively no time, and territorial frontiers present no particular impediment.

He concludes that 'while territoriality may continue to be important, globalization has brought an end to territorialism (that is a condition where social space is reducible to territorial coordinates alone).'

In all three accounts, it is the spatiality of globalization that is stressed, albeit in very different ways, from the emphasis on new territorial scales and connections, to warnings of interterritorial clash or territorial annihilation.

The interpretation shared, and explored further, in this contribution is the one offered by Giddens and Held. It sees the growth of world-level influences (e.g., transnational corporations and banks, common consumption patterns, world ideologies, international authority structures) and the intensification of global spatial connectivity (e.g., flows of people, goods, ideas, and information, facilitated by rapid transport and communications technologies) less of a threat to, or reinforcement of, territoriality, than a loosening and stretching of social geographies of linkage and identification. This is because a key aspect of globalization is:

A process which embodies a transformation in the spatial organization of social relations and transactions—assessed in terms of their extensity, intensity, velocity, and impact—generating transcontinental or interregional flows and networks of activity, interaction, and the exercise of power (Held et al. 1999, p. 16).

The first part of this article discusses recent claims in geography on the changing relationship between place

and space associated with globalization. It argues that the legacy of seeing places in scalar terms, either as geographical sites of proximate links or as territorial units, remains strong within the discipline. The second part—using the example of contemporary urbanity—dissents with the latter interpretation, to advance a very different sense of place. It argues for the need to see places under contemporary globalisation in non-scalar terms, and as perforated entities: as sites immersed in varied geographies of proximate and distant links, within which things no longer cohere as a bounded whole.

2. Geographies of Scale

Marston (2000) has reviewed the work of political geographers such as Neil Smith, Eric Swyngedouw, John Agnew, Kevin Cox, Peter Taylor, and Neil Brenner, to conclude that for them spatial scales—from home and locality to city, region, nation and continent—have no pre-given or fixed ontological status, but are produced and reproduced and continually transformed socially by the imperatives of capitalism. Importantly, they associate contemporary globalization with the rise of new territorial scales of organization, and new territorial boundaries, rather than the replacement of old, national scales and boundaries by a unitary global field of meaning and action. They do not see the replacement of an old world 'space of places' with a new 'space of flows' (Castells 1996).

Brenner (1999, p. 435), for example, asserts that 'the post-1970s wave of globalisation has significantly decentred the role of the national scale as a self-enclosed container of socio-economic relations,' and we are witnessing a 're-scaling of territoriality' which includes the increased 'importance of both sub- and supranational forms of territorial organisation.' Similarly, Swyngedouw (1997) mobilizes the term 'globalisation' to indicate that globalisation represents the break down and reconstitution of spatial scales; producing a new 'politics of scale' involving clashes of scale and contested boundaries. Globalization, thus, represents the 'relativisation of scale' (Jessop 1998, p. 90), with governance and political conflicts articulated at urban, regional, national, and supranational level, with no one scale privileged. The result is a re-scaling of territories, such that, for example, national states gear towards enhancing international competitiveness and capturing global investments cities, and cities or regions decouple from national economic circuits, to become locked into global ones (e.g., as investment sites).

The emphasis of these geographers on the produced and contingent nature of territories and scales marks, in Agnew's (1999, p. 504) terms, the 'historicity of spatiality.' He demonstrates, for example, that the present period contains a number of historical models

of power organized spatially: the model of dynastic preservation in a world of distant communities; the model of state–military conflict in a world of rigidly defined territorial units; the model of hierarchical conflict between centers and peripheries in an international economic system; and the model of nodal conflict within the networks of an integrated world society. The ‘geographical embeddedness of power relationships today, therefore, signifies something quite different from what was the case in the past’ (p. 512), for example, contemporary globalization elevates the tensions between state territorial relationships (e.g., equating nation with citizenship or property rights) and transterritorial developments (e.g., the rise of globally mobile identities and property rights).

Through these readings, globalization can be seen as the multiplication of geographical scales of social organization, linked to the changing spatial requirements of the latest phase of capitalist development. These are not seen as parallel or mutually exclusive scalar geographies, but intersecting and overlapping ones, with places/territories implicated in the full variety of scalar reorientations. Thus, not an end to geography in a globalizing totality, but a rescaling of boundaries (upwards, outwards, and downwards) and a reconfiguration of territories as they engage in multiscale processes and politics.

The strength of this interpretation of the spatiality of globalization lies in the scope it leaves for recognizing the persistence of geographical diversity, and the varied ways in which different places, with their unique legacies, are implicated in the balance between territorial and transterritorial influences. Its weakness, however—although Agnew is an exception—is that it fails to address whether the ontological status of place/territoriality itself is significantly altered as a result of the rise of world-scale processes and transnational networks and flows. The language continues to remain that of scalar geographies and territorial units. Reterritorialization follows deterritorialization and spatial scales are relativized under globalization.

Places such as cities and nations continue to exist as territorial units, but now with different external orientations (e.g., as sites in global production networks or places dependent upon international investment or competitiveness) and different scalar involvement (e.g., national welfare policies, continental trade agreements, global environmental regulations, local tax regimes). The boundary between place and space is not seen to be bending or disappearing—a line in shifting sand—but as a remaining line of distinction between ‘in-here’ relations and institutions, and those ‘out-there.’ Global connectivity is seen as leading to the nesting of territories into multiple scales, not territorial disorientation away from a geography of scalar nesting.

Cox’s (1998) work on localities as particular sorts of spaces (of association and politics) in a multiscale world provides a good example of this lingering

territorial logic. He distinguishes within a locality the spaces of dependence or localized connections that he claims define ‘our sense of significance,’ from the spaces of engagement within global relationships, which he sees as threatening the place-specific ‘essential interests’:

Spaces of dependence are defined by those more-or-less localized relations upon which we depend for the realization of essential interests and for which there is no substitute elsewhere; they define place-specific conditions for our material well being and our sense of significance. These spaces are inserted in broader sets of relationships of a more global character and these constantly threaten to undermine or dissolve them... In so doing they construct a different form of space which I call here a space of engagement: the space in which the politics of securing a space of dependence unfolds (Cox 1998, p. 2, cited in Marston 2000, p. 226).

In this account, localized relations are posed as distinguishable from global relations, and somehow more authentic and politically progressive: the local can be mobilized to resist the global. While Cox is clearly right to note that local societies continue to retain political potential in an age of allegedly globalized power, his dualist distinction between the ‘local’ and the ‘global’ as separate scalar fields remains methodologically problematic. To highlight the weakness of this dualist logic, should we, for example, judge the often charged diaspora connections of ethnic minorities or cosmopolitans in a locality to be less significant or progressive than the daily contact among local residents? Surely a key aspect of the transnationalisation of local relations is that we can no longer make an easy distinction between local and global geographies? How localized or global, for example, are the associational politics of worker, immigrant, and NGO groups campaigning for local recognition, but relying on international financial and other support networks?

Or, what makes everyday practices and customs grounded in local encounters and traditions any more ‘essential’ than nonlocal practices and customs? This desire to see places as territorial units in which local relations are counterposed to global relations leads to a peculiar politics of place in which relations within localities are cast as good and meaningful, and contrasted to bad and totalizing external relations. Its peculiarity follows from the ontological separation of place (read as in here, and intimate) and space (read as out there, and intrusive) as distinctive scalar realms.

Taylor’s (1999) recent attempt to conceptualize place and space in nonscalar terms helps us get around this particular ontological straightjacket, but replaces it with another rather stark binary distinction. Following the pioneering work of the humanist geographer Tuan (1977), Taylor claims that both terms should be used to identify human activity at all geographical scales—local, national, and global—but that they are distinct and always in dialectical tension

with each other. Taylor defines space in modernity as the realm of abstract principles, rules, rationality, science, administration, bureaucracy, and institutions. He judges it to be politically disenabling. In contrast he sees place as the substantive, politically enabling, lived realm involving intimacy, experience, belonging, feelings. The two are said to always exist in tension with each other, for example, the abstract realm of the state, combined with the felt realm of the nation, has yielded the nation-state, in which 'sovereign territory has been merged with sacred homeland to convert a space into the place.' Similarly, Taylor sees the modern home converting the space of the household into a place, and the techno-social space of the one-world (capitalist) system becoming a place through developments such as the rise of a global consciousness (e.g., global ecological fallibility).

Taylor's is an odd usage of the terms space and place, counterposing as it does, the abstract (disenabling) from the lived (enabling). Are not the spaces of science and technology, for example, R&D laboratories or domesticated science, just as lived and 'enabling' as more obviously emotional sites? Isn't the realm of experiences one that blurs the distinction between formal and informal practices? Why, for that matter, choose the labels of space and place to distinguish between the abstract and the lived? What remains helpful in Taylor's schema, however, is his rebuttal of space and place as separate geographical sites or realms, thus allowing the ontological presence of both the proximate and the remote at the same geographical level (e.g., the recognition of systemic rationality or state-sponsored science in the home, or the proximate warmth/odium of nationalist sentiment or universal causes).

The next section pursues this reasoning, but retains the geographical connotations of the terms space and place, rather too quickly abandoned by Taylor. It retains a geographical imaginary because it considers the reconfiguration of the spatiality of social relations a central aspect of contemporary globalization, and importantly, because it sees this reconfiguration shaping the meaning of what goes on in places (localities, cities, regions), how places coalesce or not as entities, how social relations in places are constructed, and how the politics of place matter. In short, it illustrates that place still matters, but no longer as a scalar unit.

3. 'A Global Sense of Place'

The section heading borrows a deliberately oxymoronic phrase of Massey (1994) to conceptualize localities in nonscalar terms, or, in her words, 'relationally,' as sites of juxtaposed social realities, shaped by the layered historical connections and geographical links that are to be found within and beyond them. Their character as sites of agglomerated activity, thus, has far less to do with the properties of a territorial system

(e.g., localized linkage, self-regeneration, local governance, etc.) than with the effects of spatial and temporal exposure and connectivity (e.g., continual and open-ended change, juxtaposition of difference, overlap of networks of different global connections). Two of Massey's now famous examples of places as 'open, porous, hybrid' (Massey 1999, p. 21) are Mexico City, with its varied architectures revealing layers of global Spanish and American-Indian historical connections, and Kilburn High Road in London, with its juxtaposition of English, Irish, and Indian diaspora lifestyles and connections.

It is important to note that since Massey defines the spatial in general 'as the sphere of the juxtaposition, or co-existence, of distinct narratives, as the product of power-filled social relations' (Massey 1999, p. 21), a global sense of place is needed not only for cosmopolitan places, but all places. For, each and every one:

...may be imagined as particular articulations of these social relations, including local relations 'within' the place and those many connections which stretch beyond it. And all of these embedded in complex, layered, histories. This is place as open, porous, hybrid—this is *place as meeting place*... This is a notion of place where specificity (local uniqueness or sense of place) derives not from some mythical internal roots nor from a history of relative isolation—now to be disrupted by globalisation—but precisely from the absolute particularity of the mixtures of influences found together there (Massey 1999, p. 22, emphasis in original).

Through this lens we might re-imagine Bamako as richly as Bombay, Boston as richly as Bermuda, and mobilize them against persisting views of globalization as place homogenisation. Massey defines places as socially constructed 'envelopes of space-time': relational settings and meeting places, not closed local systems. How does contemporary globalization fit into this picture? Massey's response is:

...globalisation, imagined through the lens of this conceptualisation of space-time, the globalisation we are facing now, is a thoroughgoing, world-wide, restructuring of those space-times, along particular lines. It is a re-making of those, inherited but always temporary and provisional, spaces, places, cultures which are themselves the hybrid products of previous restructurings.

We can take this to mean that a relational or global sense of place will interpret the spatiality of globalization—however defined—rather differently from a territorial perspective. Let us stay with the example of cities, and focus in turn on the spatiality of two aspects—economic transactions and urban politics—to illustrate this point.

3.1 Urban Economic Space

In economic discussions, the language of scales or territorial boundaries, as discussed above, has tended

to interpret the rise of transnational phenomena such as global informatics, international production and finance, and powerful international organizations, in terms of the boundary between a territorial inside and a territorial outside. Consequently, accounts which continue to see an economic place for cities contrast the global as the realm of distanced links, fast flows, and world-scale processes, with cities as the sites of face-to-face-contact, or centers of agglomeration among global networks (e.g., clusters of HQs, creativity and knowledge, communication hubs). Against the economics of space and distance are posed the economics of agglomeration and clustering, the powers of spatial proximity and interiority. The argument posed is that in the global age, the locally embedded still counts as a source of economic dynamism and competitiveness because spatial proximity encourages trust and reflexivity, reduces transaction and communication costs, facilitates technological and knowledge spill-over, and offers specialized services and institutions to the business and professional communities (Maskell et al. 1998, Krugman 1991, Audretsch 1998, Glaeser 1998).

A sense of place that sees cities as a meeting place for diverse geographies of linkage, in contrast, offers a different reading of the urban economy. The geography of cyberspace is a good example for questioning the distinction between an alleged real world of direct contact and a virtual, telemediated world. As Kitchin (1998, p. 387) argues, 'whilst information on-line might seem geographically dislocated, information is only as useful as the locale within which the body resides', and in any case, 'cyberspace depends on real-world spatial fixity—the points of access, the physicality and materiality of wires.' Consequently, there is a continuity of communication in geographical space and place, such that, as Graham summarizes, 'power to function economically and link socially increasingly relies on constructed, place-based, material spaces intimately woven into complex telematics infrastructures linking them to other places and spaces' (Graham 1998, p. 175). Similarly, in daily contact, there is an unbroken line of communication linking physical and human technologies, and it is of varying geographical reach, mobilized to source information, strike a deal, exchange knowledge, and so on.

The different realms of communication do not offer different social worlds, for example, 'there is not one single, unified cyberspace; rather there are multiple, heterogeneous networks, within which telecommunications and information technologies become closely enrolled with human actors, and with other technologies, into systems of sociotechnical relations across space' (Graham 1998, p. 178). And this is precisely why the speeding up of time within communications networks should not be read as the speeding up of people and their lifestyles (Thrift 1996).

Seen in the context of linked communications

architectures, there can be no privileged economic role for spatial proximity. First, co-presence or agglomeration need not imply association and interaction, for firms can mobilize a variety of contact networks to establish economic links with firms, markets, and institutions located elsewhere nationally and internationally. The near immediacy of distanced communications networks enables economies of association to be established without the need for spatial propinquity. The city is no longer a bounded, but spatially stretched, economic sphere. Second, spatially proximate interactions, where they appear, need not privilege particular types of economic transactions such as those based on trust, reciprocity, and reflexivity through physical interaction.

Intimacy may be achieved through the frequent and regular contacts enabled by the distanced networks of communication and travel (how else do transnational firms, institutions, and social movements work?), as well as the unbroken interplay between face-to-face and telemediated contact. The city is, therefore, not a place of unique or systemic transactional assets. Third, the agglomeration in some parts of some cities of large volumes of specialised inputs—from knowledgeable and creative people and clusters of inter-related industries, to sites of R&D, corporate power, public bureaucracy, and business services—no longer implies increasing local returns (as theorized by agglomeration economists). These sources of economic dynamism (which may well draw on local resources, for example, schools and college, leisure amenities, transport hubs) are nested into economic networks that stretch beyond the city (from transnational production networks and national or international supply chains, to global market relations).

A global sense of urban economic space recognizes places as clusters of overlapping network sites, without seeing in them only local economies of association feeding a localised economic system. It views with scepticism the territorial metaphors such as 'local fixity vs. global flow,' 'embedding the global,' or 'the local in the global' which have grown in recent years to 'defend' the space of local economy against global 'encroachment.' Instead, it highlights that local economic activity is part of, and inseparable from, proximate and distanced transactions, and that whatever counts as the local has to be seen as a network site, not a territorial system.

3.2 Urban Political Space

Similarly, a global sense of place implies an interpretation of the implications of globalization on urban politics that is different from one based on a territorial sense of place. The territorial perspective anticipates either the loss of the urban as a political space under the barrage of national and global institutions and

movements, or, as exemplified by the work of Cox above, as a particular zone of engagement. It sees politics as undergoing scalar differentiation, as exemplified by waning interest in universalist causes and nation-state institutions, and the rise of ethno-nationalist sentiment at regional level or global political institutions and movements.

In this schema, localities, when viewed positively, are expected to perform a certain kind of place-based politics: tribal, as implied by Cox, through his claims for local authenticity and feelings; formative, through involvement in 'militant particularisms' (Harvey 2000); lived, through conflicts over lifestyle, consumption, and social reproduction issues; experimental, via either the politics of difference forged around the diverse social and cultural voices gathered in the city (Sandercock 1998), or the politics of citizenship based on participation in the city's public spaces and public sphere (Sennett 1977). In short, a new politics of place serving unique purposes in a world of multiscale politics (note, however, the different position of those political geographers who mobilize the concept of relativization of scales, in order to capture the new combined politics of local, national, and global action).

There is, however, another way of viewing place politics in age of global connectivity. Instead of seeing local political activity as somehow unique, places might be seen as sites that juxtapose or combine the varied politics—local, national, and global—that we find today. What matters is this juxtaposition/combination.

This forces, for one, recognition of actor participation in nonlocal political activity. To return to the urban example, as Sassen (1998) puts it, if globalization involves the urbanization of a world capitalist class (at least in some global cities) as well as people and cultures from different parts of the world, it also generates the possibility of claims in the city which are at once local, national, and transnational in orientation. This might include, for example, business leaders campaigning against the local homeless, national government tax policies, or international trade agreements; ethnic minorities campaigning against local racism, but also through the support of national antiracist groups, and international diaspora organizations; office workers gathered together with other groups outside the municipal offices protesting against local closures, but also linked up with workers and unions elsewhere through Internet discussions of world trade barrier removals. In this perspective the city is no longer a site of place politics, but a place of engagement in plural politics, and through multiple spatialities of involvement.

Why does this shift in emphasis from the politics of place to politics in place matter? Because, it marks the necessity to see political activity in places as plural, open, and contested (Amin et al. 2000). First, the juxtaposition of difference in close spatial proximity

generates political challenges associated with this juxtaposition, for example, the contest between the business community and the homeless over the spaces of the city, or the competing demands of different classes, social and ethnic groups over the city's cultural resources. There is no automatic intimacy or consensus here against any perceived global intrusion.

Second, the idea of politics in place marks a significant change in the nature of contemporary politics, from one based on participation in singular, all-encompassing, political institutions and movements, to another based on participation in multiple causes and multiple networks of affiliation, and resulting in people, more than ever before, bearing multiple political identities as well as forging new ones. It recognizes a cosmopolitan politics that features at all spatial scales of organisation and activity (Soja 2000). Third, however, such a politics remains open-ended, and highly dependent upon cultural context and the balance of power, for its outcomes. Thus, no formative place politics of militant particularisms, but a cosmopolitan politics that is xenophobic and reactionary here, progressive and hybrid there, tolerant or particularly militant elsewhere.

4. Conclusion

This article has staked out a case for a nonscalar interpretation of the geographies of place and space associated with contemporary globalization. Its principal claim has been that the rise of a world of global flows and processes, and intensified transnational connections, signals the perforation of boundaries and the sharp combination of variety and difference at different geographical sites, rather than a shift in the balance of power between different spatial scales, or the deterritorialization and reterritorialization of activity. Thus, it has stressed the fluidity of cultural/political practices in space and place, against the emphasis elsewhere in geography on new scales of fixity. Of course, in matters of regulation and governance associated with globalization, there is a very real and felt contest of jurisdiction between local, national, and global state and nonstate organizations. It has not been the intention here to deny the state, but simply its legal and institutional inscription on other on-goings within places. Instead, the aim has been to look into how the very terminology of geographical boundaries, caught up in concepts of place and space, proximity and distance, has been challenged by globalization.

Beyond these largely silent differences of opinion within geography, the distinctive contribution of this discipline within the congested study of globalization, has lain in the study of the spatiality—social, economic, cultural, and political—of what is increasingly being seen as a single and interdependent world. The work of geographers has shown that this new world remains as varied, uneven, and contested as ever

before, with little evidence of spatial homogenization or the end of geography.

See also: Borders, Anthropology of; Capitalism: Global; Development and the State; Globalization and World Culture; Globalization, Anthropology of; Globalization: Political Aspects; Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity; Place in Geography; Territoriality: Geographical

Bibliography

- Albrow M 1996 *The Global Age: State and Society—Beyond Modernity*. Polity Press, Cambridge, UK
- Agnew J 1999 Mapping political power beyond state boundaries: Territory, identity and movement in world politics. *Millennium—Journal of International Statistics* 28(3): 499–521
- Amin A, Massey D, Thrift N 2000 *Cities for the Many not the Few*. Polity Press, Bristol, UK
- Audretsch D 1998 Agglomeration and the location of innovative activity. *Oxford Review of Economic Policy* 14(2): 18–29
- Brenner N 1999 Globalisation as reterritorialisation: The re-scaling of urban governance in the European Union. *Urban Studies* 36(3): 431–51
- Castells M 1996 *The Rise of the Network Society*. Blackwell, Oxford, UK
- Cox K 1998 Spaces of dependence, spaces of engagement and the politics of scale or looking for local politics. *Political Geography* 17: 1–23
- Dicken P 1998 *Global Shift*, 3rd edn. Chapman, London
- Drache D 1999 *Globalization: Is there anything to fear?* CSGR Working Paper No 23/99, Centre for the Study of Globalization and Regionalisation. University of Warwick, Coventry, UK
- Garrett G 1998 *Partisan Politics in the Global Economy*. Cambridge University Press, Cambridge, UK
- Giddens A 1990 *The Consequences of Modernity*. Polity, Cambridge, UK
- Glaeser E L 1998 Are cities dying? *Journal of Economic Perspectives* 12: 139–60
- Graham S 1998 The end of geography or the explosion of place? Conceptualizing space, place and information technology. *Progress in Human Geography* 22(2): 165–85
- Hannerz U 1996 *Transnational Connections: Culture, People, Places*. Routledge, London
- Harvey D 2000 *Spaces of Hope*. Edinburgh University Press, Edinburgh, UK
- Hay C, Marsh D 1999 Demystifying globalisation. In: Hay C, Marsh D (eds.) *Demystifying Globalisation*. Macmillan, Basingstoke, UK
- Held D 1995 *Democracy and the Global Order: From the Modern State to Cosmopolitan Governance*. Polity, Cambridge, UK
- Held D, McGrew A, Goldblatt D, Perraton J 1999 *Global Transformations: Politics, Economics and Culture*. Polity, Cambridge, UK
- Hirst P Q, Thompson G 1996 *Globalization in Question*. Polity, Cambridge, UK
- Jessop B 1994 Post-Fordism and the state. In: Amin A (ed.) *Post-Fordism: A Reader*. Blackwell, Oxford, UK
- Jessop B 1998 The narrative of enterprise and the enterprise of narrative: Place marketing and the entrepreneurial city. In: Hall T, Hubbard P (eds.) *The Entrepreneurial City: Geographies of Politics, Regime, and Representation*. Wiley, Chichester, UK
- Keil R 1998 Globalization makes states: Perspectives of local governance in the age of the world city. *Review of International Political Economy* 5(4): 616–46
- Kitchin R M 1998 Towards geographies of cyberspace. *Progress in Human Geography* 22(3): 385–406
- Krugman P 1991 *Geography and Trade*. MIT Press, Cambridge, MA
- Maskell P, Eskelinen H, Hannibalsson I, Malmberg A, Vatne E 1998 *Competitiveness, Localised Learning and Regional Development*. Routledge, London
- Marston S A 2000 The social construction of scale. *Progress in Human Geography* 24(2): 219–42
- Massey D B 1994 *Space, Place and Gender*. Polity, Cambridge, UK
- Massey D B 1999 *Power-geometries and the Politics of Space Time*. Hettner Lecture 1998, Department of Geography, University of Heidelberg, Germany
- Morley D, Robins K 1995 *Spaces of Identity: Global Media, Electronic Landscapes, and Cultural Boundaries*. Routledge, London
- Robertson R 1992 *Globalization: Social Theory and Global Culture*. Sage, London
- Sassen S 1998 Whose city is it? Globalization and the formation of new claims. In: Holston J (ed.) *Cities and Citizenship*. Duke University Press, Durham
- Sandercock L 1998 *Towards Cosmopolis: Planning for Multi-cultural Cities*. Wiley, Chichester, UK
- Schulte J A 2000 Global civil society. In: Woods N (eds.) *The Political Economy of Globalization*. Macmillan, London
- Sennett R 1977 *Fall of Public Man*. Knopf, New York
- Soja E W 2000 *Postmetropolis*. Blackwell, Oxford, UK
- Swyngedouw E 1997 Neither global nor local: 'Glocalization' and the politics of scale. In: Cox K R (ed.) *Spaces of Globalization*. Guilford, New York
- Taylor P J 1999 Places, spaces and Macy's: Place-space tensions in the political geographies of modernities. *Progress in Human Geography* 23(1): 7–26
- Therborn G 1998 Globalizations. Swedish Collegium for Advanced Study in the Social Sciences. Uppsala, Sweden
- Thrift N 1996 New urban eras and old technological fears: Reconfiguring the goodwill of electronic things. *Urban Studies* 33: 1463–93
- Tuan Y-F 1977 *Space and Place*. Arnold, London
- Weiss L 1998 *The Myth of the Powerless State*. Cornell University Press, Ithaca, NY

A. Amin

Globalization: Legal Aspects

1. Introduction

In light of the dramatic intensification of transnational interactions since the 1970s and its impact on the legal field, one might be tempted to conceive of the 'globalization of the legal field' as a radically new phenomenon without any roots in the past. Though

such legal globalization, or, at least, some aspects of it represent a qualitatively new development, it should be borne in mind that the modern world system, within which the globalization of social interactions occurs, has been in place since the sixteenth century and that the roots of the most recent legal, as well as cultural, social, political, and economic transformations are to be located in this historical development. Moreover, the existence of a transnational legal culture antedates the modern world system, as is dramatically illustrated by the reception of Roman law in the twelfth century and onwards.

Even at the end of the nineteenth century, at a time when liberal political theory had imposed the equation among nation, state, and law, the idea of a global or world law continued to flourish as a legal subculture. Though cultivated at the time mainly by legal comparativists who were quite marginal to mainstream legal science, this legal subculture was part and parcel of a widely hegemonic cultural constellation, the idea of the seamless and irreversible uniformization of the conditions of social life throughout the world brought about by progress, capitalism, and scientific and technological development. In contrast with earlier movements for transnationalization that were the products of an intellectual crusade by well-meaning jurists and philosophers, the transnationalization of the legal field at the beginning of the twenty-first century is promoted by multinational corporations, practicing lawyers, state bureaucrats, and international institutions, as well as by popular movements and transnational advocacy NGOs. It is an extremely diverse phenomenon combining uniformity with local differentiation, top-down imposition with bottom-up creation, formal declaration with interstitial emergence. It cannot be accounted for by any monocausal explanation and defies integration in any unidimensional cultural postulate, be it the idea of progress or the critique of progress, capitalist mass culture or anticapitalist popular culture, modern science, or postmodern knowledge. It mirrors the complexity and ambiguity of the much broader seemingly all-encompassing process of transnationalization, of which it is only a very partial manifestation. Seven major types of legal globalization may be distinguished.

2. Transnationalization of Nation-state Regulation

Transnationalization of nation-state legal regulation refers to changes in the state law of a given country that have been decisively influenced by formal or informal international pressures by other states, international agencies or other transnational actors. Despite the fact that the impact of the international context on nation-state legal regulation is inherent to

the interstate system and can be traced back to the Westphalia Treaty itself (1648), the current process of transnationalization of state legal regulation seems to be a qualitatively new phenomenon for two main reasons. First, it is a very broad and far-reaching phenomenon covering a wide range of state intervention and calling for drastic changes in the pattern of intervention. The core pressure is relatively monolithic as it stems from the 'Washington consensus,' in whose terms the market oriented development model is the only feasible model for a new global regime of accumulation and, accordingly, the structural adjustment it calls for must be carried out worldwide. Other pressures reinforce it, such as the end of the Cold War, the dramatic innovations in communication and information technologies, the new systems of flexible production, the emergence of regional blocs, a newly packaged ideological mix of economic liberalism, and liberal democracy. The scope of these pressures is wider ranging than previous processes of transnationalization, because the current ones take place after decades of active state regulation of the economy in core, peripheral, and semiperipheral countries. The creation of the 'market friendly' model involves massive normative and institutional destruction and construction. The second factor for the novelty of current legal transnationalization is that the asymmetries of transnational power between the core and the periphery of the world system, that is, between the North and the South, are more dramatic today than ever. The sovereignty of the weaker states is now threatened directly not so much by more powerful states, as used to be the case, but rather by international financial agencies and other 'private' transnational actors, such as the multinational corporations.

In the field of the economy the legal transnationalization of state law assumes greatest salience. The policies of 'structural adjustment' cover an enormous range of economic, commercial, and social intervention of the state, provoking turbulence in wide legal fields and institutional settings. Trade liberalization, privatization of industries and services, agricultural liberalization, dismantling of regulatory agencies and licensing mechanisms, deregulation of the labor market and 'flexibilization' of the wage relation, reduction and marketization of social services, less concern with environmental issues, educational reforms geared to job training rather than to citizenship building, family policies that further aggravate the condition of women and children—all these intended or unintended features of the 'Washington consensus' require massive legal changes. Because these changes take place at the end of long periods of state intervention in economic and social life (notwithstanding sizable differences across the world system) the shrinking of the state cannot but be achieved by wide-ranging state intervention. The state must intervene in order not to intervene.

3. The Law of Regional Integration

This type of transnationalization of the legal field occurs whenever a group of states combines to create supranational institutions and legal competences which will assume direct regulatory functions that previously did not exist or, if they did, were performed by the individual states as prerogatives of their sovereign powers. Though there have been attempts at regional integration in different parts of the world system, in South and Southeast Asia, in the Middle East and the Magreb, and in the Americas (lastly, the NAFTA), the European Union remains so far the most advanced instance of regional integration and, for that reason alone, deserves to be singled out. Built over four decades by a mixture of diplomacy and interstate democracy, of intergovernment politics and supranational institutionalism, the EU is today an innovative transnational entity both in legal as well as in political terms (see *European Union Law*).

According to the Treaty of Rome, the balance between the market and regulation was to be accomplished by the member states. The harmonization of laws that should have ensued failed, not only because the member states revealed a nationalistic bias, but also because the legislative processes at the European level suffered recurrent bottlenecks. As a consequence, the European Court of Justice, in co-operation with the Commission, had to assume the role of decision maker, a role that has been very active indeed. Performed both with a federalist and a deregulatory bias, it has favored the Union to the detriment of individual members, and the market to the detriment of the state. This development is significant in itself because it reveals a new dimension of the relations between the legal and the political fields. The judicial engineering carried out by the Court from the 1960s onwards decisively affected the European integration process. Through the doctrines of direct effect and supremacy of the European law, the court 'constitutionalized' the community legal structure as a legal order of its own. Much of this occurred without political debate, as a judicial rather than as a political process, which led Weiler (1991) to state the apparent paradox that while European law developed firmly in a continuing process of evolution, the Union went through one political crisis after another.

4. Global Capital's Own Law: Lex Mercatoria

Understood as a set of customary principles and rules which are widely and uniformly recognized and applied in international transactions, *lex mercatoria* or law merchant is probably the oldest form of transnationalization of the legal field (see *Lex Mercatoria*). *Lex mercatoria* was a supranational law whose most distinct features were the following: the ease with which it permitted binding contracts; the stress on

security of contracts; the speed of adjudication; the variety of mechanisms for establishing, transmitting, and receiving credit; the normative value of customs and usages of the commercial world. Medieval *lex mercatoria* underwent profound changes in the modern period, which indeed caused its seemingly irreversible decline. On the one hand, as trade expanded and more and more diverse trading communities emerged, customs changed and became more diversified themselves. The diversity of customs made *lex mercatoria* less predictable, less transparent, and thus vulnerable to the critique of lacking impartiality. On the other hand, as the modern states gained control over their territories, the existence of a deterritorialized law, which did not derive its normative claims from treaties among sovereign states, was viewed as a threat. As a result, merchant courts were often either forbidden or assimilated into the domestic court system, and transnational business customs were made to comply with law. The expansion of transnational practices and the need to protect them legally led the states to develop a private international law. Since this was a national state law and could therefore conflict with the private international law of other states, efforts were made to harmonize these bodies of laws through the creation of international uniform laws.

The scope of uniform laws has always remained very limited and unable to account for the tremendous growth, in number, complexity, and variety, of transnational contracts and other business transactions, especially after World War II. These relationships, which committed generally larger amounts of money for longer periods than domestic relationships ever did, and engaged partners often separated by great distances, and cultural and linguistic differences, had a greater degree of insecurity and required, therefore, a common normative foundation. Such foundation was sought in a deterritorialized set of normative principles and rules, expressed in such formulas as 'general common principles,' 'principles of equity,' 'principles of good faith and good will,' 'principles of international law,' 'international trade usages,' etc. Because such normative references aimed at circumventing submission to national laws and to the traditional conflicts of law, transnational contracts were deemed to be self-regulatory, subjected only to their own provisions: *contrats sans loi, contratti senza legge, rechtsordnunglose Verträge*. In fact, a new supranational legal order was emerging. It was the new *lex mercatoria*, which was to expand enormously in the last quarter of the twentieth century, by reason of the sheer intensification of transnational transactions and in the wake of a new worldwide regime of accumulation in search of adequate institutional structures. The new *lex mercatoria* comprises several elements, including general principles of law recognized by commercial nations, rules of international organizations, customs and usages, standard form contracts, and reports of arbitral awards. Though this

is disputed, *lex mercatoria* may be conceived of as comprising also uniform laws and public international law. Concerning the latter, though public international law governs the relations among nation-states, and not among private parties, *lex mercatoria* shares with it the 'general principles of law recognized by civilized nations,' such as *pacta sunt servanda*, *rebus sic stantibus*, and the prohibition of undue enrichment. But the relation between *lex mercatoria* and public international law may also be conflictual whenever the first is used to create forms of immunity which operate both vis-à-vis the national law and the public international law. As regards international uniform laws, once they have been adopted by the nation-states and become part of the national law, they are not, in themselves, part of *lex mercatoria*. Before that process of adoption and to the extent that they represent a crystallization of international trade usages, such laws are *lex mercatoria*.

However informal, the new *lex mercatoria* is neither amorphous nor neutral, the customs and usages being not necessarily universal and much less traditional or immemorial. The new *lex mercatoria*, as an emerging transnational legal field, is constituted by thick cognitive expectations and thin normative allegiances reproduced by the routine repetition of myriads of transnational contractual relations originally framed by transnational corporations and their lawyers, international banks, and international organizations dominated by both. Depending on the power relations between the parties and their stakes in the transaction, the new *lex mercatoria* can operate in either a high rigidity mode (the iron cage mode) or in a high flexibility mode (the rubber cage mode). Even in its rigidity mode, *lex mercatoria* is often ephemeral, an instantaneous customary law, so to speak, as in the case of a new type of contract drafted by a leading transnational corporation and its lawyers, which must be accepted as drafted by the weaker partner, irrespective of the fact that such contract type may not be used again in its entirety. Two issues raised by *lex mercatoria* deserve special attention: *lex mercatoria* and the world system and *lex mercatoria* and legal cultures.

4.1 *Lex mercatoria and the World System*

It has been claimed that transnational contracts are purely contractual in that they contain their own rules of recognition and validation and that, as a result, *lex mercatoria* is nonpolitical, needing no reference to noncontractual elements to sustain itself as a normative order. Grounded on a legal formalistic reasoning, this conception unduly abstracts from the hierarchies and unequal exchanges that characterize the world system. A relative balance of power between partners in the transactions is only rarely obtainable. In most cases, the dominant practices at the basis of

this global law are the practices of the dominant actors. The most prominent among these do indeed enjoy state-like political prerogatives, immunity privileges, special access to political resources (tax incentives, special rights on infrastructures, etc.), and they often negotiate directly with the host states the conditions under which the private transactions will be carried out. The political facilitation made possible by the state is a crucial noncontractual dimension of the transnational contracts.

Since stable property rights are exceedingly difficult to establish across national boundaries, the international extension of capital, which is an original feature of the modern world system, has always been an inherently problematic activity. This interstate legal field, created by the core states and imposed on the semiperipheral and peripheral states, has changed substantially in the last 150 years, but has prevailed as a framing, boundary-setting legal and political structure within which *lex mercatoria* has evolved apparently unencumbered by noncontractual constraints. The use of overt or covert interventionary force and economic sanctions by the core countries in the case of expropriations or debt renunciations in the periphery and semiperiphery has been for a long time an integral part of an 'international regime,' closely reflecting the hierarchies in the world system.

4.2 *Lex Mercatoria and Legal Cultures*

The internationalization of capital has always been the motor behind the development of *lex mercatoria*. *Lex mercatoria* is, thus, basically a transnational business law or, more broadly, an economic law. Within the national legal field, economic law traditionally has been one of the laws most permeable to foreign influences and transplants, an area of law in which the ethos of the national legal cultures has always been less determinant. Not by mere coincidence has this area of law been the area of harmonization, uniformization and conventionalization *par excellence*. For this reason, and also because of its deterritorialized character, *lex mercatoria* has often been considered as the expression of a 'global legal culture,' a kind of 'third culture,' independent of the various national legal cultures and hovering above them in its specific domain.

If a new regime of global capital accumulation is indeed emerging, its most visible institutional features, aside from *lex mercatoria* itself, are the rise of transnational corporations and the international organizations that back up their activities, the adaptation of the nation-state regulatory legislation according to the transnational finance capital requirements, the globalization of the market for legal services, and the rise of international commerce arbitration. Particularly visible is the American or American-style law firm. It stands at the core of the

new international market in consultancy and its rise to prominence contributes to the Americanization of international legal services. The type of juridicization it produces is symbiotic with the economic relations it covers and leads toward the infusion of high level competition and market imperatives in the practice of law itself.

5. *Transnational Third Worlds: The Law of People on the Move*

Greatly increased cross-border movements raise myriad sociolegal issues, from international contracts, binational marriages, adoptions of foreign children, legal protection of tourists, and cross-border consumer rights, to civil, political, and social rights of legal and illegal foreign migrant workers, refugees, and asylum seekers. But the international community has paid relatively little attention to the movement of people across national borders, compared with the elaborate regulation of the movements of goods and services. Since traditional international private law only covers a very small number of issues, the international movements of people are in many respects a legal no-man's land. As a result, the legal protection of human beings remains much more territorialized than the legal protection of goods and services. International movements and interactions of people, thus, imply a net loss of legal protection. The challenges posed by this 'black hole' heavily underscore the need to analyze the movements of people across national borders, particularly of those who are less autonomous and run higher risks in moving across borders: migrants and refugees.

5.1 *International Migration*

The slave trade was the first major form of international migration in the modern world system. From the nineteenth century until the 1960s, a period of dramatic emigration from Europe, the major form of international migration emerged based on migrant recruitment through economic inducements. This form of international labor flow was at mid-point between the coerced labor extraction of the slave trade and the self-initiated or spontaneous flows that came to dominate in more recent years.

The state and the national boundaries it controls have, thus, performed crucial roles in the creation of international migration regimes throughout the history of modern capitalism. This control has become more central lately in light of the convergent impact of two recent trends. On the one hand, the trend of core states to favor, if at all, only temporary migration and only in restrictive terms (available only to those whose skills and expertise are in demand), with the consequent tighter control over borders; on the other hand,

the increased pressure to emigrate and the potential for large-scale population movements motivated by the struggle for survival resulting from the immiseration of significant regions of the world system and the growing disparities between the North and the South. Whence the potential increase of illegal, clandestine, or undocumented migration and the consequent strengthening of the receiving states as monopoly holders of the legal status of migrants and of migrant life chances and expectations attached to it.

No international consensus has been reached on the possible terms of a compromise between the principles of international human rights applicable to nationals and nonnationals, and to legal and illegal aliens, on the one hand, and the principles of territorial sovereignty, on the other; in other words, no international migration regime has come into existence.

5.2 *Refugees and Displaced People*

As in the case of undocumented migrant workers, refugees concern us insofar as they raise specific new issues in the broad conceptualization of the transnationalization of the legal field outlined here. Such issues have to do, in this case, with two clear-cut distinctions upon which the international refugee regime was built in the postwar period: between voluntary and involuntary international migration, on the one hand, and between economically motivated and politically motivated migration, on the other. These two distinctions have collapsed in recent years, thereby producing new inadequacies between the hegemonic principles of territorial sovereignty and the dynamics of transnational movements.

According to the United Nations Convention of 1951 and 1967, a refugee is any person who, 'owing to a well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group or political opinion, is outside the country of his nationality and is unable or, owing to such fear, is unwilling to avail himself of the protection of that country; or who, not having a nationality or being outside the country of his former habitual residence as a result of such events, is unable or, owing to such fear, is unwilling to return to it.' This definition excludes most of the world's displaced populations who are mainly victims of massive conflicts, human rights abuses, civil wars, external aggression and occupation, foreign domination, etc. Since they have not been personally persecuted and often have not crossed borders, they are not entitled to the refugee status, even though the political nature of the causes of their flight offers no doubt. They are considered internally 'displaced persons' and at the most may receive some humanitarian assistance. The core countries repeatedly have expressed their opposition to an expansion of the concept of refugee; in recent years, when

confronted with the potential rise of applications for asylum, they have even adopted more restrictive legislation on (or rather, against) refugees and asylum seekers.

The last three decades of the twentieth century have witnessed other 'involuntary' flows of people still more at odds with international and domestic refugee legislation: hundreds of thousands of people fleeing from famine, starvation, and natural disasters (droughts, earthquakes, volcano eruptions). In many cases extreme poverty and environmental crises have acted together with civil war or political repression to provoke massive displacements of people. In these dramatic situations the distinction between economic and political factors have been blurred, if not totally dissolved.

One type of 'involuntary' transnational migration is most likely to grow in the coming years. It will be caused by environmental catastrophes, directly or indirectly caused by anthropogenic climate changes and often coupled with over-population. Refugees will be environmental refugees, seeking ecological asylum. Environmental refugees occupy the outer fringes of a generally precarious system of international protection. Environmental refugees are most likely to collide with the territorial logic underlying social protection as a matter of right.

The high vulnerability of refugees derives from a double adverse condition. On the one side, international human rights applicable to them are highly restricted by a highly restrictive definition of refugees that leaves out most of the current and massive situations of forced migration. On the other side, the few rights they are entitled to are grossly violated with almost total impunity. Refugees are often confined to camps and detention centers where, aside from the lack of freedom of movement and access to the outside world, refugees are subjected to the most degrading human conditions.

6. *Ancient Grievances and New Solidarities: The Law of Indigenous People*

The collective rights of indigenous peoples, is surely an old issue—at least, as old as European expansion from the fifteenth century onwards—which, however, gained a new prominence in the last quarter of the twentieth century. This new prominence is due to a transnational-local linkage which has to do with the most recent step of the secular trend of capitalist world economy to incorporate the whole planet, even its most remote regions, in the process of capital accumulation. As a result, many indigenous peoples saw their traditional lifestyles, customary rights, and peasant economies threatened by new or more powerful development claims by multinational corporations or the national state or, in most cases, the two combined. They represent another and probably more

powerful wave of imperial globalization. One of its specific features is the extent to which the national state itself intervenes as a transnational actor. Pressed by the 'Washington consensus' on debt, stabilization, and structural adjustment, the peripheral and semi-peripheral states embark in new 'development projects,' often in joint ventures with transnational corporations and with the support of international financial institutions, be they extractive—oil, minerals, biodiversity—agro-industrial, road construction, or hydroelectric projects. Some of these projects are to be carried out on indigenous lands, further destroying their traditional lifestyles and economies, forcing them into new and massive relocations.

What is truly new in the new prominence of the indigenous question is the emergence of a transformative, transnational coalition consisting of indigenous and nonindigenous NGOs, which have been drawing the attention of world public opinion to the violations of indigenous historical rights and pressing both international organizations and the national states to condemn the violations, stop the destruction, and take active measures to effectively right (or begin to right) such massive historical wrongdoing. The struggles of the indigenous peoples thereby are transnationalized. The legal dimension of this transnational transformative action lies in the struggle for the transnational recognition of indigenous collective rights and local autonomous legal regulation within territorially defined areas.

7. *Cosmopolitanism and Human Rights*

In the post-World War II period a new cosmopolitan, legal culture emerged out of a transnational understanding of human suffering and the transnational constellation of progressive (legal, political, humanitarian) social actions devised to minimize it; it has gradually evolved into an international human rights regime supported by transnational coalitions of local, national and transnational nongovernmental organizations, which have been growing dramatically, in number, range, and effectiveness in recent years.

The international human rights regime is built around three major documents and sets of standards which together comprise 'the global human rights regime,' a system of rules and implementation procedures centered on the United Nations, and whose main organs are the UN Commission on Human Rights and the Human Rights Committee.

This global human rights regime allows for internal differentiation across the interstate system. Four regional subregimes usually are identified and ranked as follows by decreasing order of strength: the European, the Inter-American, the African, and the Asian and Middle-Eastern (lack of) regime. What distinguishes the *European regime*—established by the members of the Council of Europe in 1950—is the very

strong monitoring powers of the European Commission of Human Rights and the binding decision-making powers of the European Court of Human Rights. Without questioning the overall strength of the European human rights regime when analyzed in comparative terms, one of its weaknesses, which is indeed bound to expand in the coming years, refers to the massive violations of human rights that may be in store for noncommunity citizens, as the internal borders are eliminated and the external borders strengthened.

The *Inter-American regime* is substantively narrower than the European one. Similarly to the European regime, its main organs are the Inter-American Commission of Human Rights and the Inter-American Court of Human Rights. In practice, however, the two institutions have been operating very differently from their European counterparts. As to the *African regime*, the African Charter on Human and Peoples' Rights (1981) contains two major innovations: as the title of the Charter itself indicates, to individual human rights it adds collective rights, as well as the right to development; it also introduces the concept of duty of individuals vis-à-vis the family, the community, and the state, but the provisions that establish the African Commission on Human and Peoples' Rights are exceedingly vague, conferring it no enforcement powers of any sort and only very modest investigatory powers, and there is no regional human rights court. Though in *Asia* and *The Middle East* the human rights debate is widening and deepening, there are, among other reasons, important cultural barriers to the establishment of a human rights regime. Human rights are often viewed as an exotic, foreign, Western conception whose worldwide circulation is nothing but a manifestation of the overall cultural imperialism of the West.

International regimes refer to international normative consensus among nation-states and, as such, they raise two interrelated questions: the extent to which the normative consensus collapses whenever the overriding imperatives of national sovereignty are considered to be better served by the violation of human rights; and the extent to which the inherent statism of implementation and enforcement mechanisms stays in the way of the emergence of new international legal subjects with a more cosmopolitan orientation and a transnational advocacy of greater efficiency. The optimistic predictions of a brighter future for human rights, in the wake of democratic transitions and the collapse of the communist bloc, may prove to be displaced and utterly wrong themselves.

8. *The Global Commons: Jus Humanitatis*

Of all forms of transnational legality *jus humanitatis* is the one that pushes the idea of globalization further, for it takes the globe itself as the object of its

regulation. *Jus humanitatis* expresses the aspiration to a form of governance of natural or cultural resources which, given their extreme importance for the sustainability and quality of life on earth, must be considered as globally owned and managed in the interest of humankind as a whole, both present and future. In this sense, *jus humanitatis* clashes with two fundamental principles of the dominant paradigm: property, upon which the capitalist world system is based; and sovereignty, upon which the interstate system is based. Little wonder, then, that its application has been so little so far. Nevertheless, its prolegomena are already quite visible in the doctrine of the *common heritage of humankind* that has been adopted by international law. Though the doctrine has its own limitations, it is grounded on principles which, were they to be fully developed, would bring about the bankruptcy of the dominant paradigm, and hence of international law itself as it is conceived today.

The concept of the common heritage of humankind was formulated for the first time in 1967 by Malta's Ambassador to the United Nations, Pardo (1968), in relation to UN negotiations on the international regulation of the oceans and the deep seabed. Pardo's purpose was 'to provide a solid basis for future worldwide cooperation ... through the acceptance by the international community of a new principle of international law ... that the seabed and ocean floor and their subsoil have a special status as a common heritage of mankind and as such should be reserved exclusively for peaceful purposes and administered by an international authority for the benefit of all peoples.' Since then, the concept of the common heritage of humankind has been applied not only to the ocean floor but also to other 'common areas' such as the moon, outer space, and Antarctica. Five elements usually are associated with the concept of the common heritage of humankind: nonappropriation; management by all peoples; international sharing of the benefits obtained from the exploitation of natural resources; peaceful use, including freedom of scientific research for the benefit of all peoples; conservation for future generations.

The concept of the common heritage of humankind transcends by far the field of traditional international law. International law deals traditionally with international relations among nation-states, which are supposed to be the main beneficiaries of the regulation agreed upon. Such relations are based chiefly on reciprocity, that is, granting advantages to another state or states in return for equivalent advantages for oneself. The concept of the common heritage of humankind is different from traditional international law on two accounts: as far as the common heritage of humankind is concerned, there is no question of reciprocity; and the interests to be safeguarded are the interests of humankind as a whole, rather than the interests of states. The Law of the Sea Convention, signed in Montego Bay on December 10, 1982,

contains in part XI the most developed formulation of the concept of the common heritage of humankind to date.

See also: Global Governance; Globalization and World Culture; Globalization, Anthropology of; Globalization: Geographical Aspects; International Justice: Philosophical Aspects; International Law and Treaties; International Relations, History of; International Relations: Theories; International Trade: Commercial Policy and Trade Negotiations; International Trade: Economic Integration; Markets and the Law; Science and Technology: Internationalization; World Systems Theory; World Trade Organization

Bibliography

- Bilder R 1992 International law in the 'New World Order': Some preliminary reflections. *Journal of Transnational Law and Policy* 1: 7–21
- Brysk A 2000 *From Tribal Village to Global Village. Indian Rights and International Relation in Latin America*. Stanford University Press, Stanford, CA
- Cohen R, Deng F 1998 *Masses in Flight. The Global Crisis of International Displacement*. Brookings Institution Press, Washington, DC
- Dezalay Y, Garth G B 1996 *Dealing in Virtue. International Commercial Arbitration and the Construction of a Transnational Legal Order*. University of Chicago Press, Chicago
- Donnelly J 1992 Human rights in the New World Order. *World Policy Journal* 9: 249–77
- Draetta U, Lake R, Nanda V 1992 *Breach and Adaptation of International Contracts. An Introduction to Lex Mercatoria*. Butterworth, Salem, NH
- Harvey C J 2000 Dissident voices: Refugees, human rights and asylum in Europe. *Social & Legal Studies* 9: 367–96
- Hathaway J C (ed.) 1997 *Reconceiving International Refugee Law*. Kluwer, The Hague
- Joyner C 1986 Legal implications of the concept of the common heritage of humankind. *International and Comparative Law Quarterly* 35: 190–9
- Keck M, Sikkink K 1998 *Activist Beyond Borders. Advocacy Networks in International Politics*. Cornell University Press, Ithaca, NY
- Kiss A 1985 The common heritage of mankind: Utopia or reality? *International Journal* 40: 423–41
- Pardo A 1968 Whose is the bad of the sea? *American Society International Law Proceedings* 62: 216–29
- Picciotto S 1997 Fragmented states and international rules of law. *Social & Legal Studies* 6(2): 259–79
- Rogers R 1992 The politics of migration in the contemporary world. *International Migration* 30: 33–52
- Santos B de Sousa 1997 Toward a multicultural conception of human rights. *Zeitschrift für Rechtssoziologie* 18: 1–15
- Sassen S 1999 *Guest and Aliens*. New Press, New York
- Tigar M, Levy M 1977 *Law and the Rise of Capitalism*. Monthly Review, New York
- Van Cott D L 2000 A political analysis of legal pluralism in Bolivia and Colombia. *Journal of the Latin American Studies* 32: 207–34

Weiler J 1991 The transformation of Europe. *Yale Law Journal* 100: 2403–83

B. de Sousa Santos

Globalization: Political Aspects

Globalization is the process by which the economies of many different nations and regions are being transformed into a single, integrated economy which is worldwide or global in scope. The elimination of national barriers and the expansion of international transactions are being carried to their ultimate conclusion, where the exchange of goods, capital, and services occurs within a global market. This profound economic transformation is having great consequences for national states and their societies. Some nations and groups are winners in globalization and some are losers. The result is that globalization is accompanied by major political actions and by intense intellectual debate.

1. The Term 'Globalization'

The term 'globalization' came into common usage in the early 1990s. The decline of Marxism, the end of the Cold War, and finally the collapse of the Soviet Union in 1991 meant that there was no longer any formidable ideological and political competitor to liberal democracy and the free market. The end of the socialist or Second World meant that the capitalist or First World could expand into the entire world or globe, hence 'globalization.' Also in the 1990s, new information technologies reached the critical mass where they became an entire information economy. The conjunction of computers, telecommunications, the internet, and the worldwide web greatly facilitated and expanded the exchange of goods, capital, and services on a global scale.

Globalization is often described as a process: steadily progressing over time, pervasively spreading over space, and clearly inevitable in its development. In this conception, the economics of globalization are all-important; the politics of globalization are practically nonexistent. But globalization can also be seen as a revolution, one of the most profound revolutions the world has ever known. (It is certainly more of a world revolution than communism ever was.) Revolutions produce reactions, however, and the interactions between revolutionaries and reactionaries produce a particular kind of politics. The first decades of the twenty-first century are likely to be dominated by the politics of globalization, so much so that they could give the name to the era, in the same way that the first

half of the twentieth century was known as the era of the World Wars and the second half was known as the era of the Cold War.

2. Previous Eras of Globalization

The contemporary period is not necessarily the first era of globalization; in some ways it can be seen as the third. There have been at least two previous periods when there was a rapid and massive expansion in economic exchange on a world, even global scale. Consideration of these earlier eras may help in understanding the current one.

The first era of globalization was in the sixteenth century, at the time of the commercial revolution and when European powers opened up the Americas, Africa, and Asia to new or greatly expanded trade. Facilitated by new technologies in shipbuilding and navigation, a vast system of seaborne commerce in goods, capital, and slaves came into being, connecting five continents. This first global economy did not come into being by the spontaneous operations of the free market, however. It had to be promoted and protected by a great political and military power, one that was worldwide in scope. This global hegemon was the Habsburg empire, whose centers were Spain and Austria but which ruled large territories on or near the five continents (Wallerstein 1974, 1980).

The first wave of globalization washed over many old and traditional political authorities. Many feudal and local entities were diminished or disappeared entirely, but others reinvented themselves to become absolutist monarchies and national states. The process was not steady and gradual, but disruptive and violent. It contributed to the Wars of Religion and culminated in the catastrophe of the Thirty Years War (1618–48). At the end of that war and with the Peace of Westphalia, the major surviving powers created an entirely new political order, one composed of sovereign and independent states. Most of these states now developed systematic policies (known as mercantilism) to protect local interests from foreign imports. Thus, the first era of globalization ended in what can be seen as a new era of localization.

The second era of globalization was in the nineteenth century, at the time of the industrial revolution and when European powers and the USA opened-up even larger (now inland as well as coastal) areas of the Americas, Africa, Asia, and Australia to new or greatly expanded trade. Facilitated by the steamship, the railroad, and the telegraph, a vast system of seaborne and overland commerce in goods, capital, and immigrant labor connected six continents. This second global economy also did not come into being by the spontaneous operations of a free market (although the economic theories of the time often assumed that it had). It had to be promoted and

protected by a great political and military power that was worldwide in scope. This global hegemon was the British Empire, which in the late nineteenth century ruled one-fourth of the land surface of the globe and which was the supreme naval power in every ocean (Polanyi 1944).

The second wave of globalization also washed over many old and traditional political authorities. Again, many local entities were diminished or disappeared entirely (particularly in Africa, the Middle East, and East Asia), but others reinvented themselves to become strong national states (particularly in Europe but also Japan). The process was not steady and gradual, but disruptive and violent. It contributed to a series of nationalist wars and culminated in the catastrophic World War I and II, which some historians have called a second thirty years war (1914–45).

At the end of that war and with the onset of the Cold War, a bipolar system was created, composed of the USA (liberal-democratic in its politics and free-market in its economics) and the Soviet Union (authoritarian in its politics and statist in its economics). The Americans promoted free markets and open economies in their alliance system (the Free or First World); the Soviets promoted command and closed economies in theirs (the Socialist or Second World). The existence of the Soviet Union and the Second World acted to contain the spread of the free-market policies of the USA, even within the First World itself. Thus, the second era of globalization ended in another era of economies that were closed or protected in important ways.

3. The USA as the Leader of Globalization

We are now in the third era of globalization. With the collapse of the Soviet Union, the USA could pursue its grand project of promoting and protecting free markets throughout the globe, unconstrained and uncontained by any comparable power or 'peer competitor'; it has been the leader and the energizer of the process of globalization. It serves as the new global hegemon, the successor to the Habsburg and the British Empires. Acclaimed as 'the sole superpower,' the 'high-technology economy,' and 'the universal nation,' the USA has vigorously and consistently supported this great transformation by systematically pressing to remove any national barriers to the free movement of capital, goods, and services. Its leadership is facilitated and legitimated by the three great international, now global economic institutions, the International Monetary Fund (IMF), the World Bank, and the World Trade Organization (WTO).

US leadership in globalization is also facilitated and legitimated by the particular American worldview or ideology. Americans have long advocated liberal

democracy and free markets as political and economic ideals. In foreign affairs, this ideology has been known as liberal internationalism, and its most prominent proponents were Presidents Woodrow Wilson and Franklin D. Roosevelt (T. Smith 1994). This familiar ideology has been revised in the era of globalization into a new American ideology which promotes human rights and the open society. This is a version of liberal internationalism, one that is more accurately termed liberal globalism; its most prominent proponents in the 1990s were President Bill Clinton and Secretary of State Madeleine Albright, but it is held by most American political, economic, and intellectual elites. The ideology of human rights and the open society advocates the limitation of state sovereignty and the weakening of the nation state. As such, it poses a fundamental challenge to traditional conceptions of international relations.

This ideology of human rights and the open society—liberal globalism—also provides the justification for a new kind of US military operation. This is humanitarian intervention, as in Somalia, Haiti, and Bosnia; the culmination of this development was the US-led NATO war against Serbia over Kosovo. Intervention in support of human rights is not supported by all parts of the American public; it is primarily advocated by political and intellectual elites, especially those in the media and in academia. Political opposition to humanitarian interventions and more generally to globalization comes from other American groups and interests, especially those who see themselves as losers in the global economy (e.g., labor unions and domestic industries). But the major sources of political opposition to globalization are found outside the USA.

4. The Resistance to Globalization

The US-led revolution of globalization has generated resistance and reaction. Immediately after the collapse of Soviet communism, the principal sources of resistance seemed to be major nation states, especially Japan, France, and reunited Germany. However, by the end of the 1990s these nations offered little resistance to globalization; one reason was the stagnation of their economies, compared with that of the USA.

With the paralysis of the major nation states, the principal sources of resistance to globalization have become several of the great religions. Some of these have undergone a revival and transformation into a political religion. Especially strong in their resistance have been revivalist Islam and neo-Confucianism (sometimes referred to as 'Asian values'). Also resistant have been revivalist Hinduism and Eastern Orthodoxy. The great religions see the globalization led by the USA to be closely connected with secularization and therefore a threat to themselves. These

four sources of religious opposition correspond to four large countries or areas. Two of these, Russia and the Middle East, have been losers in the new global economy, and two, China and India, have been winners.

Russia and, more generally, most countries with an Eastern Orthodox religious tradition, for a variety of reasons have been unsuccessful in making the transition from a communist regime to a liberal democracy and market economy, which are able to adapt well to an open society and global economy (e.g., Russia, Belarus, Ukraine, Moldova, Romania, Bulgaria, and Serbia). Conversely, most countries with a Roman Catholic or Protestant tradition have been successful in making this transition. This contrast among ex-communist countries, between the more western and the more eastern, between the Roman Catholic or Protestant and the Eastern Orthodox, means that the political and economic developments of the 1990s revived and reinforced an historic divide, one that corresponds to the great schism between Western Christianity and Eastern Christianity and even to the ancient division between the Latin and the Greek halves of the Roman Empire.

With their condition of political and economic weakness, the governments of Russia and the other Orthodox countries cannot mount an effective and sustained opposition to globalization. But among the populations of these countries, there is now substantial resentment and resistance directed at the USA and its global project. The Middle East and, more generally, most countries with an Islamic religious tradition have also been unsuccessful in establishing a viable liberal democracy and market economy, which are able to adapt well to an open society and global economy. For those that experienced the heady years from 1973 to 1985, when OPEC seemed to be a world power, a long decline in oil prices deepened their sense of failure and their alienation from globalization. Here, too, the governments are not willing and able to mount opposition that is effective and sustained. But among their populations there is widespread resentment and resistance, whose most obvious representatives are militant organizations based upon Islamic revivalism.

Conversely, there is growing opposition even from some of the winners in the global economy. These include China and, more generally, most countries with a Confucian tradition. Most Chinese, both in China itself and among the 'overseas Chinese,' see their economic success as resulting from their own culture or 'Asian values,' from engaging in the global economy in their own way. This conception has been reinforced by their interpretation of the Asian economic crisis of the late 1990s. The countries with the most open currency markets (South Korea, Thailand, Indonesia) suffered the greatest disruptions and declines. Those whose currency markets were most regulated (China and, ironically, Taiwan) experienced little disruption and continued their economic growth.

The entry of India into the global economy has been more recent, and its benefits have been unevenly distributed. It has occurred, however, at the same time as the growth of Hindu nationalism, the development of Indian nuclear weapons, and an increased confidence in rejecting certain Western ways. The sources of opposition to globalization led by the USA are thus very disparate. But together these sources of opposition comprise a vast region, really all of Eurasia and more, which stretches from Russia and Eastern Europe, through the Middle East, through South Asia, to China and East Asia. In this vast region are found the four great civilizations that Samuel Huntington has identified as most likely to oppose the West in 'the clash of civilizations' (Huntington 1996). These are, in his terms, 'Slavic-Orthodox,' Islamic, Hindu, and 'Sinic-Confucian' civilizations. In this region, too, are four nuclear powers, Russia, Pakistan, India, and China, each of them seeing themselves as the center of their civilization. Singly or collectively, they may develop the capacity and will to disrupt and contain globalization. If so, it will take American leadership of the highest ability to insure that the third era of globalization does not end with a clash of civilizations which culminates in a third thirty years war.

See also: Cold War, The; Global Governance; Global Population Trends; Globalization and World Culture; Globalization, Anthropology of; Globalization: Legal Aspects; Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity; International Relations, History of; World Systems Theory

Bibliography

- Cox R W (ed.) 1996 *Approaches to World Order*. Cambridge University Press, Cambridge, UK
- Eichengreen B 1996 *Globalizing Capital. A History of the International Monetary System*. Princeton University Press, Princeton, NJ
- Friedman T 1999 *The Lexus and the Olive Tree*. Farrar, Straus, and Giroux, New York
- Gray J 1998 *False Dawn: the Delusions of Global Capitalism*. New Press, New York
- Held D, McGrew A, Goldblatt D, Perraton J 1999 *Global Transformations: Politics, Economics, and Culture*. Stanford University Press, Stanford, CA
- Hoekman B M, Kostecki M M 1995 *The Political Economy of the World Trading System: From GATT to WTO*. Oxford University Press, Oxford, UK
- Huntington S P 1996 *The Clash of Civilizations and the Remaking of World Order*. Simon & Schuster, New York
- Kapstein E B 1994 *Governing the Global Economy: International Finance and the State*. Harvard University Press, Cambridge, MA

- Luard E 1990 *The Globalization of Politics*. Macmillan, London
- Polanyi K 1944 *The Great Transformation*. Beacon Press, Boston
- Smith A D 1995 *Nations and Nationalism in a Global Era*. Polity Press, Cambridge, UK
- Smith T 1994 *America's Mission: The United States and the Worldwide Struggle for Democracy in the Twentieth Century*. Princeton University Press, Princeton, NJ
- Strange S 1996 *The Retreat of the State: the Diffusion of Power in the World Economy*. Cambridge University Press, Cambridge, UK
- Wallerstein I 1974 *The Modern World-System: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century*. Academic Press, New York
- Wallerstein I 1980 *The Modern World-System II: Mercantilism and the Consolidation of the European World-Economy 1600-1750*. Academic Press, New York

J. Kurth

Globalization, Subsuming Pluralism, Transnational Organizations, Diaspora, and Postmodernity

Globalization, the historical process by which all the world's people come to live in a single social unit, brings with it religious changes as well. Prime among these is that globalization encourages religious pluralism such that religions identify themselves in relation to one another. Religions also become less rooted in particular places because of diasporas and transnational organization. In postmodern fashion, global religious pluralism creates a situation in which absolute visions end up living side by side.

1. Globalization and Religion

Since the 1980s, the term globalization has taken on a certain currency in social scientific as well as in popular discourses. Its meaning in both these areas varies; but what unites economic, political, cultural, and even theological conceptions of globalization is the sense that, for better or worse, all people around the world live today in a single social unit much more than they ever have before. There is now and has been developing for some time a world society which serves as the encompassing social context of all more local social units. In this regard, most of the debates about globalization, whether scholarly or popular, have centered around the relation and even polarity between the global and the local. Robertson (1992), perhaps the first social scientist to use the term globalization in a technical and consistent sense, talks about the dialogical relation between the universal and the particular. What this means is that the global

context, rather than only leading to a certain homogenization of culture and social structure around the world, also provides new conditions for constructing social and cultural difference. Global society, while single, is inherently pluralistic, not only permitting but even favoring a multitude of local and particular identities, both individual and collective.

The status of religion in debates about globalization has varied, to a significant extent according to academic discipline. Perspectives that see the development largely in economic terms, for instance as the increasing dominance of global capitalist institutions, tend to ignore religion almost entirely. Many of the approaches originating in the fields of international relations, political geography, and communication studies also pay little attention to religion, with the occasional exception of the role of Islam in many Muslim states (see, e.g., Anderson et al. 1995). Overall, it is primarily the works of certain anthropologists and sociologists that accord religion a more visible place. Yet even here, analyses vary a great deal, for the most part depending on how the term is understood. Two overlapping lines of variation appear to be central: the distinction between functional and substantive notions of what counts as religion and whether one sees religion as an aspect of culture or as a differentiated and independent social domain.

A functional conception of religion regards as religion any social process that addresses certain problems, for instance the question of the ultimate meaning of life, or social integration and control. Here religion, or at least the religious, almost automatically becomes a necessary dimension in the formation of world society because problems of meaning or of self-description (integration) have to be addressed in any society, especially one as diverse as the global one. As examples under this heading, some authors have sought to analyze the emergence of a global culture that is the equivalent of a new global 'sacred canopy' (Meyer et al. 1997). One might even include a growing discussion about the development of a global 'civil society' under this heading (Rudolph and Piscatori 1997, Braman and Sreberny-Mohammadi 1996). In addition, a functional approach allows questions of national, regional, ethnic, or civilizational identity in world society to take on religious character (Huntington 1996, Robertson 1992); and thus will tend to see religion largely as an aspect of culture. Substantive ideas of religion, by contrast, insist on the presence of identifiable and institutional sorts of activity that refer to a posited supraempirical, transcendent reality. From this perspective, the relation of various specifically religious groups to globalization becomes the focus of attention, groups ranging from the Roman Catholic Church and Islamic organizations to Western Buddhist groups and Pentecostal movements (Cox 1995, Robertson 1992, Rudolph and Piscatori 1997). Such approaches perforce also see religion as an independent domain of social activity and as more

than an expression of something else, whether economics, politics, or culture (see Beyer 1998, Hexham and Poewe 1997).

This variety in ideas about what counts as religion in the global context by itself already points to important aspects of the relation between globalization and religion. The fact that most of the literature which deals explicitly with globalization ignores religion or functionalizes it suggests that substantive or institutional forms are for these observers somewhere between irrelevant and antiglobal. One manifestation of this orientation is the great amount of attention that so-called 'fundamentalisms' receive, and this, ironically, very often from scholars who are oriented quite positively toward institutional or substantive religion. Thus, e.g., the volumes issuing from Appleby and Marty's *Fundamentalism Project* (see Marty and Appleby 1994), although not directly about globalization, nevertheless assume globalizing processes. They judge a wide variety of important religious movements all around the contemporary world as particularistic and usually local reactions against them (see also Juergensmeyer 1993, cf. Beyer 1994). Among some of these authors, however, designating 'fundamentalisms' as, in effect, antiglobal movements parallels a positive attitude to the idea of a single world society, especially as manifest in the search for a specifically global ethos. In these cases, the search is on an explicitly religious basis, albeit not on the basis of a particular religion. What is somewhat ironic is that the resulting global ecumenical ethic (see especially Küng 1991) corresponds reasonably well to the intention of those who functionalize religion at the global level: i.e., it avoids reference to a particular religion while addressing what it assumes to be the urgent and globally relevant religious task of providing ethical ideals and values that people all around the world can accept. In short, this sort of discussion seems to suggest that the distinction between functional and substantive conceptions of religion overlaps in a certain way those between universal/global and particular/local. Since these latter are principal organizing polarities in the entire globalization debate, it is best not to decide for one conception of religion or the other; but rather to consider the potential insights that flow from both.

2. Theoretical Perspectives

An excellent example of how a functionalized conception of religion can enter the globalization debate at the theoretical level is to be found in the work of Meyer and his co-workers. In a 1997 article that clearly presents their basic outlook, they carry to the level of global society an analysis that the classical sociology of Max Weber and Emile Durkheim made for modern societies. For these scholars, the world polity, as they call it, is an extension of modern

Western values and their corresponding rationalized structures to the rest of the world. Western religion had a critical role in bringing about this modernization and rationalization; but today the old religious elites have lost their power and have in any case also become believers in salvation through the pursuit of modern global values, especially progress (development) and equality. In their place there is now a new 'religious' elite who are 'the professionals, researchers, scientists, and intellectuals who write secularized and unconditionally universalistic versions of the salvation story, along with the managers, legislators, and policymakers who believe the story fervently and pursue it relentlessly. This belief is worldwide and structures the organization of social life almost everywhere' (Meyer et al. 1997, p. 174). Meyer and his co-workers support the existence of this world polity by showing how the concrete policies of most countries around the world conform to its goals and values. In arguing in this fashion, however, they present a rather alternativeless and homogenizing picture which leaves room neither for the assertion of any meaningful particular or local variation nor, not coincidentally, for any role for institutional religion other than as futile reaction or powerless supporter.

A significantly more nuanced vision, both as regards the question of homogeneity/heterogeneity and the possible role of institutional religion, is contained in the work of Robertson. This author also emphasizes what he calls the 'globalization of societal modernity' (Robertson 1991), which includes the sort of global religious function that Meyer and co-workers see. In addition, however, Robertson points much more positively to the proliferation of religious movements in the contemporary world, movements ranging from Latin American liberation theology and Kūngian 'ecumenism' to Islamic 'fundamentalism,' Japanese Soka Gakkai, and the Unification Church of Sun-Myung Moon. His theoretical model makes room for these more institutional and substantive forms of religion, and not just as local defensive reactions against incorporation into the global system. Instead, Robertson allows that they can also be attempts to define the 'global-human' situation as a whole; and at the same time ways of identifying a particular culture, region, or (in his terms) society in relation to the global whole and other subunits of it. The task of constructing such universalistic religious visions—whether functional or substantive—is always done on the basis of particular traditions, cultures, or movements. In consequence, the forms of institutional religion themselves change in the context of globalization, leading, for instance, to revisions of old traditions, syncretic or hybrid movements, and the rise of entirely new religions. For Robertson, then, the World Parliament of Religions, the Vishva Hindu Parishad (World Hindu Organization), the Church of Scientology, a peripatetic Pope or Dalai Lama, and Chinese Qi Gong groups are just as much aspects of

globality and contributors to the shape of global society as are the aims and pronouncements of social scientists and technocrats. They all reflect and share a common social context; and one is not more modern or global than the other.

3. Religious Pluralism

From these theoretical starting points, one of the important questions that emerges concerns religious pluralism. Only with a radically functionalized concept of religion, such as that of the world polity school, can it appear possible for global society to share a single religious vision, a single world view. All other conceptions point to the near impossibility of a hegemonic, global religion. Substantively speaking, contemporary global society shows no sign of moving in this direction. If anything, the opposite is the case: not only is there an increasing number of identifiable religions or religious traditions in the world, the variety of expressions within those religions is also growing. Indeed, some sociologists of religion, notably Luckmann (1967), have long suggested that the typically modern—and therefore now global—form of religious expression is a highly individualized one that carries pluralism to the point where Luckmann can call it 'invisible religion.' Put somewhat differently, the question of religious pluralism is not simply one of how different religions, each one claiming an absolute vision, can live side by side in world society; but also to what extent this society still favors authoritative institutions worthy of the name, religions. A partial answer is possible through a comparison of the idea of religion with those of culture and nation.

The model of global culture argues that, in the contemporary social world, the ideas of culture and nation serve as names for types of legitimate and fundamentally plural collective identities: cultures and nations, like individuals, have the right to equal existence. Within the context of homogenizing forces like global capitalism and mass media, one can assert one's difference and defend that difference in the name of one's culture or nation. Cultures and nations, like individuals, are legitimated agents or actors in global society: it is part of the world polity of which Meyer and co-workers speak, that their right to 'self-determination' should only be limited by the right of other such identities to assert themselves in like fashion (see Lechner 1989). Religion is both a similar and different concept. On the one hand, religion is often a part of cultural and national identities: it can help define them. On the other hand, global culture also allows that religion is something independent of these other two notions. Religions are something to which people assumedly adhere voluntarily as individuals; and this may or may not flow from their cultural or national identities. Religions also enjoy the same 'protection' as cultures and nations; at least such

freedom of religion is an ideal expressed in most of the world's state constitutions. Correspondingly, global cultural ideals regard religions, like cultures and nations, as inherently plural. Two seeming paradoxes emerge from this view. First, religious freedom is grounded in the individual's right to choose, but also in the right of the religious collectivity to exist and determine itself. Religious collectivities can therefore claim authority over their members, but only with the consent of these latter! Second, religions are particular and plural; and yet it is understood that they each offer a universal or cosmic vision, a total way of life. This holist emphasis therefore includes the possibility of exclusivist religions, ones that deny the validity of all others. Thus, in true postmodern fashion, religions provide ways of multiplying ultimate narratives about how the world is and they do this in such a way that each is regarded as of equal value, the choosing of one or the other being deemed a matter of preference, or upbringing, or culture. Ultimately, each is grounded only in the conviction of its adherents, and this in the knowledge that most others in global society have different convictions: even by the most generous estimates, the largest religion, Christianity, can claim far less than half the world's population.

4. Religion in Postmodernity

The postmodern character of religion in the context of world society is manifest in other features as well. Embedded in the idea that religion is fundamentally plural are the questions of how religions distinguish themselves from one another and what the difference is between religion and nonreligion. Neither question is susceptible to a simple and clear answer. The idea that there are a plurality of distinct religions is one that dates historically only from the European seventeenth century in the context of post-Reformation religious conflicts and as European expansion brought its people into protracted contact with a much wider variety of other cultures, civilizations, and hence religious traditions. Which traditions emerged from this historical process as 'one of the religions' depended both on what Westerners observed as religions and on what non-Westerners were willing to style according to the implied model of a religion. Different and even diametrically opposite outcomes resulted. Thus, the elite carriers of Hinduism took up the suggestion of Western observers and cooperated in reimagining South Asian religious traditions as somehow a single religion, whereas corresponding Chinese intelligentsia refused, in spite of Western observation, to reconstruct their Confucian thought and ritual traditions as a religion (see Beyer 1998). The contemporary list of (world) religions therefore has an inductive and even arbitrary character that cannot be reduced to abstract characteristics. Like nations and cultures, what comes to be seen as one of the religions

depends in large part on how well the supposed members succeed in having others accept it as one.

The boundary between religion and nonreligion exhibits comparable ambiguities, the difficulty of deciding between functional and substantive ideas of religion being one example. At the core of this problem lies the notion that religion refers to an overarching view of the world and thus to a complete way of life; and yet religions are, in principle and in fact, plural. A prime symptom of this ambiguous status of religion in world society is that, over the past century and a half, key representatives of almost all the so-called world religions have claimed that 'ours is not a religion.' Most often, protests of this sort favor a broad functional conception of religion for their tradition, and reject the sort of substantive one that limits religion to a special category of human action, say one dealing only with the interaction with gods or spirits. Similarly, opponents of various new religions have sought to exclude these from the category, in these cases so as to eliminate them from the range of legitimate religious choices. Such conflicts over categorization are, on occasion, ways of circumventing the global cultural value of 'freedom of religion,' justifying in these instances the obligatory adherence of everyone to the tenets and practices of a particular 'religion'; or preventing such adherence to another. Thus, does the category of religion in global culture allow a post-modern juxtaposition of absolute narratives without offering an overarching narrative through which to reconcile them.

5. Religious Authority, Transnational Organization, and Diaspora

Accompanying these various ambiguities regarding the category of religion and religions, the carrying out of institutional religion under conditions of globalization brings a number of other challenges and changes, especially with respect to questions of religious authenticity and authority. In any society, there exist a limited number of ways through which conformity to authoritative religious traditions can come about. Thus, e.g., in small scale societies a relative lack of alternatives, and social structures centered on face-to-face interaction could assure reasonable religious conformity. After the advent of writing and large scale political empires, dedicated priesthoods or monastic orders, political state support, along with distinctions between elite and popular religiosity (e.g., religion vs. superstition) could handle the problem according to hierarchical principles. Under conditions of globalization, neither relative social isolation nor hierarchical solutions are possible. Globalized communication media, from telephones and air travel to television and the internet assure that religious social networks need not be limited to local interaction or particular places. In fact, the existence

of these possibilities positively encourages global or translocal religious networks and religious organizations. Moreover, the global cultural valuing of equality for both individuals and collectivities severely militates against the notion that the orthodoxy practiced or advocated by a certain group of religious professionals or adherents in a certain place should serve as the unquestioned standard for all others. Even in religions that have traditional centers of such authority, e.g., Rome for Roman Catholics, al-Azhar for Sunni Muslims, or Salt Lake City for Mormons, these centers have limited resources for enforcing their authority; and those that choose to ignore them can use the same techniques of individual religious freedom, (alternative) tradition, separate organization, and transnational networking to circumvent it. The only real exceptions are situations in which state power helps enforce religious authority; but in those cases, such as Iran or Russia, the state boundaries themselves set effective limits to such compulsion.

A prime illustration of the way that globalization encourages the multiplication of local centers of religious authenticity is the effect on religion of global migration and transnational religious networking and organization. Although the migration of people from one part of the world to others is a constant of world history, what is new under conditions of globalization is both the extent of the phenomenon and the fact that migrants leave one place for another far less definitively. The possibilities for maintaining communicative and even physical ties with the places of origin are far more developed, meaning that migrants are less dependent on their new contexts and can still influence the places that they have left. From a different angle, the difference between migrants and nonmigrants is not as great as it used to be, just as the difference between travel/tourism and migration is not as clear. And the chief source of this new situation is the much greater and more rapid communicative links that now exist throughout the world.

Among the consequences for religion of migration and enhanced transnational or translocal linkages, two are of the greatest potential significance. The first is that, as migrants from a particular religion settle in different parts of the world outside of the traditional 'heartlands,' these diaspora communities will develop religious variations that respond to their new locations, and may furthermore re-export these to their places of origin as authentic and authoritative versions. Thus, e.g., the temples that Hindu migrants typically set up in diaspora communities of Europe or North America often allow the worship of a range of gods and goddesses in the Hindu pantheon, unlike in India where temples usually are dedicated to one deity only. The diasporic 'one-stop supermarket' temple is primarily a response to the fact that Hindus from different parts of India migrate to the same places and that these communities do not have the resources to build temples for each subgroup. In spite of these

strategic origins of the variation, however, the model is being transferred back to India, providing another authentic way of 'doing' this religion. Similar analyses could be made for Muslim, Christian, Buddhist, or other migrants as well.

The second major consequence for religion is related but also in a sense opposite. It consists in the possibility of tying together disparate versions of a religion across the globe, expressly in spite of the difference in local variations. The Roman Catholic church is perhaps the prime example of a religious organization that has taken great advantage of this possibility. The links among the world's Islamic populations have also strengthened significantly in this context (see Levitt 1998, Rudolph and Piscatori 1997). Another very illustrative example is global Christian Pentecostalism. Here, the concrete ways of doing Pentecostal Christianity vary enormously when one compares places like Korea, South America, North America, Asia, and Africa. Yet transnational linkages in such forms as common publications, regular world conferences, connections among professionals, and migration encourage a consciousness in most Pentecostals of being part of the same worldwide religious movement, and this in spite of the differences among them (see Van Dijk 1997).

6. Conclusion

What the foregoing demonstrates is that globalization creates and enhances conditions in which religion can become more pluralistic and yet at the same time more unified. The local and the global in religion, the universal and the particular, are not opposite poles between which religious forms oscillate. Rather local religion can now increasingly only be done with reference to the global context, especially in the awareness of religious pluralism. And global religious forms expressly include and are the expression of local manifestations. Global society does not present the choice between antiglobal, perhaps fundamentalist resistance and proglobal, liberal accommodation. Instead, it recreates the conditions under which religions form and operate in relation to one another.

See also: Diaspora; Globality; Globalization and World Culture; Globalization, Anthropology of; Multinational Corporations; Pluralism; Pluralism and Toleration; Postmodernism in Geography; Postmodernism in Sociology; Religiosity; Modern

Bibliography

- Anderson J, Brook C, Cochrane A (eds.) 1995 *A Global World? Re-ordering Political Space*. The Open University and Oxford University Press, London
Beyer P 1994 *Religion and Globalization*. Sage, London

- Beyer P 1998 The modern emergence of religions and a global social system for religion. *International Sociology* 13: 151–172
- Braman S, Sreberny-Mohammadi A (eds.) 1996 *Globalization, Communication and Transnational Civil Society*. Hampton Press, Cresskill, NJ
- Cox H G 1995 *Fire From Heaven: The Rise of Pentecostal Spirituality and the Reshaping of Religion in the Twenty-first Century*. Addison-Wesley, Reading, MA
- Hexham I, Poewe K 1997 *New Religions as Global Cultures: Making the Human Sacred*. Westview Press, Boulder, CO
- Huntington S P 1996 *The Clash of Civilizations and the Remaking of World Order*. Simon & Schuster, New York
- Juergensmeyer M 1993 *The New Cold War? Religious Nationalism Confronts the Secular State*. University of California Press, Berkeley, CA
- Küng H 1991 *Global Responsibility: In Search of a New World Ethic*. Crossroad, New York
- Lechner F 1989 Cultural aspects of the world system. In: Swatos Jr W H (ed.) *Religious Politics in Global and Comparative Perspective*. Greenwood Press, Westport, CT
- Levitt P 1998 Local-level global religion: the case of United States–Dominican migration. *Journal of Scientific Religious Studies* 37: 74–89
- Luckmann T 1967 *The Invisible Religion: The Problem of Religion in Modern Society*. Macmillan, New York
- Marty M E, Appleby R S (eds.) 1994 *Accounting for Fundamentalisms: The Dynamic Character of Movements*. University of Chicago Press, Chicago
- Meyer J W, Boli J, Thomas G M, Ramirez F O 1997 World society and the nation-state. *American Journal of Sociology* 103(1): 144–81
- Robertson R 1991 The globalization paradigm: thinking globally. In: Bromley D (ed.) *Religion and the Social Order. New Developments in Theory and Research*. JAI Press, Greenwich, CT, Vol. 1
- Robertson R 1992 *Globalization: Social Theory and Global Culture*. Sage, London
- Rudolph S H, Piscatori J (eds.) 1997 *Transnational Religion and Fading States*. Westview Press, Boulder, CO
- Van Dijk R A 1997 From camp to encompassment: discourses on transsubjectivity in the ghanian pentecostal diaspora. *Journal of Religion in Africa* 27: 135–59

P. Beyer

Goddess Worship (Old and New): Cultural Concerns

Goddess worship is integral to many religions, old and new. The ancient Mediterranean world contained abundant Goddess lore; today, Hinduism and Buddhism are richest in female divinities. Academic studies of these traditions trace their texts, histories, rituals, iconographies, theologies, intersections with economics and politics, and—most recently—implications for female status. This article builds upon such scholarship, but it aims to describe the contemporary phenomenon of Goddess worship as a choice by women and men who believe that worshipping

God in a female form will empower them spiritually, socially, or politically. Who is turning to Goddess worship, and why? What does this say about history, religion, and society? Are Goddesses ‘good’ for women? Although the material investigated here derives largely from Western contexts, one can detect the same recourse to Goddess figures by people in other cultures who seek transformation through potent symbols and personalities. But the ‘new’ Goddess worship of today, embracing and refashioning the Goddesses of ‘old,’ is not everywhere the same. A Wiccan member of The Fellowship of Isis may differ radically from an Indian author involved with the feminist press, Kali for Women. Thus to nuance the picture of this emerging phenomenon, non-Western experiences of contemporary Goddess appropriation—particularly Indian—are brought in as test cases for claims made about Goddess worship here in the West.

1. Seeking New Symbols

In Western countries, Goddess worship is typically viewed as a religious alternative for those disillusioned with the overwhelming maleness of Jewish and Christian religious imagery and social structure. Symbols of God as father, conqueror, judge, and lord are believed to enforce the patriarchal domination of women. Moreover, Western philosophical dualisms—soul/body, sacred/profane, transcendent/immanent, rational/emotional, and changeless/changeable—justify the creation of hierarchy and demonize the lower halves, such that it becomes necessary to subdue sexuality, nature, emotion, and change. The new Goddess traditions also make historical charges—that Jews and Christians have wiped out indigenous Goddess-worshipping societies, Jews during the time of the prophets and Christians in the fourth to fifth centuries, when their emperors closed Goddess temples throughout the Roman empire. Indeed, as many critics claim, Jewish and Christian histories have systematically subordinated Goddesses and women, rather than according them equivalence in the theological and social spheres.

The most cogent justification for the need to reject the Judeo-Christian tradition for that of the Goddess draws upon theories about religious symbols. Following the work of Clifford Geertz, Carol Christ (1987, 1997) and other feminists have postulated that symbols have a direct relationship to sociocultural systems. Since symbols create powerful, pervasive, and long-lasting moods and motivations, shaping our understanding of the world and suggesting patterns of action, patriarchal symbols reinforce patriarchal social structures, which therefore appear uniquely realistic. Changes in the way women are perceived and treated can only come through the creation of new symbols, which empower women and undercut male

dominance. As Birnbaum states, quoting Gloria Anzaldúa, 'nothing happens in the "real" world unless it first happens in the images of our heads' (1993, p. 7).

2. *Varieties of Goddess Votaries*

Western Goddess spirituality is heterogenous and flexible, and sits at the intersection of four large groups (Eller 1993, Puttick 1997). The first, known loosely as spiritual feminism or women's spirituality, started in the early 1970s as a spontaneous grassroots movement. Cynthia Eller describes it as being founded on four premises: religion and ritual *can* be empowering for women; a healthy environment, peace, and community must be sought; Western history needs revising; and gender relations should be explored. Since many spiritual feminists stay within their birth traditions to work for justice and equality, not all women in this category worship Goddesses.

The second umbrella label is (neo-) paganism and Wicca, founded in the 1950s in Britain and based on the recovery of pre-Christian European religions such as Druidry and Odinism. In general, such groups believe in a dynamic balance between male and female; the centrality of fecundity; an immanent deity; care for the earth; the efficacy of ritual and magic; and the leadership of women. Goddesses are central to pagan religiosity, but not necessarily in isolation from male gods.

The concerns of spiritual feminists and Wiccans can be similar, as seen in the writings and practices of feminist witches Zsuzsanna Budapest and Starhawk. However, there are also frictions between the two groups. Since Wicca is not political, does not aim to strip away patriarchal domination, and advocates male/female complementarity, some spiritual feminists view Wiccans as sexist. Wiccans return the compliment; for them, politics and feminism are irrelevant to spiritual growth.

A third set of related groups is referred to collectively as the New Age. Famous for their crystals, chakras, channeling, past-life regressions, astrology, and mind-over-matter affirmations, New Age individuals create a pastiche of religious self-help devices, many of which involve Goddess figures. Spiritual feminists and pagans often critique the New Age for its tendency to whitewash and sweeten Goddesses; for its racism, misogyny, lack of commitment to community struggle, and exploitation of cultures from which it borrows; and for its naive optimism, unhelpful in times of crisis (Christ 1997, Sjöö 1992).

The final group—new since the 1980s—of contemporary Western Goddess worshippers are the eco-feminists, eco-spiritualists, and Gaians (from Gaia, or Earth). Though not all of them gynemorphize the earth as a Goddess (there are Jewish, Christian, and Buddhist eco-feminists as well), most worship the

earth as a mother Goddess and decry her desecration by modern industrial patriarchs. For Vandana Shiva, the Indian human rights activist, feminism is a form of ecology, and ecology is a revival of *prakṛti*, the demonized but potent Hindu feminine principle of nature (1988, pp. 38–54).

3. *Who is She?*

Like the variety of the Goddess movement, so also the Goddess herself. Her identity is multiple and fluid, changing depending upon the needs of her worshippers. Some women understand her as the affirmation of the female body, sexuality, and power, all of which have been downplayed in traditional Western religions. Although many feminists caution against too rigid a defense of female sexuality, as if biology were determinative of a women's 'essence' (di Leonardo 1998, pp. 98–108, Gross 1998, pp. 159–70), others argue that creating a space for women to reclaim the positive aspects of their sexuality is so important that the project is worth the risks.

Another method of affirming female strength is through a Jungian perspective, where the Great Goddess is an archetype of the feminine, dark side of life. Following Robert Graves, Erich Neumann, and Joseph Campbell, modern Jungians claim not only to empower women and men through the recognition of the positive aspects of innate femaleness, but also to explain, through the theory of archetypes, why women who have never encountered Goddesses in real life can suddenly start creating and dreaming about them. Not all feminists and Goddess worshippers approve of the use of Jungian archetypes, however. Some protest that in revalorizing the feminine unconscious, the banished 'other,' one is creating a new version of the old male/female dichotomy which subjugates women (Gross 1998, p. 190); others who worship a 'real' Goddess complain that Jungians relegate her to the 'unreal' realm of the mind (Christ 1998, pp. 86–8); and still others, drawing again upon Geertz's work on symbols, prefer to create new ones rather than to adopt those postulated by a patriarchal psychological system.

The Goddess also has material manifestations. As indicated above, some women and men view her as the earth. Since we are all part of nature, we must celebrate, and not attempt to control, its cycles. 'I feel deeply that the flight from finitude and death is at the root of the problems we face... Maybe something went wrong, massively wrong, when Platonism and Christianity became the dominant symbol systems of our culture. Maybe... the Goddess can help us... to love a life that ends in death' (Christ 1987, p. 210). Often linked with this conception of divinity is the assertion that the Goddess is a relational symbol of embodied, immanent love.

Although many Western Goddess votaries might assent to the interpretations given above, the majority



Figure 1
Artemis of Ephesus (reproduced by permission of the Mansell Collection/TimePix)



Figure 2
Cybele enthroned with her lions (reproduced by permission of the Mansell Collection/TimePix)

also worship the Goddess as a living presence. Some, choosing to stay within their Jewish and Christian contexts, derive inspiration from the feminine sides of God as revealed in the Bible or from their respective mystical traditions: Isaiah's 'womb language'; the emphasis on God's wisdom; the Kabbalists' Shekhinah, or presence of God; the Virgin Mary; feminine interpretations of the Holy Spirit; and Julian of Norwich's Mother Jesus.

Others, finding the traditions of their birth irredeemably patriarchal, go outside in search of Goddess figures whom they can approach in prayer and ritual. The most frequent destinations for such seekers are Western, often European, traditions, where, even if the Goddesses are no longer the center of living cults, they are believed to be waiting, still puissant, able to heal. Examples include the Greek Demeter and Persephone, Aphrodite, Artemis (see Fig. 1), and Athena; the Egyptian Isis; the Sumerian Inanna; the Canaanite Astarte; and the Anatolian Cybele (see Fig. 2). Sometimes, though more gingerly, for fear of causing offense, Western Goddess worshippers also turn to figures from Native American tribes, such as

Old Spider Woman, Corn Woman, and Serpent Woman, as well as to powerful ancestors and spirits from Africa.

Non-Western Goddesses also have devoted followings: the Japanese Amaterasu, the Hindu Durgā and Kālī, the Tibetan Buddhist Goddesses Tārā and Vajrayoginī, and the Chinese Buddhist Kuan-yin. Often these, and the Western deities mentioned above, are lumped together through superficial resemblances, archetypal links, and claims of historical connection, producing what Cynthia Eller calls the 'grab-bag approach to feminist spirituality' (1993, p. 75). Many criticize these dehistoricized approaches to Goddess worship as insulting, romantic, essentialist, and orientalist. The fact that some modern Goddesses have been created afresh, according to elite needs—for example, Asphalta, the parking Goddess (Grey 1989)—does not assuage such fears. Nevertheless, most modern Goddess worshippers are sincere seekers, turning to female embodiments of the divine in a genuine search for new life.

Before concluding this introduction to the meanings of the Western Goddess, it is worthwhile noting that

the reclamations under discussion appear not to be occurring among Muslims. There are three possible explanations: the minority status of Muslims in Western countries, which places them outside the liberal mainstream; the paucity of feminine images of the divine in the Qur'an; and the strict monotheism of Islam, which makes anathema the search for a Goddess figure.

4. *How Old is She?*

One of the most controversial elements of the Goddess spirituality movement is a claim about origins, or the 'Goddess Hypothesis.' Currently championed by authors like Christ, Eisler, Gadon, Gimbutas, Spretnak, and Stone, but relying on the earlier works of Bachofen, Engels, Graves, and Neumann, this hypothesis rests on a belief that 'there were past societies in Europe and the Near East, especially prior to the so-called invasions of the Indo-Europeans circa five thousand years ago, that were Goddess-worshipping, female-centered, in harmony with their environments, and more balanced in male-female relationships' (Conkey and Tringham 1995, p. 206). Drawing upon archaeological excavations of female figurines (fertility 'idols' of 'the Great Mother cult'), grave sites, and cave paintings, as well as anthropological research into modern-day 'primitive' societies who may be considered survivals from earlier eras, such authors see continuities from the time of the Upper Paleolithic period (32,000–10,000 BCE) and the Neolithic Revolution (9,000 BCE), down to the societies of Old Europe (6,500–3,500 BCE). What they have in common is a Great Mother cult and a social organization that is matrifocal (some claim matriarchal), earth-bound, and peace loving. Present-day clues of this earlier time, when the 'earth was believed to be the body of a woman and all creatures were equal' (Birnbaum 1993, p. 3), can be seen in the black Madonnas of Europe and South America, as well as in tribes that revere women for their reproductive powers. All such societies ended abruptly with the arrival of the patriarchal, war-oriented Indo-Europeans in the middle of the fourth millennium BCE; they subjugated human women and tamed, desexualized, or slew their Goddesses.

Although some women and men find such a vision of the past to be liberating for the future, others find much to criticize (Conkey and Tringham 1995, Eller 2000, Motz 1997, Puttick 1997). How do we know what the female figurines represent without an adequate knowledge of the social systems that produced them? Is it possible that they had the same symbolic meaning over nearly 30,000 years? Is it not feminist essentialism to argue that all the figurines connote fertility, as if women have always been defined biologically? Most historical Goddesses, indeed, are

not tied to fertility at all. And if it is allowable to read back to Old Europe by citing modern-day 'survivals,' what about contemporary groups in which there is matrilineality but male control, ritualized fertility but no Goddess worship, and Goddess veneration but little respect for women? The brunt of such criticisms is that the past is oppressive, not golden; it ill serves the women's movement to idealize something unproven.

5. *Is She 'Good' for Women?*

Definitions of women and Goddesses that rely too much on biological determinism, Jungian archetypes, and the 'Goddess Hypothesis' are not the only contested issues in the Goddess movement. Another is the relationship between Goddess-worshipping societies and the status of their women. Do Goddesses serve patriarchal interests, or feminist ones? The evidence is mixed; although some Goddesses appear to serve their human devotees by undercutting patriarchal subordination, and although some women find strength in images of the divine feminine, other Goddesses reinforce their societies' domination of women, such that feminists refuse to call upon them.

First, some positive correlations. Many modern Western women do indeed derive inspiration from their worship of Goddess figures: Luisah Teish, who looks to West African ancestor veneration; Charlene Spretnak, Christine Downing, and Carol Christ, who have found spiritual homes with the Goddesses of the Mediterranean; China Galland, whose life was changed by encountering Tārā; and Rita Gross, who testifies about the inner transformation she experiences through the worship of Vajrayoginī. Very often, one woman's metamorphosis leads to another's, as occurs through the women's pilgrimages to Crete led by Carol Christ.

It is not only Western women who are thus empowered. To take a few examples from South Asia, the audacity of the Goddess Rādhā vis-à-vis her lover Kṛṣṇa provides Bengali women with models of assertiveness and resistance (Wulff 1997); forms of Durgā empower Hindu devotees and holy women by their autonomy and puissance (Erndl 1997); women Tantrikas of the Tibetan Buddhist tradition find their spiritual aspirations mirrored in powerful images of the feminine divinity (Shaw 1994); and the activist Mother's Front in Sri Lanka worships Kālī for fortitude in the midst of the civil war (de Alwis 1999). Sometimes the Goddess herself overturns patriarchal values; the South Indian Goddess Kāṛumāriamman allows menstruating women into her temple (Narayanan 1999, p. 69).

But there are also contrary indications. All 10 Goddesses whom David Kinsley (1989) describes in *The Goddesses' Mirror* belong to societies in which men dominate women; the same is true for those

catalogued by Motz (1997). Other scholars find similarly depressing inverse correlations between Goddess worship and freedom for women. Again, from South Asia: William Sax investigates the north Indian devotion to Nandādevī and finds that while the Goddess represents the trials of out-married daughters, her cult enables rather than undercuts their sorrows (1991); China Galland visits Bhaktapur, Nepal, and describes both her delight in the beauty of Durgā worship and her horror at the thriving trade in child prostitution which exists alongside it (1998); and Lisa Hallstrom discovers that although devotees view the modern saint Ānandamayī Mā as an incarnation of the Goddess, she is not considered a 'woman,' and feminist interpretations of her significance are discouraged (1999). The frustration of such a disjunction between expectations and reality is summed up by an eight-year-old girl in a story by Mrinal Pande: 'when you people don't love girls, why do pretend to worship them?' (1990, p. 63).

Many women's activists in countries such as India, therefore, eschew Goddesses as aids in their struggles. For Goddesses can validate a sexist model, as in the case of Sītā, the virtuous wife of Rāma, or represent a romanticization of the past and a glorification of motherhood. Indian feminists, after initially turning to Kālī and Durgā as symbols of empowerment, found them usurped by Hindu nationalists and deployed for divisive ends (Agnes 1995). One example is Durgā Vāhinī, or Durgā's Army, a Hindu nationalist women's group. The Vāhinī has revived female militancy—not in the service of women's issues (caste, divorce, dowry, sati, or marital violence), which are not debated or fought for, but against Muslims (Pathak and Sengupta 1995). The eight-armed Durgā gives women a new identity, but an identity that serves patriarchal values.

6. Trends, Evaluations

Goddess worship, wherever it is found, is diverse and controversial. Particularly since 1985 in the West, it is also a place where the popular and the scholarly meet and spar. Sometimes academic authors bemoan the theoretical backwardness that characterizes the public enthusiasm for Goddess spirituality and matriarchal histories, complaining that exhorting women to realize that 'we're all Goddesses within!' vitiates the need for external change (di Leonardo 1998, pp. 104–7). But sometimes the meetings between popular and scholarly are not so antagonistic: Goddess-revering people of all stripes can and do join hands when it comes to issues of social justice and eco-spiritualism.

The revived Goddess spirituality of the West also provides opportunities for encounters of other types. The increasing ethnic mix of Western societies means that 'new' Goddess worshippers come into contact with

those for whom Goddesses are a traditional part of their religious backgrounds. Sometimes Western votaries find this an inspiring, deepening encounter, and sometimes they do not. How, for instance, is one to incorporate the traditional practice of goat sacrifice into one's Western worship of Kālī? By the same token, many first- and second-generation immigrants find the appropriation of their Goddesses by mainly white Americans to be offensive. The Goddess movement is obviously a fertile field for intercultural exchange.

How then to evaluate this refashioning of 'old' Goddesses in the service of 'new' ideals and hopes? While from an academic perspective one might deplore the 'Goddess Hypothesis,' biological determinism, the perceived exploitation of 'other people's symbols,' and a naive assumption that all women everywhere benefit from the Goddess, the debate, creativity, and diversity characteristic of 'new' Goddess worship in the West are evidence of laudable cultural vibrancy. Goddesses may not always, or even often, provide alternative models for women in today's global societies. But as various studies have shown, in religious groups where women do have a leadership role, there is typically a feminine aspect to God (Sered 1994). Rita Gross makes the point differently when she asserts that even if Goddesses do not confer status or independence to women, they do contribute to psychological well-being. 'Though the presence of goddesses does not guarantee equity for women, their absence almost certainly guarantees inequity for women' (1998, p. 161). A final example from India: just because the slain Mrs. Gandhi was depicted in the press as a mother Goddess killed by her ungrateful Sikh children, who therefore deserved the violence unleashed upon them, there is no reason that Goddess symbolism, wrested from nationalist agendas and purposely reworked, could not be used for healing and hope. To echo Christ, echoing Geertz: people change as symbols change, and vice versa. That the denizens of the Goddess movement are trying to push this mutual process forward to their own ends is neither surprising nor unprecedented, though where the Goddesses 'take hold' and in what forms are matters for future interest.

See also: Buddhism and Gender; Catholicism and Gender; Feminist Theology

Bibliography

- Agnes F 1995 Redefining the agenda of the women's movement within a secular framework. In: Sarkar T, Butalia U (eds.) *Women and Right-Wing Movements: Indian Experiences*. Zed Books, London
- Birnbaum L 1993 *Black Madonnas: Feminism, Religion, and Politics in Italy*. Northeastern University Press, Boston

- Christ C 1987 *Laughter of Aphrodite: Reflections on a Journey to the Goddess*. Harper & Row, San Francisco
- Christ C 1997 *Rebirth of the Goddess: Finding Meaning in Feminist Spirituality*. Routledge, New York
- Conkey M, Tringham R 1995 Archaeology and the Goddess: Exploring the contours of feminist archaeology. In: Stanton D, Stewart A (eds.) *Feminisms in the Academy*. University of Michigan Press, Ann Arbor, MI
- de Alwis M 1999 Motherhood as a space of protest: Women's political participation in contemporary Sri Lanka. In: Jeffrey P, Basu A (eds.) *Resisting the Sacred and the Secular: Women's Activism and Politicized Religion in South Asia*. Kali for Women, New Delhi, India
- di Leonardo M 1998 The Three Bears, The Great Goddess, and the American temperament: Anthropology without anthropologists. In: di Leonardo M (ed.) *Exotics at Home: Anthropologies, Others, American Modernity*. University of Chicago Press, Chicago, pp. 79–144
- Eller C 1993 *Living in the Lap of the Goddess: The Feminist Spirituality Movement in America*. Crossroad, New York
- Eller C 2000 *The Myth of Matriarchal Prehistory: Why an Invented Past Won't Give Women a Future*. Beacon Press, Boston
- Erndl K 1997 The Goddess and women's power: A Hindu case study. In: King K (ed.) *Women and Goddess Traditions: In Antiquity and Today*. Fortress Press, Minneapolis, MN
- Galland C 1998 *The Bond Between Women: A Journey to Fierce Compassion*. Riverhead Books, New York
- Grey M 1988 *Found Goddesses: Asphalta to Viscera*. New Victoria Publishers, Norwich, VT
- Gross R 1998 *Soaring and Settling: Buddhist Perspectives on Contemporary Social and Religious Issues*. Continuum, New York
- Hallstrom L 1999 *Mother of Bliss: Ānandamayī Mā (1896–1982)*. Oxford University Press, New York
- Kinsley D 1989 *The Goddesses' Mirror: Visions of the Divine Feminine from East and West*. State University of New York Press, Albany, NY
- Motz L 1997 *The Faces of the Goddess*. Oxford University Press, New York
- Narayanan V 1999 Brimming with Bhakti, embodiments of Shakti: Women of power in the Hindu tradition. In: Sharma A, Young K (eds.) *Feminism and World Religions*. State University of New York Press, Albany, NY
- Pande M 1990 Girls. In: Holmström L (ed.) *The Inner Courtyard: Stories by Indian women*. Virago Press, London
- Pathak Z, Sengupta S 1995 Resisting women. In: Sarkar T, Butalia U (eds.) *Women and Right-Wing Movements: Indian Experiences*. Zed Books, London
- Puttick E 1997 Goddess spirituality: The feminist alternative? In: Puttick E *Women in New Religions: In Search of Community, Sexuality, and Spiritual Power*. St. Martin's Press, New York, pp. 196–231
- Sax W 1991 *Mountain Goddess: Gender and Politics in a Himalayan Pilgrimage*. Oxford University Press, New York
- Sered S 1994 *Priestess, Mother, Sacred Sister: Religions Dominated by Women*. Oxford University Press, New York
- Shaw M 1994 *Passionate Enlightenment: Women in Tantric Buddhism*. Princeton University Press, Princeton, NJ
- Shiva V 1988 *Staying Alive: Women, Ecology, and Survival in India*. Kali for Women, New Delhi, India
- Sjöö M 1992 *New Age and Armageddon: The Goddess or the Gurus? Towards a Feminist Vision of the Future*. The Women's Press, London

- Wulff D 1997 Rādhā's audacity in Kīrtan performances and women's status in Greater Bengal. In: King K (ed.) *Women and Goddess Traditions: In Antiquity and Today*. Fortress Press, Minneapolis, MN

R. F. McDermott

Goffman, Erving (1921–82)

The work of the sociologist Erving Goffman is regarded as a classic by most of his professional colleagues. Randall Collins (1988, p. 41) even considers him 'the greatest sociologist of the latter half of the twentieth century.' Goffman's influence, particularly on many younger researchers, is still powerful at the beginning of the twenty-first century and is by no means restricted to sociology. Goffman's works have been influential in numerous parallel disciplines, above all in social psychology, linguistics, philosophy, literature studies, pedagogics, psychiatry, and media studies. It cannot be denied that Goffman made important empirical, theoretical, and methodological contributions. The following essay will discuss them against the background of a short biographical sketch and relate them to relevant sociological contexts of discussions.

1. Biographical Sketch

Erving Goffman was born in the small Canadian town of Manville on July 11, 1922, the son of Russian-Jewish immigrants. He began his academic career at the University of Toronto, where he took his BA in 1945. After completing his sociology studies at the University of Chicago, USA, where he primarily studied under Everett C. Hughes and W. Lloyd Warner, he took up his first post in Scotland, UK. From 1949 to 1951 he worked at the Department of Social Anthropology at Edinburgh University. After that he returned to Chicago, where he gained his doctorate with his thesis *Communication Conduct in an Island Community*, which he had prepared in Scotland. The empirical basis of his dissertation, which was supervised by W. Lloyd Warner, Donald Horton and Anselm L. Strauss, was a field study on the Shetland Islands. Following a spell as research assistant at the Department of Sociology at Chicago University, where his research included social stratification, Goffman was employed as a Visiting Scientist in the Laboratory of Socio-Environmental Studies at the National Institute of Mental Health (NIMH) in Bethesda (Maryland) from mid-1954 to the end of 1957. During this time Goffman carried out a one-year study in a mental hospital. This study on the basis of 'participant observation' formed the essential foundation of his book *Asylums* (1961a). In 1958 Goffman transferred to Berkeley as Assistant Professor of

Sociology and member of the Center for the Integration of Social Science Theory at the University of California. From 1962 he was a Full Professor in Berkeley. In 1968 he accepted the Benjamin Franklin Professorship at the University of Pennsylvania in Philadelphia, where he substantially advanced his research program and wrote his *chef d'oeuvre* 'Frame Analysis' (1974).

His works, which reached a wide academic and non-academic audience and have been translated into many languages, earned Goffman numerous honors. He received a number of prestigious awards and was Visiting Professor at several universities at home and abroad. In 1981 Goffman became President of the American Sociological Association (ASA). At the peak of his powers after an academic career of almost 30 years, Goffman died on November 20, 1982.

2. Goffman's Work

2.1 Research Program and Intellectual Context

Goffman's main theme was the forms of immediate social interaction, 'face-to-face' encounters. He saw in them a specific type of social system and a sociological field in its own right. In his Presidential Address delivered at the end of his term as President of the American Sociological Association shortly before his death, he declared: 'My concern over the years has been to promote acceptance of this face-to-face domain as an analytically viable one—a domain which might be titled for want of any happy name, the interaction order—a domain whose preferred method of study is microanalysis' (Goffman 1983, p. 2). This programmatic description of his position does not mean that Goffman regarded the interaction level as completely autonomous or socially fundamental. Rather he stressed the manifold 'relativity' of the autonomy of the interaction order. In many places in his works Goffman pointed out its historical variability as well as links ('a loose coupling') between the micro-level of the interaction and social macro-structures (e.g., social stratification or bureaucratic organization). However, Goffman did not systematically consider the historical and social conditions and dependencies of his subject. Rather he was primarily concerned with observing and discovering seemingly insignificant aspects of daily behavior. He focused on tiny moments of people's behavior, which they themselves find more or less natural and are unaware of: fleeting glances, physical contact and distance, apologies, greetings, compliments, silence, etc.

All Goffman's works, 11 books in total, are consistently oriented towards this empirical reality, but from the start Goffman also adopted means of interpretation from various sociological schools and traditions, including above all the 'classics' William James (see *James, William (1842–1910)*), Emile Durkheim (see *Durkheim, Emile (1858–1917)*) and

Georg Simmel (see *Simmel, Georg (1858–1918)*), structural functionalism (Alfred Radcliffe-Brown, Talcott Parsons (see *Mead, George Herbert (1863–1931)*)), representatives of the Chicago School (George H. Mead (see *Mead, George Herbert (1863–1931)*), Charles H. Cooley, Louis Wirth et al.), existentialism (Jean-Paul Sartre), ethology, ethnomethodology (Harold Garfinkel, Harvey Sacks), and sociolinguistics (Dell Hymes). But Goffman never regarded his theoretical references, models and metaphors as ends in themselves, they were rather means to the end of discovering and describing social order.

2.2 Approaches, Positions, and Results

In the following attempt to briefly summarize and appraise Goffman's work, a number of key sociological terms either coined or redefined by Goffman will be used as a guideline.

2.2.1 Dramaturgy. Goffman became known for his analysis of social life based on the model of the theater (see 1959), and his model has remained influential. It is too simple to speak of Goffman's work, as for example Jonathan Turner does, as the 'dramaturgical school of interactionism,' but it can be demonstrated that the dramaturgical perspective consistently played an important role for Goffman. It can similarly be proved that Goffman's dramaturgical approach partly anticipated and formed a basis for many of his later constructs. This particularly applies to his works on strategic interaction, ritual, territoriality, identity, and stigma.

Within the context of the theater model Goffman reconstructed above all the strategies of social 'face work': forms of concealment, lying, self-stylizing, mystification, the manipulation of 'facades,' etc. His aim here was not just to show how individuals, selfish subjects in the fight for social recognition, attempt to present themselves in the best light; rather he constantly also saw dramaturgical everyday life as producing and reproducing social order, a restrictedly open and widely endangered process that is to be permanently established.

Goffman's dramaturgical approach and his later reflections on social territories (1971) and frames (1974) also help us understand the spatial organization (social ecology) of modern society. For example, it becomes clear that normality is a fictional impression to be produced by differentiating distinct spheres of perception, audiences, and representations; it is the foundation for the actors' basic social trust. Goffman's differentiation between 'frontstage' and 'backstage' (see 1959) became famous in this context.

2.2.2 Roles and identities. Goffman's early works in particular are to be seen in the context of role theory,

influential at the time, which he criticized and developed further. Goffman treated role analysis primarily as interaction analysis, investigating the functioning and organization of the ‘actual’ practice of performing a role against the background of its normative frame. Thus Goffman developed concepts of patterns and styles of behavior, including the term ‘role distance’ (see 1961b), which has meanwhile acquired the status of a basic sociological term and which refers to a way of behavior which comments on the role and primarily serves either the interaction system or the selves relevant in the situation. For example, Goffman described a five-year-old boy riding a roundabout horse who by little irreverences in his behavior demonstrates that his current role does not correspond to his ‘true,’ ‘more adult’ self.

In *Stigma* (1963), Goffman differentiated the term ‘self’ in the context of an identity theory which distinguishes three types of identity in a basic sense that is still valid at the beginning of the twenty-first century: (a) social identity as a person’s role set, (b) personal identity as a person’s synchronic and diachronic individuality ascribed to him by observers, and (c) ego identity as a person’s inner self-reference. Goffman used these concepts of identity together with the instruments of his dramaturgical approach above all in analyzing stigmatized deviations and deviants. He focused on the techniques of information control and emotion management of the stigmatized people as well as on the consequences of the stigmatization for their identities and socialization.

2.2.3 Total institutions. Goffman’s book *Asylums* (1961a) represents an exception among his works in that it systematically goes beyond the interaction level. It describes a class of social organizations that has to do with confining, controlling and ‘altering’ people, such as mental hospitals, prisons, convents, homes, barracks, concentration camps, etc. However, Goffman’s (1961a) investigated such ‘total institutions’, where the most elementary matters of course and norms of everyday interaction are broken, with the aim of exposing the interaction order and the self connected with it. Beyond that, Goffman’s analysis demonstrated that total institutions form an ambivalent milieu for socialization. On the one hand they act as machinery for disciplining and normalizing behavior and interaction. On the other hand these institutions damage the moral self, e.g., self-respect and modesty, through various forms of humiliation.

2.2.4 Ritualizations. In his work Goffman repeatedly dealt with the moral sides of the interaction order. For Goffman the self is a moral fact which implies rights and duties when showing respect and which can be observed in ritual practices and codes of every-

day interaction. Every person appears as both subject and object of a ritual ‘language,’ a ‘language’ of politeness, decency, and tact, but also on occasion of contempt, wounding, humiliation, etc. (see Goffman 1967).

At the end of his career Goffman elaborated his early ritual model to include ethological and frame theoretical considerations, on the basis of which he analyzed gender interactions and the interpretative patterns they are based on (see 1979). Goffman depicted gender as a combination of ritual styles of presentation representing beliefs about gender-specific essential attributes. These beliefs amount to a schematic asymmetry similar to that of the parent-child relationship. In daily interaction life and its media depictions (e.g., in advertising) men appear as superior beings to women by ritualizations of dress, physical contact, smiling, activity, etc.

2.2.5 Strategic interactions. For Goffman, ritualizations represented the most important social condition of profit-oriented strategic action on the level of interaction. Totally ignoring their contents, he analyzed this action within the framework of a model of strategic interaction oriented towards game theory (see 1969). The main point of departure and aim of Goffman’s respective considerations were the social situation in which the subjects strive for control of the ‘definitions’ of the other players or rivals. Following on from his dramaturgical approach, Goffman above all investigated the conditions of competence and limits of ‘impression management,’ the strategically relevant interpretations of expression (e.g., of the body), the logic of reciprocal strategic observation and assessment, as well as the complexity of meaning and virtuosity of strategic moves. The empirical basis of the investigation consisted in particular in self-descriptions by a wide variety of strategic subjects (spies, confidence tricksters, adulterers, politicians, etc.). Goffman ‘translated’ their practical knowledge of action (knowledge of cleverness) into the categories of his analytic description, e.g., a typology of strategic moves.

2.2.6 Frames and framing subjects. *Frame Analysis* (1974) can be regarded as Goffman’s main theoretical work, because it integrates, or could integrate, the many various analytic perspectives and terminological distinctions of his earlier works. *Frame Analysis* itself therefore forms something like a frame. It is characterized by two basic orientations. On the one hand, the interest is directed at ‘structures’ and procedures for establishing meaning which are effective in generating normality in social situations and which are valid, independent of these situations. On this level, which can be compared with the grammar

of a language, Goffman revealed basic types of meaning and complex possibilities of transforming meaning (play, experiment, demonstration, etc.). On the other hand, on the level of the concrete situative events he aimed both at normal uncertainties, dangers, and the pressure to conform, as well as at extreme forms of abnormality, such as the 'negative experiences' of inmates of mental hospitals, whose habitual framings, which have proved their worth in everyday life, fail in the institution. The individual, whom Goffman revealed as the subject of the framing, is characterized by enormous powers of judgment and virtuosity, particularly in functionally complementing rules and also in manipulating them.

2.2.7 Forms of talk. In his last period, Goffman built on earlier studies and particularly concerned himself with conversations. The main result of this work is his book *Forms of Talk* (1981), which Goffman specifically placed in the context of *Frame Analysis*, one chapter of which was already devoted to conversations.

Goffman's late focusing on conversations can be seen in connection with the emergence of conversation analysis as a branch of ethnomethodology, which he approved of, but also criticized (see Bergmann 1991). In contrast to conversation analysis, which tends to conceptualize the process of communication as an autonomously organized system, Goffman stressed the importance of ritual order, which is always culture-specific and which permits selves and consequently psychological states to develop. Goffman basically rejected the assumption by conversation analysts that sections of a conversation can be analyzed without reference to their situational and extra-situational contexts. Moreover, Goffman was mainly interested in a sociological approach to the contextuality of social processes. *Frame Analysis* represents this approach.

2.3 Goffman's Methods

Goffman's work is essentially the result of the author's talents, in particular his personal sensitivity, the subtlety of his powers of observation, and his ability to classify his material in his search for order. He displayed these talents in the course of practical research, which struck most observers as unconventional, even puzzling. Among the most irritating points are Goffman's data, whose heterogeneity can hardly be exceeded. In his works the reader comes across personal everyday experiences, novels, autobiographies, the daily press, essays by film critics, photographs, etc. In many other respects Goffman also deviated from the normal forms of sociological discourse. He abbreviated the traditional terminological vocabulary of his field, avoided lengthy argumentation, frequently changed perspective, focused on seemingly unimportant

details, commented in an autoreflective manner on his own work, made ironic remarks and repeatedly told or referred to 'stories.' What is most important is that Goffman did not conform to any methodological canon, rather his methods entirely depended on the subject. This meant first of all that it was necessary to forgo certain traditional research methods. For example, the fact that those patterns of behavior that interested Goffman were generally unconscious ones meant that interviews were largely or totally impossible.

Goffman himself described his procedure as primarily 'naturalistic.' The scope of what he meant by that ranged from 'participant observation' (even in his own everyday life) to work with audio-visual recordings. His aim was always to get as close as possible to 'social nature' and to capture it as accurately as possible. On the other hand Goffman tried to achieve analytic distance and obtain information by various forms of 'alienation.' Alongside systematic work with models such as that of the theater, his most important technique was the strategy of alienation, which Paul Drew and Anthony Wootton (1988, p. 7) called 'the investigation of the normal through the abnormal.' Goffman employed totally different forms of abnormality, such as stigmatizations, espionage, accidents, or celebrities as approaches to the logic of everyday normality.

In order to convert into texts the results that were generated by his methodological dialectic of approaching and distancing Goffman combined two strategies. On the one hand he developed a formal analytical language which contains the basic features of the interaction order and the self connected with it. On the other hand he employed a strategy of 'thick description' (C. Geertz). By using a literary style, a metaphorical language suited to the specific subject, and extensive quotations from his material, cleverly combined with his analytical observations, he succeeded in making his subject appear alive and viewing it in a new light.

3. Goffman's Importance for the Social Sciences Today

It is not easy to judge how relevant Goffman's work is for the social sciences at the beginning of the twenty-first century; its main significance probably lies in his contribution to discovering the interaction order. In this way Goffman opened up a specific research area and at the same time provided central categories and strategies for exploring it.

But quite apart from interaction sociology, Goffman also had a strong influence on a number of other research areas and he remains influential today. This is especially true of many aspects of research into identity and socialization. Of prime importance among them are Goffman's reflections on institutional control and

discipline procedures, and on subjectivity. Goffman is especially appreciated today in research on normality and deviation as well as in gender studies, where some of his works are still standards now. In addition, he has a significant influence not only on the level of the everyday interaction order, but also on the sociology of certain social fields of action (politics, advertising, science, etc.) and in media research, where many of Goffman's analytical instruments have proved to be valid. Furthermore it can be seen that recently Goffman's works have increasingly been read and discussed from the point of view of general theories of civilization and society. For example, parallels have been drawn and connections made with the works of Norbert Elias (see *Elias, Norbert (1897–1990)*), Michel Foucault (see *Foucault, Michel (1926–84)*), and Pierre Bourdieu (see Willems 1997). In general it is a remarkable fact that widely differing and contrasting schools have incorporated parts of Goffman's works. The spectrum of positions Goffman is 'relevant' for, ranges from modern systems theory through critical theory, right down to the rational-choice approach.

Just as the reactions to Goffman's work have been wide-ranging and many-sided, so evaluations of its scope and its identity have differed greatly. While most people see the importance of Goffman's work essentially on the level of empirical micro-observation, others (e.g., Anthony Giddens) believe that in it there is to be found a systematic theory that goes beyond the level of micro-sociology and is highly significant for the development of the social sciences. Similarly, various labels have been attached to Goffman's thinking (symbolic interactionism, phenomenology, role theory, structuralism, ethnomethodology, etc.).

No matter how the situation of his reception is to be judged, it cannot be denied that Goffman left behind a series of studies that are still relevant and influential today and a network of concepts that is both highly complex and highly integrated. Perhaps an equally important part of his legacy is the paradigm he provided by the style and ethos of his research.

Bibliography

- Bergmann J R 1991 Goffmans Soziologie des Gesprächs und seine ambivalente Beziehung zur Konversationsanalyse. In: Hettlage R, Lenz K (eds.) *Erving Goffman—ein soziologischer Klassiker der zweiten Generation*. UTB, Bern, Stuttgart, Germany, pp. 301–26
- Collins R 1988 Theoretical continuities in Goffman's work. In: Drew P, Wootton A (eds.) *Erving Goffman. Exploring the Interaction Order*. Polity Press, Cambridge, UK, pp. 41–63
- Drew P, Wootton A 1988 *Erving Goffman. Exploring the Interaction Order*. Polity Press, Cambridge, UK, pp. 1–13
- Goffman E 1959 *The Presentation of Self in Everyday Life*. Doubleday Anchor, New York
- Goffman E 1961a *Asylums. Essays on the Social Situation of Mental Patients and Other Inmates*. Doubleday Anchor, New York

- Goffman E 1961b *Encounters: Two Studies in the Sociology of Interaction*. Bobbs-Merrill, Indianapolis, IN
- Goffman E 1963 *Stigma. Notes on the Management of Spoiled Identity*. Prentice Hall, Englewood Cliffs, NJ
- Goffman E 1967 *Interaction Ritual: Essays on the face-to-face Behavior*. Doubleday Anchor, New York
- Goffman E 1969 *Strategic Interaction*. University of Pennsylvania Press, Philadelphia, PA
- Goffman E 1971 *Relations in Public. Microstudies of the Public Order*. Harper and Row, New York
- Goffman E 1974 *Frame Analysis. An Essay on the Organization of Experience*. Harper and Row, New York
- Goffman E 1979 *Gender Advertisements*. Macmillan, London
- Goffman E 1981 *Forms of Talk*. Basil Blackwell, Oxford, UK
- Goffman E 1983 The interaction order. *American Sociological Review* 48: 1–17
- Manning Ph 1992 *Erving Goffman and Modern Sociology*. Stanford University Press, Stanford, CA
- Willems H 1997 *Rahmen und Habitus. Zum theoretischen und methodischen Ansatz Erving Goffmans: Vergleiche, Anschlüsse und Anwendungen*. Suhrkamp, Frankfurt a. M., Germany

H. Willems

Goodness of Fit: Overview

Many statistical methods make assumptions about, or have consequences for, particular distributions of certain quantities. Goodness-of-fit methods assess the extent to which such assumptions or consequences are supported by the data. Similarly, in the process of model search or model selection within a particular choice of statistical distribution or model form, goodness-of-fit approaches are used to guide the process of arriving at a sensible parsimonious model that captures the main features of the data.

This article describes the general philosophy of assessing the goodness of fit of models and distributions in the context of statistical analyses. The following sections address, in turn, the distinction between significance tests and goodness-of-fit tests, the role of alternative test criteria, and distributional plots for assessing goodness of fit, the role of frequentist goodness-of-fit criteria in model selection, as well as Bayesian approaches. Because there is an inevitable trade-off between the desire for model parsimony and models that fit well, it seems wise to give up the notion that a model might be 'true,' and instead think about choosing a model because its usefulness is reasonable.

For a detailed consideration of some specific tests and diagnostic measures in the context of assessing the adequacy of distributional assumptions see *Goodness-of-fit Tests and Diagnostics*. For detailed uses of the different approaches to goodness of fit see *Multivariate Analysis: Discrete Variables (Loglinear Models)*; *Linear Hypothesis: Regression (Basics)*; *Linear Hypothesis: Regression (Graphics)*.

1. Goodness of Fit for Categorical Data

Karl Pearson (1900) originally proposed his chi-squared test as a way to test the fit of probability model involving k parameters, by comparing a set of observations $O_1, O_2, \dots, O_t$, in t non-overlapping classes with their expectations under the model, $E_1, E_2, \dots, E_t$:

$$X^2 = \sum_{i=1}^t \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

In most practical situations, the values of $E_1, E_2, \dots, E_t$ involve estimates of the k parameters of the model. Then for large samples, the distribution of X^2 is approximated by the χ^2 distribution with $t - k - 1$ degrees of freedom when the probability model is in fact correct (see the discussion in *Goodness-of-fit Tests and Diagnostics* and in Bishop et al. 1975, for details on conditions associated with the estimates used, etc.). For categorical data problems such as those arising in the analysis of contingency table data (see *Multivariate Analysis: Discrete Variables (Overview)*), a widely used alternative to X^2 is the log-likelihood ratio test which is also referred to as the ‘deviance.’ Both of these are members of the power-divergence family of goodness-of-fit statistics proposed by Read and Cressie (1988):

$$K_\lambda(\mathbf{O}, \mathbf{E}) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^t O_i \left[\left(\frac{O_i}{E_i} \right)^\lambda - 1 \right] \quad (2)$$

where λ is a real-valued parameter in the interval $-\infty < \lambda < \infty$. The statistic X^2 corresponds to $\lambda = 1$, and the log-likelihood ratio statistic corresponds to the limit as $\lambda \rightarrow 0$. All of the members of this power-divergence family of statistics are asymptotically equivalent. Read and Cressie (1988) describe some rationales for the choice of value of λ .

To test the fit of a specific model, e.g., the model of independence in a two-way contingency table, the chi-squared test uses an omnibus alternative, e.g., dependence. The p -value of the test, is the probability of observing a discrepancy as large or larger than that observed under the hypothesized null distribution, where ‘discrepancy’ is measured by the particular distance metric $K_\lambda(\mathbf{O}, \mathbf{E})$ of equation (2). But in many circumstances testing the goodness of fit of a model is not viewed as being the same as doing a test of hypothesis or test of significance. Not rejecting the null hypothesis at some arbitrary p -value is often taken as evidence that the model provides an acceptable fit to the data (cf. the discussion in Bishop et al. 1975).

Goodness-of-fit test statistics are, of course, sensitive to the sample size n , and to the real class of alternative models of interest. Suppose the null model is not exactly true (as is nearly always the case). Then,

with a large enough value of n , one will almost surely conclude that the null model does not fit the data. For modest values of n the situation is more complex, and will result in small p -values for some alternatives, while for others the model may appear to fit the data reasonably well. And if the data are sparse, i.e., if the sample size, n is small relative to the number of cells t , or the number of parameters being estimated k , then the p -value computed by reference to the corresponding χ^2 distribution will inevitably suggest that the model fits even when it does not (cf., Koehler and Larntz 1980). Thus, the p -value of a chi-square test alone gives limited guidance on whether the hypothesized distribution is close enough to being correct that the conclusions of the analysis can be relied upon. When goodness-of-fit tests are used for model comparison and model search some of these problems can be attenuated, as is described in Sect. 3.

2. Other Goodness-of-fit Approaches Focused on Empirical Distributions

Other classes of goodness-of-fit tests are of special interest when one is focused on the appropriateness of a specific parametric model for some underlying cumulative distribution function, $F(x)$. What is often done in such circumstances is to compute the empirical distribution function $F_n(x)$, whose value at x is the fraction of the sample with values less than or equal to x , i.e., $\{\#x_i \leq x\}$, and then compare $F_n(x)$ with $F(x)$. Well known examples of such tests include the Kolmogorov–Smirnov test, the Cramer–von Mises test, the Anderson–Darling, and the Neyman smooth test. These, and related diagnostics, are described in *Goodness-of-fit Tests and Diagnostics*.

One of the most widely used diagnostic methods for checking the fit of a model is the quantile–quantile or ($Q-Q$) plot, which plots the quantiles of the empirical distribution $F_n(x)$, against the quantiles of the assumed distribution $F(x)$. A straight line through the origin at a 45 degree slope indicates a perfect fit. The advantage of this method is that it can indicate where in the distribution the assumption is weakest. $Q-Q$ and other probability plots are described in Wilk and Gnanadesikan (1968) and are used in statistical practice. The interpretation of deviations from the line in such plots is, however, often subjective.

3. Classical Approaches to Model Comparison and Model Search

3.1 Model Comparison

Model comparison and model search are ubiquitous activities in modern statistical practice. When analysts

consider developing a regression model to describe a set of data, for example, they typically have a large number of possible variables available for inclusion in the model. What is well-known in the regression context is that the use of additional regression variables typically improves the ‘predictive fit’ of the model to the data (e.g., see Copas 1983) but including as many variables as possible is a poor strategy which does not hold up to scrutiny under cross-validation or when predictions are made for new observations (e.g., see the discussion in Copas, 1983, and Linhart and Zucchini 1986). Thus, at a minimum, statistical analysts engage in a series of model comparisons or tests in order to avoid the inclusion of ‘non-significant’ variables or terms in a model. This follows the scientific dictum known as ‘Occam’s razor,’ attributed to William of Occam (but see Thorburn 1918), which suggests that the simpler of two models should be the one chosen. For example, *t*-tests and *F*-tests are widely used in model selection procedures for regression and analysis of variance models (see *Linear Hypothesis*). Similarly, there are various chi-square test-based techniques for selecting among the class of loglinear models applicable for contingency table structures, and these typically (although not always) resemble corresponding model selection procedures for analysis of variance and regression models (see the discussions in Bishop et al. 1975, and Fienberg 1980).

3.2 Model Search

Several methods of penalizing large models to avoid overfitting have been proposed, many taking the form of choosing a model to minimize

$$-2 \log l + f(p_1, p_2, n) \quad (3)$$

where l is the likelihood ratio, f is the penalty assessed for choosing a model with p_1 parameters instead of a more parsimonious one with p_2 parameters, and n is the sample size. When the models are nested, i.e., one is a special case of the other, the quantity $-2 \log l$ is the likelihood ratio test for comparing them. Under the null hypothesis that the simpler model fits, this test statistic follows a χ^2 distribution when the sample size n is sufficiently large.

The choice $f(p_1, p_2, n) = 2(p_1 - p_2)$ corresponds to Akaike’s Information Criterion (Akaike 1973) or AIC, while $f(p_1, p_2, n) = (p_1 - p_2) \log n$ corresponds to Schwarz’s (1978), also referred to as the Bayesian Information Criterion or BIC. Other classical approaches to this problem include Mallows’s (1973) C_p statistic which is similar to AIC, and Allen’s (1974) PRESS which comes from a related prediction criterion.

These and other related penalized optimization approaches to model selection have been used in the regression setting involving continuous or categorical variables. Edwards (2000), for example, uses both AIC

and BIC and suggests variations on computationally efficient model search that take special advantage of the form of graphical models.

4. Bayesian Approaches to Goodness of Fit

Faced with the question of the adequacy of a particular parametric model, a natural Bayesian impulse is to embed the model in question in a parametric model of higher dimension (see Gelman et al. 1995). Having done so, how should a Bayesian evaluate the special model against the more general alternative?

One answer, espoused especially by George Box (1980) is to use a test of significance. However, the probability of observing an outcome as less or more extreme than that observed is only distantly related to goodness of fit, and violates the likelihood principle (see *Significance, Tests of; Likelihood in Statistics*).

A second answer, more in keeping with Bayesian ideas, is to impose a prior distribution on the larger parameter space created by embeddings. Every continuous prior distribution in this larger space has the consequence that the original model would have prior, and hence posterior, probability zero. It is for this reason that L. J. Savage advocated models ‘as big as an elephant’ (see Lindley 2000). A second type of prior distribution that might be used takes the form of a spike-and-slab, i.e., one that puts positive probability on the original model (the spike) and a continuous prior on the rest of the space (the slab). The posterior odds on the original model can then be expressed as the product of the prior odds and the ‘Bayes factor.’

Bayes factors have attracted attention because, in the case of only two alternatives, they give a quantity that depends on only the likelihood function and that indicates the strength of the evidence. Faced with more than two alternatives, and especially with a continuum of alternatives, the Bayes factor critically depends on the prior distribution in the alternative space. When the prior is close to the special model, the alternative is favored; when it puts most of its mass far from the special case, the special case is favored, as is reasonable. Attempts to overcome this dependence using heuristics claimed to be robust to changes in the prior on the alternative space include fractional Bayes factors (O’Hagan 1995) and so-called intrinsic Bayes factors (Berger and Perrichi 1996). Another difficulty with Bayes factors, highlighted by Kadane and Dickey (1980) is that the loss function associated with them is two-valued. Hence if the wrong model is chosen, one is assumed not to care whether the error is quantitatively trivial or not. This is an unreasonable way to think about model choice.

Each of these methods is based on the notion that a better model should fit the data better. If this were true, the very best model for any data set would put a probability of one on the data actually observed, so that whatever happened had to have happened. While such a model is useless for prediction, it does show that

other considerations such as prior beliefs about how the data might have been generated, and utility considerations, must enter into a full explanation of model choice.

Bayesians too have developed approaches that are analogous to the classical approaches discussed above which penalize large models to avoid ‘overfitting.’ For example, for the problem of variable selection for the normal linear model, George and Foster (2000) provide interpretations of AIC, C_p , BIC, and RIC (for Risk Inflation Criterion) under a specific hierarchical Bayesian formulation. Similarly, Madigan and Raftery (1994) suggest several such strategies using posterior model probabilities, including variations of Occam’s razor in order to discard models with low posterior probability. But there is limited agreement on the theoretical basis for such a strategy. Utility considerations should weigh heavily in choosing the form of a penalized search criteria, and there has not been (and it may not be possible for there to be) a convincing argument as to why a Bayesian’s utility ‘should’ be consistent with one of these forms and not another. By contrast, Leamer (1978) presents a more flexible, but *ad hoc* Bayesian view of what he terms ‘specification searches.’

Finally, we mention the proposal to use ‘Bayesian model averaging’ (see Hoeting et al. 1999), which combines posterior information across possible models instead of choosing directly from among them. This approach is closely related to some of the efforts to do Bayesian model selection, but has a further difficulty. The interpretation of parameters in a statistical model depends on other parameters in the model. Thus it remains unclear how to interpret posterior distributions for parameters based on combining results across posteriors for different models.

5. Conclusion

Given the goal of assessing whether a hypothesized statistical model or probabilistic distribution is so far from true as to jeopardize the intended use, it would seem wise to inquire as to the intended use. It would also seem wise not to expect a satisfactory procedure to be independent of the intended use. Second, it seems necessary to address the question of jeopardy, i.e., how much of what kind of departures might make invalid the intended inference. Clearly statisticians and others need to devote more effort to this important topic.

Part of the philosophy behind the objective of model search is to do what our recently deceased colleague Herbert Simon (1959) called ‘satisficing,’ that is to find a model that is good enough to capture the main features of the data but which we know is possibly imperfect in various dimensions. The late John Tukey often reminded statisticians and others that ‘the best is the enemy of the good.’ For many data sets and substantive problems, it is often a fruitless task to try

to choose the ‘best’ model—one which demonstrably dominates all others and whose assumptions can be shown to be correct. But there may be many ‘good’ models that serve the analyst’s purposes well without one that dominates. It is a good idea to pick one of them. Thus, it seems wise to give up the notion that a model might be ‘true.’ If, instead, models are thought of as useful ways of thinking about the world, choosing a model because of its usefulness is reasonable, even if formal ways of explaining that choice are now lacking.

For a detailed consideration of some specific tests and diagnostic measures in the context of assessing the adequacy of distributional assumptions see *Goodness-of-fit Tests and Diagnostics*. Raynor and Best (1989) provide a good technical discussion of specific goodness-of-fit test statistics, and Read and Cressie (1988) give an extended treatment of the power divergence family of statistics and their use for the analysis of categorical data models. Bishop et al. (1975), discuss model search and goodness of fit in the context of categorical data in a form consistent with the philosophical position presented here. Edwards (2000) and Linhart and Zucchini (1986) provide alternative, but useful discussions of related issues for various forms of linear and generalized linear models. Gelman et al. (1995) offer extended advice on Bayesian approaches to model checking.

See also: Goodness-of-fit Tests and Diagnostics; Likelihood in Statistics; Linear Hypothesis: Regression (Basics); Linear Hypothesis: Regression (Graphics); Multivariate Analysis: Discrete Variables (Loglinear Models); Multivariate Analysis: Discrete Variables (Overview); Significance, Tests of

Bibliography

- Akaike H 1973 Information theory and an extension of the maximum likelihood principle. In: Akaike H (ed.) *2nd International Symposium on Information Theory*. Akademia Kiado, Budapest, Hungary, pp. 267–81
- Allen D M 1974 The relationship between variable selection and prediction. *Technometrics* **16**: 125–7
- Berger O, Pericchi L R 1996 The intrinsic Bayes factor for linear models. In: Bernardo J M, Berger J O, Dawid A P, Smith A F M (eds.) *Bayesian Statistics 5*. Oxford University Press, Clarendon, Oxford, UK, pp. 25–44
- Bishop Y M M I, Fienberg S E, Holland P W 1975 *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA
- Box G 1980 Sampling and Bayes’ inference in scientific modeling and robustness. *Journal of the Royal Statistical Society, Series A* **143**: 383–430
- Copas J B 1983 Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B* **45**: 311–54
- Edwards D 2000 *Introduction to Graphical Modeling*, 2nd edn. Springer-Verlag, New York
- Fienberg S E 1980 *The Analysis of Cross-classified Categorical Data*, 2nd edn. MIT Press, Cambridge, MA
- George E I, Foster D P 2000 Calibration and empirical Bayes variable selection. *Biometrika* **87**: 731–47

- Gelman A, Carlin J B, Stern H S, Rubin D B 1995 *Bayesian Data Analysis*. Chapman & Hall, London
- Hoeting J A, Madigan D, Raftery A E, Volinsky C T 1999 Bayesian model averaging: A tutorial (with discussion). *Statistical Science* **14**: 382–417
- Kadane Y B, Dickey J O 1980 Bayesian decision theory and the simplification of models. In: Kmenta J, Ramsey J (eds.) *Evaluation of Econometric Models*. Academic Press, New York, pp. 245–68
- Koehler K J, Larntz K 1980 An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association* **75**: 336–44
- Leamer E E 1978 *Specification Searches: Ad Hoc Inference With Nonexperimental Data*. Wiley, New York
- Lindley D V 2000 The philosophy of statistics (with discussion). *The Statistician* **49**: 293–337 (see p. 307)
- Linhart H, Zucchini W 1986 *Model Selection*. Wiley, New York
- Madigan D, Raftery A E 1994 Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* **89**: 1535–46
- Mallows C L 1973 Some comments on C_p . *Technometrics* **15**: 661–76
- O'Hagan A 1995 Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society, Series B* **57**: 99–138
- Pearson K 1900 On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, 5th Series* **50**: 157–75
- Rayner J C W, Best D J 1989 *Smooth Tests of Goodness of Fit*. Oxford University Press, Clarendon, Oxford, UK
- Read T R C, Cressie N A C 1988 *Goodness-of-fit Statistics for Discrete Multivariate Data*. Springer-Verlag, New York
- Schwarz G 1978 Estimating the dimension of a model. *Annals of Statistics* **6**: 461–4
- Simon H A 1959 Theories of decision-making in economics and behavioral science. *American Economic Review* **49**: 253–73
- Thorburn W M 1918 The myth of Occam's razor (in Discussion). *Mind, New Series* **27**: 345–53
- Wilk M B, Gnanadesikan R 1968 Probability plotting methods for the analysis of data. *Biometrika* **55**: 1–17

S. E. Fienberg and J. B. Kadane

Goodness-of-fit Tests and Diagnostics

The millennium year 2000 marked the 100th anniversary of the introduction of the first and possibly best known test of modern statistical inference: the chi-squared test of fit due to Pearson (1900). The purpose of a test of fit is to give an objective measure of how well a probability model agrees with observed data. Otherwise, as Pearson (1900, p. 171) said: 'the comparison of observation and theory in general amounts to a remark—based on no quantitative criterion.' Reviews of Pearson's paper may be found in Plackett (1983) and Rayner and Best (1989, Sect. 2.2).

Rayner and Best (1989) identify three benefits of having applied a goodness-of-fit test:

- a compact description of the data;
- illumination of the data generation mechanisms; and
- validation, if that is the outcome, of distributional assumptions necessary for the application of powerful parametric procedures.

Pearson's chi-squared statistic compares the *observed* counts from a histogram of the data with the counts *expected* assuming that the hypothesized probability model is correct. In this form, Pearson's test addresses the one sample goodness-of-fit problem (see also *Goodness of Fit: Overview*). The multiple sample goodness-of-fit problem seeks to assess if data from a number of populations are consistent with having come from the same population. If the answer is affirmative, then the one sample question may be appropriate. We shall not address the multiple sample goodness-of-fit problem further here, beyond recommending Scholz and Stephens (1987) and Rayner and Rayner (1988) for references and recent work.

For continuous data and for discrete data for which pooling is used, the grouping of the data used in forming the Pearson chi-squared statistic leads to a loss of information. Another approach that does not group the data is to compare the sample or *empirical distribution function* (EDF) with the distribution function specified by the hypothesized probability distribution. The Russian mathematician Kolmogorov (1933) proposed the first EDF test of fit. EDF tests of fit, especially for continuous distributions, are reviewed in D'Agostino and Stephens (1986). Tests of fit for discrete distributions using the EDF approach and Cramer–Von Mises statistics are considered in Choulakian et al. (1994). See also Spinelli and Stephens (1997).

Another problem with Pearson's chi-squared test is that it is an omnibus test, with some power against general alternative distributions to the hypothesized distribution, but with relatively poor power against important directional alternatives. Alternatives which are often of great practical importance are differences between the data and the mean, variance and higher moments of the specified distribution. A test that does have good power at detecting these moment alternatives is Neyman's (1937) 'smooth' test. It has the useful property that it can readily be partitioned into components that give a powerful and detailed scrutiny of the data. See Rayner and Best (1989, Sect. 4.1) for a review of Neyman (1937).

In a series of papers, Rayner and Best developed the Neyman smooth tests to cover both categorized distributions and distributions involving nuisance parameters. See Rayner and Best (1989) for an overview, and their subsequent related papers. In the main, Rayner and Best use orthonormal functions, as did Neyman (1937), but permit nuisance parameters, following Thomas and Pierce (1979) and Kopecky and

Pierce (1979). The Rayner and Best smooth tests may be used to test for very general distributions, provided orthonormal functions on the specified distribution may be constructed. The resulting test is omnibus, weakly optimal, and the test statistic has a convenient asymptotic chi-squared distribution. Components are readily available, and these are asymptotically independent, each asymptotically having the standard normal distribution. The components are often well known tests of fit in their own right, but more importantly, permit a detailed, convenient and informative scrutiny of the data.

LaRiccia (1991) and Pena (1998) have developed different classes of smooth tests of fit using this approach.

The chi-squared tests, the EDF tests and the smooth tests of fit are the main families of tests of fit. Other tests of fit have been suggested, but we shall not consider these here, except to note that correlation tests are sometimes suggested in conjunction with quantile–quantile or probability plots of the data. The Shapiro–Wilk test of normality, for example, may be considered a correlation test. Again see D’Agostino and Stephens (1986).

1. Chi-squared Tests

For Pearson’s χ^2 test, sometimes called the chi-squared test, observations $O_1, O_2, \dots, O_m$ are assumed to come from m non-overlapping classes that are expected to contain $E_1, E_2, \dots, E_m$ observations. The test statistic is

$$\chi_P^2 = \sum_{i=1}^m (O_i - E_i)^2 / E_i.$$

Pearson (1900) initially assumed that for his test of fit, parameters of the probability models were known, and showed that the asymptotic distribution of χ_P^2 is the χ_{m-1}^2 distribution.

There is a considerable literature on how, when there is choice, the cells may best be constructed. The recommendation that for the χ^2 approximation to be adequate, the cell expectation be at least 5, goes back to Sir Ronald Fisher writing in the 1920s. This advice is now seen as extremely conservative. Mann and Wald (1942) suggested that cells be constructed to be equiprobable and that there be a specified number, k say, of such cells. In their recommendation, k increases with the sample of size n , and for even moderate sample sizes this will lead to a relatively large number of cells. It is now known that this can lead to a considerable loss of power, that sometimes the fewer the number of cells, the greater the power. Recent studies suggest that the greatest power can be achieved for heavy tailed distributions with larger k , while lighter-tailed alternatives require smaller k . Roscoe and Byars (1971) recommend that, with equiprobable cells, the average cell frequency should be at least one

with a 5 percent significance level, and at least two with a 1 percent significance level. If cells are not approximately equiprobable, these recommended frequencies should be doubled.

Most interesting applications involve estimating parameters. Pearson asserted that estimating parameters makes no difference. If the cell expectations are calculated from parameters estimated efficiently from the grouped data, then Sir Ronald Fisher, again writing in the 1920s, proved that degrees of freedom for the asymptotic distribution of the test statistic, χ_{PF}^2 say, needed to be reduced by one for each parameter estimated. The asymptotic distribution is χ_{m-q-1}^2 if q parameters are estimated. Chernoff and Lehmann (1954) showed that Pearson’s statistic no longer had an asymptotic chi-squared distribution if parameters were estimated from the raw or ungrouped data. There is a ‘partial recovery’ of the q degrees of freedom ‘lost’ by estimation. Rao and Robson (1974) gave a test of fit statistic χ_{RR}^2 that had an asymptotic χ_{m-1}^2 distribution no matter how many parameters were estimated, provided they are estimated efficiently from the ungrouped data. So all the degrees of freedom ‘lost’ by estimation are recovered by their statistic. Rayner and Best (1989, Sect. 7.4) compared the power performance of the tests based on χ_{PF}^2 and χ_{RR}^2 , and concluded that neither is uniformly superior: recovery of the degrees of freedom does not necessarily give a more powerful test.

There are many possible generalizations of the Pearson’s test. Neyman’s version is based on

$$\chi_N^2 = \sum_{i=1}^m (O_i - E_i)^2 / O_i.$$

One interesting possibility is the Cressie–Read family, based on the test statistic

$$I_\lambda = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^m O_i \left\{ \left(\frac{O_i}{E_i} \right)^\lambda - 1 \right\}.$$

Taking $\lambda = 1$ leads to χ_P^2 , $\lambda = -1$ leads to a modified likelihood ratio statistic, $\lambda = 0$ is the usual likelihood ratio statistic, $\lambda = 0.5$ is the Freeman–Tukey statistic, and $\lambda = -2$ leads to χ_N^2 . Other choices can lead to tests with excellent properties; see Read and Cressie (1988). Read and Cressie (1988) also illustrate the application χ^2 to testing fit of loglinear models (see also *Multivariate Analysis: Discrete Variables (Overview)*; *Multivariate Analysis: Discrete Variables (Loglinear Models)*).

2. EDF Tests

If we have n data points, the empirical distribution function (EDF) is

$$F_n(x) = (\text{number of observations} \leq x) / n.$$

Tests of fit for a hypothesized distribution function, $F(x)$, can be made by looking at the difference between $F_n(x)$ and $F(x)$ for each data point. Kolmogorov (1933) suggested the statistic

$$KS = \max_x |F_n(x) - F(x)|$$

while a generally more powerful test, is based on the Anderson–Darling (1954) statistic. The formula for this statistic can be simplified for various applications, as we illustrate below.

As, for example, in Stephens (1974), it can be shown that the EDF test statistics are linear combinations of components similar to the smooth test components, with decreasing weights. If the null hypothesis specifies a discrete distribution where the number of classes is fixed and not too large (say five or less) we suggest calculating χ^2 to assess goodness-of-fit. Otherwise, the Anderson–Darling test is a good choice, as it weights the important early moment components heavily and gives a powerful omnibus assessment. Looking at the first few components of Neyman’s smooth test in an exploratory data analysis manner complements this omnibus assessment. See the discussion in Sect. 4 below.

2.1 Example 1: Testing Normality

Consider the following set of examination marks from 20 students:

53, 15, 70, 73, 79, 48, 91, 20, 24, 91,
87, 15, 3, 78, 78, 62, 16, 15, 20, 32.

Are these data normally distributed? We now apply the Anderson–Darling test to answer this question. Calculate $\bar{x} = 48.5$ and $s = 30.8$ and then find $y_i = (x_i - \bar{x})/s$ for $i = 1, \dots, 20$, where x_i are the 20 exam marks. Next use a computer routine or tables of the standard normal distribution to find

$$z_i = F(y_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y_i} e^{-\frac{1}{2}x^2} dx$$

and order the z_i as $z_{(1)}, z_{(2)}, \dots, z_{(n)}$. Then the Anderson–Darling statistic, AD say, is

$$AD = -n - n^{-1} \sum_{i=1}^n (2i-1) [\log z_{(i)} + \log(1 - z_{(n+1-i)})],$$

which, for the above data, takes the value 0.931. From tables of critical values such as D’Agostino and Stephens (1986, Table 4.9), or using MINITAB’s Normplot routine, we find a p -value of 0.014. This implies the normality hypothesis is doubtful and the marks are not well summarized by $\bar{x} \pm s$. We suggest a better summary below.

2.2 Example 2: Testing the Poisson Assumption

Consider the number of people in a shopping mall contributing to a charity collector during five-minute intervals:

Number of contributors	0	1	2	3	4	5	6	7	8
Number of intervals	9	27	26	17	10	8	4	3	1

We have $n = 105$. Let the observed and expected counts assuming a Poisson distribution be o_j and $e_j = np_j$, respectively, for $j = 0, 1, \dots, 8$. Now let $z_j = \sum_{i=0}^j (o_i - e_i)$ and $h_j = \sum_{i=0}^j p_i$, in which case

$$AD = n^{-1} \sum_{j=0}^8 z_j^2 \hat{p}_j / [\hat{h}_j (1 - \hat{h}_j)]$$

where the ‘ $\hat{\cdot}$ ’ indicates we have estimated the Poisson parameter by the sample mean. We find that $AD = 1.17824$, and, using Spinelli and Stephens (1997 Table 1), the p -value is between 0.025 and 0.050. Thus the distribution is not well fitted by a Poisson model. We discuss a possible alternative distribution in Sect. 4. It seems plausible that contributors come in clusters rather than at random, as a Poisson model would imply.

3. Neyman Smooth Tests

Suppose we wish to test for a probability density function $f_x(x; \beta)$, where $\beta = (\beta_1, \dots, \beta_q)^T$ is a q by 1 vector of parameters (such as μ and σ in the normal case). This null probability density function is nested in the order k probability density function

$$C(\theta, \beta) \exp \left\{ \sum_{i=1}^k \theta_i \pi_i(x; \beta) \right\} f_x(x; \beta),$$

where $\theta = (\theta_1, \dots, \theta_k)$ is a vector of k real parameters, $C(\theta, \beta)$ is a normalizing constant and $\{\pi_i(x; \beta)\}$ is a set of functions orthonormal on $f_x(x; \beta)$, so that for $f_x(x; \beta)$ continuous,

$$\int_{-\infty}^{\infty} \pi_i(x, \beta) \pi_j(x, \beta) f_x(x; \beta) dx = \delta_{ij} = 1$$

for $i = j$ and zero otherwise.

Sometimes $C(\theta, \beta)$ may not exist but the statistic we are about to define will. Using a random sample $X_1, \dots, X_n$ we test for $f_x(x; \beta)$ parametrically, by testing $H_0: \theta = 0$ versus $K: \theta \neq 0$ with β a vector of nuisance parameters. For small k the alternatives vary ‘smoothly’ from the null, in the sense of Neyman (1937). As an example of this, in testing for the standard normal distribution, the order one alternative is a normal distribution with a mean shift, the

order two alternative is a normal distribution with both a mean shift and a variance shift, and the third order alternative is more involved, with skewness differences to what might be expected as the distribution normal.

Neyman (1937) developed this approach for the case of no nuisance parameters. In this situation the probability integral transformation converts the problem of testing for a completely specified distribution to testing for the uniform (0, 1) distribution. Thomas and Pierce (1979) and Kopecky and Pierce (1979) considered smooth tests for general distributions involving nuisance parameters. These tests are based on a family of alternatives specified using powers of the cumulative distribution function rather than orthogonal polynomials as Neyman (1937) used, and as a consequence they require tables of coefficients for their implementation. The Rayner and Best formulation permits nuisance parameters and uses orthonormal functions, eliminating the need for tables of coefficients. For a more detailed account of the history of the smooth tests, see Rayner and Best (1989, Sect. 1.2).

When nuisance parameters are present, the smooth tests involve their maximum likelihood estimators. In several important cases, such as the binomial, Poisson and normal, these estimators are the solutions to the equations

$$\hat{V}_r = \sum_{j=1}^n \pi_r(X_j, \hat{\beta}) / \sqrt{n} = 0, \quad r = 1, \dots, q.$$

In such cases the smooth test is now based on the statistic

$$\hat{S}_k = \hat{V}_{q+1}^2 + \dots + \hat{V}_{q+k}^2.$$

The $\hat{V}_r$ are components of $\hat{S}_k$, and provide directional tests with high power in one direction and poor power elsewhere, complementing the omnibus test based on $\hat{S}_k$, which has moderate power in the k directions corresponding to the non-zero components. The omnibus test statistic $\hat{S}_k$ has the χ_k^2 distribution, central under the null distribution and noncentral under contiguous alternatives. It can be shown that the score test for testing $H_0: \theta_r = 0$ against $K_r: \theta_r \neq 0$ is based on the statistic $\hat{V}_r^2$. Being score tests, these tests are weakly optimal. If there are no nuisance parameters, no estimation is required and the results indicated hold with $q = 0$ and the $\hat{\beta}$ s removed from the test statistics. Bouleau and Ducharme (1995) give an example of what happens when the maximum likelihood estimators are *not* solutions of $\hat{V}_r = 0$ for $r = 1, \dots, q$.

We should emphasize that different $\{\pi_i(x; \beta)\}$ detect different alternatives using different tests. Which $\{\pi_i(x; \beta)\}$ should be chosen depends on which alternatives one hopes to detect most powerfully. If the distribution were uniform, as in Neyman (1937), we could take the (non-orthogonal) series $\{x^i\}$. The r th component would still compare the data and moments of the hypothesized distribution up to the r th. How-

ever, we prefer, like Neyman, to use orthonormal functions because the components are simply defined and are, at least asymptotically, independent. If it were desirable to test for periodic alternatives we would consider using the series $\{\sqrt{2} \sin(i\pi\sqrt{x})\}$. The aim is to find an orthonormal series that represents the alternatives of interest in as few terms as possible. Greater power results from doing this.

Each family of distributions to be detected requires an independent study. Given the choice of orthonormal family, the small sample distributions of the test statistics and their components should be investigated. Usually we use regression techniques to calculate corrections to the asymptotic null critical points; see Rayner and Best (1989) for details.

Again for each family of distributions we may compare, via simulation studies, the powers of the smooth tests we propose and the commonly used competitor tests. We give the general advice that in a testing context the most powerful goodness-of-fit test is likely to be the sum of the squares of the first two, three, or four non-zero components; $\hat{S}_4$ perhaps. We recommend augmenting this by the use of its components $\hat{V}_{q+1}, \dots, \hat{V}_{q+k}$ in an exploratory data analysis fashion. However, in particular cases there are variations to this advice.

We have so far studied the continuous and discrete uniform, univariate, bivariate and multivariate normal, exponential, geometric, binomial, univariate and bivariate Poisson. Details are in Rayner and Best (1989) and their subsequent relevant papers. In all cases the smooth tests are generally at least as powerful as the competitors studied, and the smooth tests have the added flexibility of meaningful components being available. So if the null hypothesis is rejected, its failure can be attributed to particular θ_i being non-zero, and an alternative model is suggested, namely

$$C(\hat{\theta}, \hat{\beta}) \exp \left\{ \sum_{i=q+1}^{q+k} \hat{\theta}_i \pi_i(x; \hat{\beta}) \right\} f_X(x; \hat{\beta})$$

where the $\hat{\theta}_i = \hat{V}_i / \sqrt{n}$ are unbiased, consistent and efficient estimators of the θ_i . This density estimate is a fortuitous by-product of the testing, and may be used to graph the data or as a basis for further inference, as in Carolan and Rayner (2001).

4. Diagnostics

In Sect. 2 we suggested use of the Anderson–Darling statistic as a test of fit. However, this will only tell us whether a given probability model fits the data well or not. It does not suggest or ‘diagnose’ how the data differs from the probability model. A histogram, density estimate or a Q–Q plot might shed some light on this difference, but interpretation of these plots is subjective. A more objective data scrutiny is given by calculating the components of Neyman’s smooth test which are approximately distributed as independent

standard normal random variables. We now illustrate this ‘diagnostic’ approach in our previous examples and one further example. There is no guarantee the diagnosis is correct (see Rayner et al. 1995), but it will be suggestive.

4.1 Example 1: Testing Normality Revisited

For the exam mark data in Section 2.1, calculate $y_i = (x_i - \bar{x})/s$. The orthonormal polynomials appropriate for testing normality are often called Hermite–Chebyshev polynomials. We have

$$\begin{aligned}\pi_1(y) &= y \\ \pi_2(y) &= (y^2 - 1)/\sqrt{2} \\ \pi_3(y) &= (y^3 - 3y)/\sqrt{6} \\ \pi_4(y) &= (y^4 - 6y^2 + 3)/\sqrt{24} \\ \pi_5(y) &= (y^5 - 10y^3 + 15y)/\sqrt{120} \\ \pi_6(y) &= (y^6 - 15y^4 + 45y^2 - 15)/\sqrt{720}\end{aligned}$$

and given these values we can calculate, for the above data, $\hat{V}_1 = \hat{V}_2 = 0$, due to the estimation of μ and σ from $\bar{x}$ and s respectively, and $\hat{V}_3 = -0.0361$, $\hat{V}_4 = -1.4656$, $\hat{V}_5 = -0.0462$, $\hat{V}_6 = 1.8907$. Note that $\hat{V}_3$ and $\hat{V}_4$ are standardized versions of the usual skewness and kurtosis coefficients. As $\hat{V}_3, \dots, \hat{V}_6$ have asymptotic standard normal distributions, we see that $\hat{V}_6$ is near significance at the 5 percent level and that $\hat{V}_4$ and $\hat{V}_6$ are much larger than the odd-order moment statistics $\hat{V}_3, \hat{V}_5$. These observations suggest an alternative probability model that is symmetric and shorter tailed than the normal. Thus a uniform distribution is a possibility so that we could summarize the data as ‘evenly’ spread over the range (0, 100) rather than as 48.5 ± 30.7 . If the odd order $\hat{V}_r$ had dominated the even order values then a skewed alternative model is suggested, while if both odd and even $\hat{V}_r$ are large, $|\hat{V}_r| > 1.65$ say, then a symmetric alternative with longer tails than the normal, such as the Laplace, is indicated.

What should be done if one or more $\hat{V}_r$ are significant but AD is not? One possibility would be to obtain more data and test the significance of those $\hat{V}_r$ previously significant.

4.2 Example 2: Testing the Poisson Assumption Revisited

For the Poisson, the appropriate orthonormal polynomials are known as Poisson–Charlier polynomials. We have

$$\pi_r(y) = \sqrt{(\lambda^r/r!)} \sum_{j=0}^r [(-1)^{r-j}] ({}^r C_j) (j!) (\lambda^{-j}) ({}^r C_j)$$

where $y = 0, 1, 2, \dots$, from which the first two orthonormal polynomials are

$$\begin{aligned}\pi_1(y) &= \sqrt{\lambda} (y/\lambda - 1) \\ \pi_2(y) &= \sqrt{\lambda^2/2} [y(y-1)/\lambda^2 - 2y/\lambda + 1].\end{aligned}$$

Due to the estimation of λ , $\hat{V}_1 = 0$. For the data here $\hat{V}_2 = 2.12$, indicating a model which has greater dispersion than the Poisson. Note that $\hat{V}_2$ is a standardized version of Fisher’s Poisson Index of Dispersion.

4.3 Example 3: Testing the Ordered Categories Discrete Uniform Distribution

In a market research trial, 40 consumers were asked for their color preference of five orange cordial drinks with orange colors ranging from ‘pale’ to ‘deep.’ Arranged in order of increasing orange color, the preference counts were 5, 10, 11, 10 and 4. If there were no color preference, the counts would all be expected to be eight. As there are only five unambiguous classes, a Pearson χ^2 could be used here. We have

$$\chi^2 = (5-8)^2/8 + (10-8)^2/8 + \dots + (4-8)^2 = 5.25$$

which, on using the asymptotic χ^2_4 distribution, gives a p -value of 0.27. Using the same formula for AD as that used for the Poisson test, we find the Anderson–Darling statistic takes the value 0.89. Consulting tables in Choulakian et al. (1994) we find the p -value is about 0.30. It appears the data are consistent with a discrete uniform distribution. However, we now compare expected and observed moments via components of Neyman’s smooth test. Here the orthonormal polynomials are the Chebyshev polynomials and the first two are

$$\begin{aligned}\pi_1(y) &= \sqrt{[12/(m^2-1)]} [y - (m+1)/2] \\ \pi_2(y) &= \sqrt{\{180/[(m^2-1)(m^2-4)]\}} \{[y - (m+1)/2]^2 \\ &\quad - (m^2-1)/12\}\end{aligned}$$

where m is the number of categories and $y = 1, 2, \dots, m$. The higher order $\pi_r(y)$, if needed, can be found by recurrence relation as in Emerson (1968). As in Stephens (1974), it can be shown that for the discrete uniform model

$$\chi^2 = \sum_{i=1}^m V_i^2 \quad \text{and} \quad AD = \sum_{i=1}^m [i(i+1)]^{-1} V_i^2$$

where the components are defined by

$$V_r = \sum_{j=1}^n \pi_r(X_j) / \sqrt{n}, \quad r = 1, \dots, m$$

since no nuisance parameters need estimation.

For these data, neither χ^2 nor AD was significant. However $V_1 = \sqrt{0.05} = 0.22$ and $V_2 = \sqrt{5.14} = 2.27$, indicating possible extra dispersion in the data compared with what was expected for the discrete uniform. We find that V_2^2 explains almost all χ^2 , so that dispersion differences are all that can be important. The consumers seem to have preferred a mid-range orange-colored drink. Unless it is the first component, AD will not be powerful if only one component is important, as in the present example. However, for many data sets, more than one component will be important, so that AD should usually have reasonable power.

In the above examples, examination of smooth test components gives an informative and thorough scrutiny of the data. Formal tests of significance would require the Bonferroni adjustment to the α level; here we are recommending an informal exploratory data analysis approach.

5. Conclusion

Statistical model assessment is at the heart of good statistical practice, and is the genesis of modern statistics (see *Goodness of Fit: Overview*). In general we suggest using the Anderson–Darling test as an omnibus test, augmented by the use of the components of the smooth tests in an exploratory data analysis fashion. This will assist in diagnosing the type of deviation from the hypothesized distribution. To find out more about the chi-squared and EDF tests see D’Agostino and Stephens (1986). For more about the smooth tests see Rayner and Best (1989).

Bibliography

- Anderson T W, Darling D A 1954 A test of goodness of fit. *Annals Mathematical Statistics* **23**: 193–212
- Boulerice B, Ducharme G R 1995 A note on smooth tests of goodness of fit for location-scale families. *Biometrika* **82**: 437–8
- Carolan A M, Rayner J C W 2001 One sample score tests for the location of modes of nonnormal data. *Journal of Applied Mathematics and Decision Sciences* **5**(1): 1–19
- Chernoff H, Lehmann E L 1954 The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Annals Mathematical Statistics* **25**: 579–86
- Choulakian V, Lockhart R A, Stephens M A 1994 Cramer–Von Mises statistics for discrete distributions. *Canadian Journal of Statistics* **22**: 125–37
- D’Agostino R B, Stephens M A 1986 *Goodness-of-fit Techniques*. Marcel Dekker, New York
- Emerson P L 1968 Numerical construction of orthogonal polynomials from a general recurrence formula. *Biometrics* **24**: 695–701
- Kolmogorov A N 1933 Sulla determinazione empirica di una legge di distribuzione. *Giornale dell’ Istituto Italiano degli Attuari* **4**: 83–91
- Kopecky K J, Pierce D A 1979 Efficiency of smooth goodness-of-fit tests. *Journal of the American Statistical Association* **74**: 393–7
- LaRiccia V N 1991 Smooth goodness-of-fit tests: a quantile function approach. *Journal of the American Statistical Association* **86**: 427–31
- Mann H B, Wald A 1942 On the choice of the number of intervals in the application of the chi-squared test. *Annals Mathematical Statistics* **13**: 306–17
- Neyman J 1937 ‘Smooth’ test for goodness of fit. *Skandinavisk Aktuarietidskr* **20**: 150–99
- Pearson K 1900 On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Philosophy Magazine*, 5th series **50**: 157–75
- Pena E A 1998 Smooth goodness-of-fit tests for the baseline hazard in Cox’s proportional hazards model. *Journal of the American Statistical Association* **93**: 673–92
- Plackett R L 1983 Pearson Karl and the chi-squared test. *International Statistical Review* **51**: 59–72
- Rao K C, Robson D S 1974 A chi-square statistic for goodness-of-fit tests with the exponential family. *Communication in Statistics*. **3**: 1139–53
- Rayner J C W, Best D J, 1989 *Smooth Tests of Goodness of Fit*. Oxford University Press, New York
- Rayner J C W, Best D J, Mathews K L 1995 Interpreting the skewness coefficient. *Communications in Statistics—Theory and Methods* **24**(3): 593–600
- Rayner J C W, Rayner G D 1988 S-sample smooth goodness of fit tests: Rederivation and Monte Carlo assessment. *Biometrical Journal* **40**: 651–63
- Read T R C, Cressie N A C 1988 *Statistics for Discrete Multivariate Data*. Springer-Verlag, New York
- Roscoe J T, Byars J A 1971 Investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic. *Journal of the American Statistical Association* **66**: 755–9
- Scholz F W, Stephens M A 1987 K-sample Anderson–Darling tests. *Journal of the American Statistical Association*. **82**: 918–24
- Spinelli J J, Stephens M A 1997 Cramer–Von Mises tests of fit for the Poisson distribution. *Canadian Journal of Statistics* **25**: 257–68
- Stephens M A 1974 Components of goodness-of-fit statistics. *Annals of the Institute of Henri Poincaré* **10**: 37–54
- Thomas D R, Pierce D A 1979 Neyman’s smooth goodness-of-fit test when the hypothesis is composite. *Journal of the American Statistical Association* **74**: 441–5

J. C. W. Rayner and D. J. Best

Government Lawyers

Government lawyers are those officers and agents that provide legal counsel to government policy makers and represent governments in judicial proceedings. The term ‘government lawyers’ is not common in social science literature, and, generally speaking, studies of government lawyering focus on individual

offices (such as attorneys general, ministers of justice, solicitors general, district attorneys, etc.), or the legal policies of particular governments or administrations (such as the Thatcher Government or the Bush Administration, etc.). The reluctance to treat government attorneys as a group stems from the fact that they perform a wide variety of roles, from routine criminal prosecution to advising presidents or prime ministers on the broad policy concerns embedded in constitutional or statutory law. Moreover, the role of government lawyers, like the role of lawyers and courts generally, varies a great deal from one political system to the next. Nevertheless, commonalities do exist, and scholarly attention to the role of government lawyers has grown as their political importance has expanded alongside the legal-administrative and regulatory state.

1. Size, Structure, and History of Government Legal Offices

The structure, number, and size of government law offices usually reflects the broader constitutional structures and policy making processes of the regime. Thus, federal or confederal systems usually have separate law officers at the national level and at the state, *Länder*, or canton levels. Likewise, separation of power systems tend to have more legal offices than parliamentary systems, as during periods of conflict each branch of government finds it necessary to access independent legal advice and representation. In countries such as the United States, where there is federalism, a separation of powers, and where courts exercise sweeping policy making powers, governmental lawyers are both numerous and important policy makers (Clayton 1995).

There is no accurate count of the total number of lawyers employed by governments around the world, but in the USA alone the Department of Justice employs nearly 11,000 attorneys and has a budget in excess of \$21 billion (DOJ website 2000). Thousands more are employed in the general counsels' offices of various executive branch agencies and departments, including the White House Legal Counsel's Office. Congress too employs numerous attorneys, either in the House or Senate Legal Counsel's offices, or as staff counsel to various committees. Even the Supreme Court, since 1972, has employed attorneys in its own in-house legal counsel's office to advise the Court and its members. At the state level, governmental law offices range in size from the Department of Law in California, which employs nearly 1,000 lawyers, to relatively small Attorney General's offices in states such as Montana. In addition, many more attorneys work for cities, counties, municipalities, and other local governmental units (Clayton 1994).

The premier government attorney in the USA is the Attorney General. The office was established under

the Judiciary Act of 1789, and originally operated as a quasi-judicial institution. By the time Congress created the Department of Justice in 1870, the office had become thoroughly entrenched as part of the executive branch establishment. The Attorney General is appointed by the President and confirmed by the Senate. The office is a member of the Cabinet and a key advisor to the President on judicial selection and the making of administration policy. The Attorney General's transformation from quasi-judicial to executive, however, left deep tensions in its normative relationship to the President, especially during periods of separation of powers conflict (see Baker 1992). The office's relationship to other government lawyers has also changed over time. Originally, the Attorney General had little authority over the US District Attorneys (prosecutors in each federal judicial district), or over the solicitors that existed in executive branch departments and agencies. The early history of the Attorney General's office is marked by struggles to consolidate control over the federal governments' legal activities. Eventually, a division of labor emerged wherein agency legal counsel have assumed primary responsibility for advising and counseling government agencies and departments, and the Justice Department, under the direction of the Attorney General, has assumed nearly complete responsibility for conducting federal litigation and prosecution (see Horowitz 1977).

The Attorney General is an administrator of a large department. The office is assisted by a Deputy Attorney General, an Associate Attorney General, and a Solicitor General. The latter has evolved into an elite barristers office, responsible for conducting nearly all of the federal government's appellate litigation, enjoying a special relationship with the Supreme Court (Sakokar 1992). The Justice Department is divided into 15 major divisions and offices, including important policing and corrections divisions such as the Federal Bureau of Investigation, the Drug Enforcement Administration, the Marshals Service, the Immigration and Naturalization Service, and the Bureau of Prisons. The most important legal divisions, each headed by an Assistant Attorney General, include: Antitrust, Civil Rights, Land and Natural Resources, Criminal, Civil, and the Tax divisions. Additionally, the Office of Legal Counsel, also headed by an Assistant Attorney General, is responsible for authoring the official opinions of the Attorney General. Typically, these are considered binding within the executive branch but not binding on courts, although they are usually accorded respect by the judiciary.

In the USA, most state attorney generals' offices or law departments mirror the subject-area structure of the Justice Department's legal divisions. Thus, they are likely to have divisions for antitrust, tax, consumer protection, corporate regulation, etc. Several important distinctions exist at the state level in the USA, however. First, in most states the attorney general is responsible for providing legal advice to all branches

of government. Second, responsibility for nonlawyering functions (police, prisons, etc.) and for most criminal investigations and prosecution are not usually under the control of the attorney general, but under the control of other state-level administrators or locally elected prosecutors. Third, most state attorneys general are elected directly and politically independent of the governor and legislature (Clayton 1994).

In most other countries, the lawyering activities that involve broad policy concerns are usually also separated from those related to routine prosecutions, corrections, and policing. In many, an attorney general, or some similar office, is responsible for advising the government on legal matters, and for litigation enforcing public or political rights, while responsibility for prosecution, police, corrections, and the administration of justice is vested in a Minister of Justice or a Home Secretary.

In England, the Attorney General's title dates back to 1472. The modern office typically is held by a member of the House of Commons and an elected member of the majority political party. By convention, the Attorney General is not usually a member of the Cabinet, although most attend Cabinet meetings. The Solicitor General serves as a chief deputy. These two officials are known collectively as 'the Law Officers of the Crown' (Edwards 1964). Unlike its American counterpart, the Attorney General in England is not an administrator or department head. However, the office has supervisory responsibility over the Treasury Solicitor, the Director of Public Prosecutions as head of the Crown Prosecution Service, the Director of the Serious Fraud Office, and the Director of Public Prosecutions in Northern Ireland, and answers for each of these departments in Parliament.

The Law Officers' primary responsibility is to represent the public interest in court and act as the chief legal advisers to the government. On matters of political importance, the Attorney General personally will offer legal advice to the government. Routine legal advice, however, is provided to each department or ministry by the Treasury Solicitor's Department, or by small in-house departmental legal staffs, which are analogous to the general counsel's offices in various departments in the USA. The Treasury Solicitor's Department is the largest in-house legal organization and provides legal advice to administrative officials across the executive, including the departments of Treasury, the Cabinet Office, the Ministry of Defence, and the Departments for Education and Employment, and Culture, Media and Sport. The Department employs several hundred attorneys and it also assists the Attorney General in litigation to protect the public interest.

The Home Secretary is responsible for the police, and for administering immigration, the correctional system, and other law enforcement agencies. The appointment of judges is handled exclusively by the Lord Chancellor's Office, a part of the House of

Lords. The chief prosecutor in England is the Director of Public Prosecutions, an officer appointed by the Home Secretary and located in the Home Office, but who is responsible to the Attorney General in the exercise of prosecutorial discretion. The Director heads the Crown Prosecution Service, which is divided administratively into 42 separate areas, each corresponding to the boundaries of local police forces. Each area is headed by a Chief Crown Prosecutor who works with panels of local attorneys and case workers to prosecute crime in the crown courts. Crown Prosecutors report to the Director of Public Prosecutions, but in routine matters they are locally accountable to the public and cooperate closely with the area Chief Constable and other agencies in the criminal justice system. If criminal prosecutions raise serious political concerns, however, the Director of Public Prosecutions or the Attorney General may become involved and assume responsibility for the prosecution. As with much else in English government, prosecutorial independence is protected through custom and convention. Ever since the famed Campbell case, which brought the downfall of the Macdonald government in the 1920s, the constitutional rule is that the Attorney General may seek advice from the Cabinet when making prosecutorial decisions but contacts initiated in the other direction give rise to serious questions of propriety.

In countries influenced by the French and German models, prosecutors are often a part of the judiciary, or under judicial supervision in an essentially inquisitorial system. In such systems, judges and government prosecutors are often located administratively in the Ministry of Justice, but they are buffered from direct political control. In Italy, for instance, public prosecutors are under the constitutional control of the independently elected Supreme Judicial Council, or *Consiglio Superiore delle Magistratura*. Likewise, in Japan, a hybrid system, the Office of the Public Prosecutor is part of the Ministry of Justice (*Homusho*), but the Prosecutor General (*Kenji Socho*) is appointed directly by the Cabinet, and the Minister of Justice (*Homu Daijin*) is prohibited by law from exercising control over specific cases.

Even in systems more clearly based on the Anglo-American model, certain prosecutions, especially those involving government officials, are institutionally protected from political control. In England, the Attorney General's prosecutorial decision-making is protected by custom and convention. In the USA, following the infamous Watergate scandal, Congress established the office of the Independent Counsel. Independent counsels were appointed by a special division of the court and buffered from direct political control of the President or the Attorney General (Harriger 1992). Controversy surrounding the Independent Counsel's investigation of the Whitewater scandal and the eventual impeachment of President Clinton, however, led to the office's disestablishment

in 2000. Nevertheless, Attorney Generals continue to have authority to appoint 'special prosecutors' inside the Department of Justice, who enjoy nearly complete independence from political direction by the elected executive.

2. *The Intersection of Politics and Law in Government Lawyering*

Protecting prosecutorial independence is one part of a larger concern regarding the role of politics in government lawyering. In this respect, two separate problems dominate the scholarly literature in this area. The first has to do with the separation of powers and the litigating role of government lawyers during periods of constitutional conflict. Administrations who use government litigation to challenge established judicial precedents, for example, are often charged by critics with violating the rule of law or other constitutional constraints (Caplan 1988). Moreover, in the USA it is not uncommon for government lawyers representing different branches to disagree in court about the meaning of a statute or a constitutional provision. In *Morrison vs. Olson* (1988), for instance, Justice Department attorneys argued against the constitutionality of the Independent Counsel Act, while counsel for Congress defended the statute.

Conflicts such as these put government attorneys in an awkward ethical position, and raise serious questions about the appropriate role of government lawyers in a system of separated powers. Moreover, in the USA such conflicts have become more common as prolonged periods of divided government have led partisans increasingly to turn to courts and the administrative bureaucracy, rather than the routine legislative process, for policy making. Scholars and practitioners have written a great deal about the principles that should guide government lawyers when two or more branches disagree about the law. However, such principles are usually just as controversial as the contested notions of separation of powers upon which they rest.

While government lawyers in parliamentary systems escape the political entanglements caused by a separation of powers, they cannot avoid the 'politics' of partisan abuse, corruption, and favoritism. There has been a great deal of scholarly debate about which institutional structures best safeguard government lawyers from 'inappropriate' political pressures. Anglo-American systems have experimented with structural reforms, such as the Independent Counsel's Office, as well as informal protections such as conventions that prevent political leaders from contacting government prosecutors in particular cases. All such efforts, however, suffer from the difficulty of distinguishing between proper and improper political influences.

In a well-known case in the USA in 1971, for

example, President Nixon ordered the Justice Department to drop a pending antitrust suit against ITT Corporation in exchange for a \$400,000 donation to Nixon's re-election campaign. This led critics to argue for the construction of a 'fire-wall' that would prevent White House officials from contacting the Justice Department in particular cases. However, contrast that White House intervention with President Carter's 1978 decision to reverse the Justice Department's position in *Regents of the University of California vs. Bakke* (1978), the first major affirmative action case to reach the Supreme Court. Or consider whether President Clinton should have had no role in deciding whether the Justice Department should file antitrust charges against Microsoft Corporation in 1999. Some cases, whether civil or criminal, present legal issues or involve individuals in a way that the rule of law properly requires direction from democratically accountable officials. Others do not. However, there is no entirely satisfactory way to distinguish *a priori* between proper and improper political influences. Moreover, removing prosecutors from political control only shifts, but does not remove, the influence of partisanship. Critics of the Independent Counsel's office in the USA, for example, long complained that most investigations, such as those surrounding Watergate or the Iran-Contra scandals, were politically motivated (Eastland 1989).

By contrast, countries relying on the French and German models tend to protect prosecutorial independence by trying to remove it entirely from 'political' influence and lodging the power under the direction of the judiciary. Such systems thus trade off democratic accountability over criminal prosecution for judicial control. However, they too only shift, rather than remove, the sources of political influence and must also confront the difficulty of distinguishing between proper and improper influences. Even judicial officials are accountable to broader regime structures and they too can be subject to corruption and partisan abuse. Determined political leaders have had no more trouble influencing prosecutors administratively lodged in the judiciary as those in the executive. In the well-known Credit Lyonnais case in France, for instance, President François Mitterrand was able to use party influence to control decisions in the Ministry of Justice over the prosecution of corruption charges involving government ministers, and the French and Italian Socialist parties, and leading businessmen with alleged Mafia ties. Likewise, in Italy, where the independence of prosecutors is fixed constitutionally, a well-known effort to prosecute political corruption charges against more than 400 political figures, including several ministers and members of Parliament in the mid-1990s foundered when Parliament approved a law fixing a deadline for the disposal of cases. In a system notorious for delay and inefficiency, the law effectively gave immunity to many of those who had been indicted.

The intersection of law and politics in government lawyering, and the administrative structures governing it, will likely continue to dominate scholarly attention in the future. As noted above, existing studies of government lawyering tend to focus on particular offices or administrations. Little work of a comparative nature, either within, or across, different political systems has been undertaken, and this remains perhaps the most fruitful direction of future research.

See also: Counsel, In-house; Governments; Lawyers; Legal Professionalism; Political Lawyering; Professions in Organizations; Public Interest Law

Bibliography

- Baker N 1992 *Conflicting Loyalties: Law and Politics in the Attorney General's Office*. University Press of Kansas, Lawrence, KA
- Caplan L 1988 *The Tenth Justice: The Solicitor General and the Rule of Law*. Vintage, New York
- Clayton C W 1992 *The Politics of Justice: The Attorney General and the Making of Legal Policy*. Sharpe, New York
- Clayton C W 1994 Law, politics and the new federalism: State attorneys general as national policymakers. *The Review of Politics* 56: 525
- Clayton C W 1995 *Government Lawyers: The Federal Legal Bureaucracy and Presidential Politics*. University Press of Kansas, Lawrence, KA
- Eastland T 1989 *Ethics, Politics and the Independent Counsel: Executive Power, Executive Vice*. National Legal Center, Washington, DC
- Edwards J L J 1964 *Law Officers of the Crown*. Sweet & Maxwell, London
- Edwards, J L J 1984 *The Attorney General, Politics, and the Public Interest*. Sweet & Maxwell, London
- Eisenstein J 1978 *Counsel for the Untied States*. Johns Hopkins University Press, Baltimore
- Fionda J 1995 *Public Prosecutors and Discretion: A Comparative Study*. Oxford University Press, Oxford, UK
- Harriger N 1992 *Independent Justice: The Federal Special Prosecutor in American Politics*. University Press of Kansas, Lawrence, KA
- Horowitz D 1977 *The Jurocracy*. Heath, Lexington, MA
- Perrot R 1995 *Institutions Judiciaires*. Montchestrien, Paris
- Sakokar R M 1992 *The Solicitor General: The Politics of Law*. Temple University, Philadelphia

C. W. Clayton

Government Statistics

Government statistics are vital to advance the economic well-being and quality of life of a nation's citizens. Without such statistics, policy making would be less effective, economic markets would be less informed, and the countless decisions of private

businesses and individuals would allocate economic and other resources less efficiently. But more, government statistics are vital to a democracy, as the information they convey provides the means by which citizens can hold their government accountable.

This article provides a general review of government statistics largely in the context of the USA, but it also addresses broader issues relevant to other countries. The focus is on the national government and, therefore, mostly on national statistics, but much is also relevant to governments at subnational levels, which also produce statistics.

The article begins with the purposes of government statistics and the rationale for why governments produce them. Next it takes up the principles by which the statistics should be produced to serve these purposes effectively; this is followed by descriptions of various types of data collection, and of how the responsibilities for collection, analysis, and reporting of statistics are organized among statistical agencies. Effective practices of these agencies are identified. The entry concludes with a discussion of some challenges faced by statistical agencies.

Statistics from government sources are referred to in many entries in this Encyclopedia (see *Censuses: History and Methods*; *Censuses: Comparative International Aspects*; *Economic Panel Data*; *Sample Surveys: Cognitive Aspects of Survey Design*; *Large-scale Social Survey, Analysis of*).

1. What are Government Statistics?

'Government statistics' means factual information that is reported upon for statistical purposes by a unit of government. Some countries also use the term 'official statistics.' In this articles, 'statistics' and 'data' are used interchangeably to mean factual information, and the terms are not restricted to quantitative information. Usually, the unit or agency of government that issues the statistics has collected or compiled the data to produce them. In most cases, the agency is a distinct entity of government entrusted with reporting the statistics. In some cases, the agency may analyze its data to provide more information about the statistics in its reports, especially information that makes the statistics more relevant to public policy concerns. For the agency to be effective, the focus of any analysis must be on ways to improve the relevance and accuracy for use in public policy analysis and decision making, independently of any particular policy.

The *sine qua non* of the definition of government statistics is that they are reported upon only for *statistical purposes*:

'Statistical purposes include description, evaluation, analysis, inference, and research. For these purposes, a statistical agency may collect data from individuals, establishments, or

other organizations directly, or it may obtain data from administrative records, but it does not do so for administrative, regulatory, or enforcement purposes. Statistical purposes exclude any interest in an individual person or economic unit. The data are used solely to describe and analyze statistical patterns, trends, and relationships involving groups of persons or other units' (Martin et al. 2000)

2. *Why do Governments Produce National Statistics?*

A government provides statistics to inform its citizens, satisfying an essential requirement of a democracy (see, e.g., United Nations Economic and Social Council 1994). The government, however, is most often the primary user of the statistics it produces. The statistics are used in the formulation of, and decision making on, public policies, as well as in their implementation and evaluation. And the statistics are used for much private-sector and individual decision making. Information from statistics is, for example, vital to inform economic markets.

In most cases, a government collects and reports the statistics itself, either directly or through the services of a private organization. In some cases, the private sector satisfies the demand for national statistics. Examples are the Conference Board's compilation of US leading economic indicators, abortion statistics for many countries compiled by the Alan Guttmacher Institute, and many private market-research surveys and reports of economic forecasting firms. But even when the private sector can do so, there are reasons for a government to collect the statistics. Alternative sources may not exist. Timeliness may be a critical factor. Statistics collected by government agencies are often deemed to be of greater quality. And the government may collect statistics because they contain confidential information.

A more compelling reason may be trust and credibility. Potential subjects of government data collection may be more willing to volunteer their time and effort, and divulge confidential information, for a government survey rather than for a private one. And the public and other important users may find data collected by the government to be more credible.

A fundamental reason, however, for a government to collect data has to do with the very nature of data. If a commodity can be consumed by many and consumption by one does not reduce the consumption of others, then the commodity is a 'public good.' Data are manifestly public goods: use of data by one party in no way diminishes their usefulness to other parties. Individuals benefit from a public good but, unless each individual is charged for the consumption of the good, they are unlikely to pay for them voluntarily. Consequently, the goods may not be produced, unless the government intervenes. In the case of data, the government may intervene directly by collecting the data, or it may intervene indirectly, for example,

by defining rights for a license or copyright to provide the means for the private sector to charge for the value added in producing secondary data. For these reasons, a government typically provides primary data and the information infrastructure. The private sector typically provides secondary data and may build on the infrastructure.

3. *Principles of Government Statistics*

The US National Academy of Sciences–National Research Council report (Martin et al. 2000) defines a 'statistical agency' as a unit of government whose principal function is the compilation and analysis of data, and the dissemination of information for statistical purposes. The unit is recognized generally as a distinct entity, which may be located in a cabinet department or independent agency. The mission of a distinct entity or agency of government that issues government statistics may or may not be solely statistical. The USA has a highly decentralized statistical system, with over 70 agencies having direct funding of over \$500,000 per year for statistical activities. Only some of these agencies are statistical agencies.

Statistical agencies are established to be a credible source of accurate, timely, and relevant statistics in one or more subject areas. Accuracy requires proper concern for consistency in spatial and temporal dimensions, as well as statistical measures of error. Timeliness requires concern for the frequency of issuing data, as well as disseminating data to users when they need them. Relevance requires concern for anticipating future data needs, as well as meeting current ones. And credibility requires concern for both the reality and appearance of impartiality, as well as the reality and appearance of independence from political control. It is essential to the mission of statistical agencies to work to ensure the goals of accuracy, timeliness, relevance, and credibility of statistical information.

Martin et al. (2000) identify three principles for a statistical agency. First is that the agency is in a position to provide information relevant to issues of public policy. That requires the agency to be knowledgeable about public policy issues and be able to provide objective information that is relevant to the needs of formulating, implementing, and evaluating public policies. Second is that the agency has a relationship of mutual respect and trust with those who use its data and information. To develop that relationship, an agency must strive to maintain credibility for itself and for its data. In particular, the agency must be perceived to be free of political interference and policy advocacy. Other aspects that contribute to credibility are a commitment to quality and professional standards, and openness about all aspects of the data, especially the quality of the data.

The third principle is that the agency have a relationship of mutual respect and trust with respondents who provide data, and with all data subjects whose information it obtains. Statistical agencies rely in large part on information supplied by individuals and organizations outside the federal government, such as state and local governments, businesses, and other organizations. Some of this information is required by law, such as employers' wage reports, and some of it is compiled from records of government programs. But much of it is obtained through the voluntary cooperation of respondents to surveys. Even when the response to a survey is required by law, cooperation by respondents promotes accuracy (Edmonston and Schultze 1995), which, in turn, promotes the credibility of the agency's data. In encouraging cooperation, it is important to statistical agencies that respondents perceive that the information they are being asked for is important; that it is being collected in an impartial, professional manner; and that their personal information will remain confidential.

Confidentiality of individual responses means that the information will be used only for statistical purposes and that, in particular, it will not be used for any administrative, regulatory, or enforcement purpose regarding the individual (see *Confidentiality and Statistical Disclosure Limitation*). Statistical agencies go to great lengths to maintain the confidentiality of individual responses (see Duncan et al. 1993, Mackie and Bradburn 2000). Some, such as the US Bureau of the Census, have legislation to protect confidentiality, and some, such as the US Bureau of Labor Statistics, have a history of successful reliance on executive order, custom, and case law. A Federal Statistical Confidentiality Order (US Office of Management and Budget 1997) directs 12 statistical agencies not to disclose or use data acquired for exclusively statistical purposes in identifiable form for any other purpose unless otherwise compelled by law. In Europe, the 1995 European Community Directive on Data Protection governs the transfer of individually identified or identifiable data.

The respect and trust of data providers are also essential for statistical agencies to achieve credibility for their data. Protecting the confidentiality of individual responses engenders trust, but, as was noted, agencies may be limited in their abilities to do so. In that case, agencies need to provide respondents with realistic promises of confidentiality that the agency can reasonably be expected to honor.

Respect for data providers also means valuing their time and effort to respond—in particular, designing data collection so that it is not unduly burdensome—and valuing their responses so that they are put to important uses. It is important too that respondents perceive that the agency collecting their information is trustworthy and credible; in particular, that the information will not be used for a particular partisan

interest. A statistical agency must be perceived, by its data providers as well as its data users, to be working in the national interest.

4. *How are Government Statistics Collected and Reported?*

Statistical agencies collect data through censuses, surveys, and compilations from administrative records. Many countries conduct a census of their population. In the UK and USA, one is conducted every 10 years; in Canada, every five years. Other examples of US censuses are the quinquennial census of agriculture, which covers agricultural production on farms that are defined as places from which \$1,000 or more of agricultural products were produced and sold in the census year; the quinquennial economic census, which covers the economic activity of nonfarm establishments; and the quinquennial census of governments, which covers employment and finances of state and local governments.

Examples of surveys in the USA include the Current Population Survey, from which monthly employment and unemployment data are derived; the Consumer Expenditure Survey, from which data on expenditure patterns are used periodically to adjust the weights of goods and services in producing the Consumer Price Index; and the National Health Interview Survey, which provides annual estimates of health conditions and use of healthcare services. Data in government surveys may be collected by field staff of an agency (e.g., the Census Bureau conducts the Survey of Income and Program Participation with its field staff); by state or local governments under a cooperative program (the Current Employment Statistics program of the Bureau of Labor Statistics is a monthly survey, based on payroll records of business establishments, conducted by state employment security agencies); by a private organization under a contract or grant from the agency (the University of Michigan conducts the Health and Retirement Survey under a grant from the National Institute on Aging); or by another federal agency (the Census Bureau conducts the Current Population Survey and the Consumer Expenditure Survey for the Bureau of Labor Statistics and the National Health Interview Survey for the National Center for Health Statistics). Many of these surveys have counterparts in other countries: for example, the Labour Force Survey of the Office of National Statistics in the UK and the Survey on Employment of the National Institute for Statistics and Economic Studies in France.

The units of observation in surveys may be individuals, households, families, schools, establishments, or other economic units. Surveys may be cross-sectional to provide an estimate at a point or interval of time, or they may be longitudinal in design in that they follow the same units of observation over time.

Longitudinal surveys can provide information on changes in a unit over time: for example, transitions of families in an out-of-poverty status from the Survey of Income and Program Participation (see *Longitudinal Data; Economic Panel Data*). Some surveys, however, may have a longitudinal component in their design to reduce error or increase efficiency, such as the American Housing Survey.

Statistical agencies also compile information from administrative records. For example, vital statistics are compiled by the National Center for Health Statistics from state agencies, and administrative records are the source of basic statistics for very small firms in the quinquennial economic census, thus reducing or eliminating the need to send these firms census questionnaires.

Statistical agencies may be dependent on the censuses and surveys of other agencies as well as their own data collection in order to produce national indicators. The Bureau of Economic Analysis relies heavily on economic statistics from other sources in compiling the national income and product accounts. And most agencies rely on the Census Bureau to produce the population denominators for reported rates.

Since the US statistical system is so decentralized, important information on a population of interest, such as children or the elderly, is often collected by many agencies and in different ways that may make relating the data difficult, if not impossible. Sometimes, agencies collaborate to fill important information gaps. Two examples are key indicators of well-being for older Americans (Federal Interagency Forum on Aging-related Statistics 2000) and for children (Federal Interagency Forum on Child and Family Statistics 2000).

Whatever the source of data, one of the most important activities of a statistical agency is measuring, controlling, and reporting upon error. Statistical agencies have been pre-eminent in the development of survey designs to reduce sampling error. And many require, as a standard for some forms of publication, that statistical errors be reported. The agencies too are concerned about measurement and other forms of nonsampling errors, and employ various means to study them, including reinterview studies, record checks, multiple indicators, and cognitive studies, which focus on how errors arise in the process of answering questions. Three statistical agencies, the Census Bureau, the Bureau of Labor Statistics, and the National Center for Health Statistics, have cognitive laboratories to evaluate questionnaires and conduct other research to better understand and reduce error (see *Sample Surveys: Cognitive Aspects of Survey Design*).

5. How are Government Statistics Organized?

Government statistical activities may be centralized into a single agency or not. The organization in the

USA, which is described first, is highly decentralized with activities dispersed among many agencies. Some of the trade-offs between centralized and decentralized systems are described at the end of this section.

5.1 The US Statistical System

The highly decentralized statistical system in the USA arose historically because, as new subject matter areas emerged as a national priority, new statistical agencies were created to meet the needs for new data. Little attention was paid to the relevant data-collection activities of other agencies. Efforts at collaboration have met with some successes, such as the interagency forums on aging-related statistics, and on child and family statistics, but efforts to assure systemic collaboration have failed. Attempts to achieve collaboration through centralization, or at least a major consolidation, have run into stakeholders with a special interest in retaining strong, relatively independent agencies within a particular department of government (see Norwood 1995).

Responsibility for coordinating the myriad statistical activities of the US government rests with a small unit in the Office of Management and Budget, the Statistical Policy Branch in the Office of Information and Regulatory Affairs. Authorization of that responsibility derives from legislation principally concerned with paperwork reduction. Despite its size, however, the Statistical Policy Branch has brought statistical agencies together to address many matters of common concern.

The 1995 reauthorization of the Paperwork Reduction Act created an Interagency Council on Statistical Policy, formalizing an arrangement whereby statistical agency heads participate with the Statistical Policy Branch to coordinate their activities. Ten agencies are recognized as having statistical activities as their principal mission: the Bureau of Economic Analysis and the Bureau of the Census in the Department of Commerce; the Bureau of Justice Statistics in the Department of Justice; the Bureau of Labor Statistics in the Department of Labor; the Bureau of Transportation Statistics in the Department of Transportation; the Economic Research Service and the National Agricultural Statistics Service in the Department of Agriculture; the Energy Information Administration in the Department of Energy; the National Center for Education Statistics in the Department of Education; and the National Center for Health Statistics in the Department of Health and Human Services. The heads of these principal statistical agencies, together with the head of the Office of Management and Budget Statistical Policy Branch and the heads of statistical units in the Environmental Protection Agency, the National Science Foundation, the Internal Revenue Service, and the Social Security Administration, are the members of the Interagency Council.

Through this council, its Federal Committee on Statistical Methodology, and other interagency committees it convenes, and its encouragement of the Committee on National Statistics of the National Academy of Sciences–National Research Council, the Statistical Policy Branch has addressed many methodological and other issues that cut across the interests of most statistical agencies. These include issues of poverty measurement, and the definitions of race and ethnicity and of metropolitan areas. It is also concerned with international comparability of statistics, and it represents the USA at the United Nations Statistical Commission and other international organizations. Moreover, the Statistical Policy Branch has striven to extend by regulation and law the protection of confidential data in ways that would allow statistical agencies to share needed information under that protection.

The Statistical Policy Branch is the only unit of government charged to be concerned with the government statistical system as a whole. It conducts an annual review of all major statistical programs for budget directors in the Office of Management and Budget, and publishes a special analysis of funds proposed for statistical activities in the President's budget (US Office of Management and Budget 2000).

5.2 Centralized vs. Decentralized Systems

Most large countries of the world have centralized most or all of their statistical activities into a single agency with broad authority for collecting and compiling government statistics. Coordination of the different activities occurs within the agency itself. Examples are Statistics Canada, the statistical agency of Canada, and the Office of National Statistics in the UK.

A central statistical agency is usually seen to be more efficient in resource allocation; in reaping the benefits of scale; in sharing information across different subject matter areas in order to enhance data and to avoid duplicating activities; in coordinating data activities, such as developing a single longitudinal survey of youth that serves the purposes of both health and education policy makers; and in arranging for the effective use of administrative data. Advantages of a decentralized system are that priorities for budget, relevance, and timeliness, and the urgency to meet needs for new data are set within the department with the most concern for the information and with policies that are most affected by it, and that different perspectives and ways of conducting statistical activities can bring strength to the system as a whole.

6. Practices of Effective Statistical Agencies

The operation of an effective statistical agency begins, first, with a clearly defined and well-accepted mission

and, second, with a strong standing of independence. With these prerequisites, a number of practices contribute to the agency's effectiveness: continual improvement of and development of more useful data, cooperation with data users, openness about its data, commitment to quality and professional standards, wide dissemination of data, advancing the effective use of statistics, analysis and interpretation, an active research program, a highly professional staff, and coordination and cooperation with other statistical agencies.

6.1 Mission

An agency's mission should include responsibility for assessing and meeting needs for information in its purview, determining the sources of the data and the measurement and collection methods, and ensuring that its data are available to the public. An agency mission statement often goes beyond its charge in legislation. In particular, it should articulate the principles that will govern all of its work. The mission should be well accepted externally by the agency's department and others who oversee the agency, and especially by data users. It should also be well accepted internally, so that all agency employees understand how their work furthers the mission.

6.2 Independence

A strong and widely acknowledged standing of independence is necessary for a statistical agency to have the credibility to carry out its mission effectively. In particular, the agency must be independent of other parts of its department engaged in policy making, enforcement, regulatory, or administrative activities. It must be impartial and avoid even the appearance that its data might be manipulated for political purposes, or that information on an individual might be used for enforcement, regulatory, or administrative purposes.

Forms of independence differ among statistical agencies. Agency heads may be required by law to be professionally qualified in the collection and analysis of the relevant statistics. They may be appointed by the highest levels of government for a specific term not coincident with the administration. And they may have direct access to the head of the cabinet department in which the agency is located.

Other characteristics of independence include authority for the following: to decide the scope, content, frequency, and manner of data collected, compiled, or reported; to select and promote professional staff; to release statistical information without prior clearance and with a clear separation from policy interpretations by others; to represent the agency's statistical program before legislatures and the public; and to establish and follow predetermined schedules for the release of important economic and other indicators.

6.3 *Continually Developing More Useful Data*

In order to make their data more useful and to meet changing public policy needs, statistical agencies must continually look to develop data systems that can serve broader purposes. Integrating data from surveys conducted by different agencies or integrating data from surveys with data from administrative records, such as tax or social welfare program records, are ways to achieve this goal. When separate data sets are collected and analyzed in such a way that they may be used together, the value of each may be greatly enhanced. Data sharing and statistical uses of administrative records, with appropriate safeguards for confidentiality, also make a statistical agency more efficient as well as more effective.

6.4 *Cooperation with Data Users*

An effective statistical agency consults with a broad range of users of its data in order to make its products more useful. It seeks advice on data concepts, methods, and products in a variety of formal and informal ways, from data users as well as subject-matter and other experts. It endeavors to meet the needs for access to data while safeguarding the confidentiality of individual responses. And it strives to provide equal access to all potential users.

6.5 *Fair Treatment of Data Providers*

To maintain a relationship of mutual respect and trust with data subjects and other data providers, an effective statistical agency observes fair information practices, which include: protecting the confidentiality of data identified with an individual, whether the data are collected directly or obtained from administrative record sources; informing respondents of the purposes of data collection and the expected uses of the information; informing those asked to report whether their participation is voluntary or mandatory, and using appropriate informed consent procedures for voluntary data collections; minimizing the contribution of time and effort asked of respondents, which may be achieved in seeking the advice of respondents in the planning and design of data collection; and respecting the willingness of respondents to contribute to society by representing accurately the information they provide, and by making it available for important uses.

6.6 *Openness*

An effective statistical agency operates in a completely open manner. Openness requires providing full documentation of the data with acknowledgment of any uncertainty, a description of the methods used and assumptions made, and measures or other descriptions of error, for both sampling and nonsampling error.

When a mistake occurs, the agency should acknowledge it promptly and publicly. For a correction to an important indicator, the agency should strive to inform as many current and potential users as possible.

6.7 *Quality and Professional Standards*

An effective statistical agency is committed to the highest standards of quality work and professional practice and ethics, in statistics and other relevant professions. For this reason, a strong and continuous relationship with appropriate professional organizations in statistics and relevant subject-matter areas is important.

6.8 *Dissemination of Data*

An effective statistical agency seeks the widest possible dissemination of its data through a variety of means, including public-use data files and the agency's website, and has policies to guide what data to archive and how.

6.9 *Advancing Effective Use of Statistics*

Statistical agencies should also advance the uses of their data through making them easily accessible to users in convenient forms, involving data users in planning data collection and dissemination, and providing full documentation and other forms of technical support to users.

6.10 *Analysis and Interpretation*

Agencies can make their data more relevant to public policy concerns by analyzing and interpreting the data. Care must be exercised, however, that analyses not be perceived as policy advocacy.

6.11 *Research*

An active research program is integral to the activities of an effective statistical agency. Research may be conducted on the substantive issues for which the data were compiled, especially concepts and definitions; methods of measurement, collection, analysis, and dissemination; and topics to improve the utility of the agency's data in policy analysis and decision making.

6.12 *Professional Staff*

A statistical agency must recruit and retain a professional staff of high caliber in statistics and fields relevant to its mission. Opportunities should be provided for staff to maintain and advance their capabilities, such as through appropriate activities of professional organizations.

6.13 Coordination and Cooperation

In the US context, an effective statistical agency continually seeks opportunities to cooperate with other statistical agencies to enhance the value of its own and other agencies' information. Statistical agencies cooperate, not only with each other, but also with state and local statistical agencies in the provision of data for subnational areas. They cooperate, too, with foreign and international statistical agencies to exchange information, on both data and methods, and to develop common classifications and procedures to promote international comparability of information.

Integrating data collected or compiled by different statistical agencies invariably requires much coordination and cooperation among the agencies, each of which have different missions. But the rewards are data more relevant to policy concerns, and a stronger statistical system as a whole. For these reasons, statistical agencies must cooperate as partners with one another, not only on the development of data, but also on the entire panoply of statistical activities, including definitions, concepts, measurement methods, analytical tools, professional standards, dissemination modes, ways to advance the effective use of their data, and means to protect the confidentiality of responses.

7. Challenges Faced by Statistical Agencies

We live in a rapidly changing, increasingly interdependent world with a global economy. Many countries face profound demographic changes. Primarily fueled by technological advances, widespread changes are occurring in society, the environment, and international relationships. Governments struggle to keep up with the changes. They are concerned with the need for new property rights, laws governing new markets, new forms of businesses, and investments in new technology for education. They are concerned too about mitigating human error in transportation, communications, healthcare, and other complex, critical systems; avoiding harmful effects on the environment and sources of food; and guarding against the loss of individual privacy. And governments are also concerned with disparities in economic growth and the distribution of income and wealth, to which disparities in access to technology may contribute, since these disparities could increase tensions among social groups and nations.

An efficient, flexible, and modern information system is essential for a government to address these and other concerns. The greatest challenge facing any statistical agency is for it to keep up with the times. 'Public policy analysis often cannot wait until new surveys are undertaken, but must rely on knowledge drawn from the existing statistical database' (Cook 2000). Those in charge of government statistics must anticipate and prepare for future needs.

Technology, which has induced so many rapid changes, itself can give statistical agencies the means to respond to those changes. The agencies must not only incorporate new technology into collecting, processing, analyzing, and disseminating government statistics, but they must also prepare for the plethora of new data in digital form.

Advances in information technology are also making it more and more difficult to protect confidentiality. If a statistical agency is to respect its data providers by putting their information to use in the best way to benefit society, stronger ways to protect confidentiality must be sought. Statistical agencies should lead in confidentiality research, including research on data encryption and altering or reconstituting data, and in fostering legal and administrative means to protect confidentiality.

8. Conclusion

Statistics are essential for a government to function effectively and to advance businesses and the well-being of its citizens. The nature of data as a public good provides an economic argument for a government to collect statistics, but there are many other reasons for it to do so. One is to assure accurate, timely, and credible information that its citizens may use to hold their government accountable on any issue. A statistical agency must be able to provide information relevant to current and future policy issues, and to establish credibility for itself and its data. Finally, the agency must be respected and trusted by respondents who provide it with data and with all data subjects whose information it obtains. That respect and trust can be fostered by protecting the confidentiality of individually identifiable information and by putting the information it obtains to use in the best ways to benefit society, including research.

More information on the principles and practices of statistical agencies can be found in Martin et al. (2000), and on the organization of US statistical agencies in Norwood (1995). Statistics for over 200 countries and areas of the world are compiled by the United Nations Statistics Division (2000), which provides links for the national statistics of countries online at www.un.org/Depts/unsd/ through a link to 'Global Statistics.' Links to the websites of statistical agencies in the USA are available at www.fedstats.gov.

See also: Databases, Core: Demography and Registers

Bibliography

Cook L 2000 Statement by the first National Statistician on the launch of National Statistics on 7 June, www.statistics.gov.uk/nsbase/about_ns/About.asp

- Duncan G T, Jabine T B, De Wolf V A (eds.) 1993 *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. National Academy Press, Washington, DC
- Edmonston B, Schultze C (eds.) 1995 *Modernizing the U.S. Census*. National Academy Press, Washington, DC
- Federal Interagency Forum on Aging-related Statistics 2000 *Older Americans 2000: Key Indicators of Well-being, 2000*. US Government Printing Office, Washington, DC, <http://www.agingstats.gov>
- Federal Interagency Forum on Child and Family Statistics 2000 *America's Children: Key National Indicators of Well-being, 2000*. US Government Printing Office, Washington, DC
- Mackie C, Bradburn N (eds.) 2000 *Improving Access to and Confidentiality of Research Data: Report of a Workshop*. National Academy Press, Washington, DC
- Martin M E, Straf M L, Citro C F 2001 *Principles and Practices for a Federal Statistical Agency: Second Edition*. National Academy Press, Washington, DC
- Norwood J L 1995 *Organizing to Count: Change in the Federal Statistical System*. The Urban Institute Press, Washington, DC
- United Nations Economic and Social Council 1994 Fundamental principles of official statistics of the Statistical Commission of the United Nations. Report of the Statistical Commission on its Special Session, New York, 11–15 April 1994. *Official Records of the Economic and Social Council*, suppl. 9
- United Nations Statistics Division 2000 *Statistical Yearbook Forty-third Issue*. United Nations, New York
- US Office of Management and Budget 1997 Notice of order providing for the confidentiality of statistical information and extending the coverage of energy statistical programs under the federal statistical confidentiality order. *Federal Register* 62: 35044–50
- US Office of Management and Budget 2000 *Statistical Programs of the United States Government, Fiscal Year 2001*. US Government Printing Office, Washington, DC

M. Straf

Governments

Governments are universal. They exist in every social organization: there is always a body, normally relatively small, which has the task of running that organization; but the most visible governments are those of nations and of other public institutions. With the decolonization process, independent national governments have come to rule practically the whole of the planet and they have more than doubled in number, from well under a hundred in the early 1950s to nearly two hundred at the beginning of the third millennium.

1. Problems of Definition

Being at the center of political life, national governments tend to be rather small and, often at least, compact. Yet it is not possible to circumscribe them

neatly as their boundaries are not precise. In many cases, indeed most, an official list exists, but this list is typically not exhaustive, as key personalities, who may formally only be ‘advisers,’ but are in fact highly influential, tend not to be included. There are also often two or more ‘levels’ of the government: thus in the UK and in countries which have followed closely the British model, for instance, the ‘cabinet’ constitutes only an inner core, with ‘ministers of state’ and ‘parliamentary secretaries’ being below this core. In the USA, the government is typically referred to as the ‘administration’ or as the ‘executive’: however, both of these expressions remain imprecise with respect to those whom they cover.

It is therefore more appropriate to look for a definition in terms of the functions which national governments fulfill. In general, these are said to ‘run the affairs of the nation,’ although they only do this up to a point since they are helped and advised by a multiplicity of individuals and of groups, and, in modern democratic countries in particular, by parties and legislatures, while everywhere the bureaucracy is at their ‘disposal.’ One can in reality distinguish among three functions of governments. First, they have to elaborate realistic policies which can be implemented: this is the function of conception. Second, they have a function of implementation, as they must find the means of ensuring that the policies which have been elaborated are indeed turned into practice. There can be tensions between these two functions. The skills required to undertake each of them successfully are different: to conceive means to ‘imagine’; to implement means to ‘manage.’ Yet governments have also a third function, that of coordination: they must ensure that the policies which are elaborated and implemented do not pull in different directions but are complementary.

2. Types of Government, Liberal-Democratic and Authoritarian

A ‘revolution’ in the characteristics of governments has occurred since the eighteenth century, when ‘modern’ government began. This ‘revolution’ has had two main aspects. First, there has been a massive increase in the complexity of governmental activities, as states have come to be increasingly involved in the economic and social life of countries, especially from the late nineteenth century to the 1980s. Second, many governments have come to be (or have had to be) responsive to the demands of citizens, in particular where the representative system had been introduced: for the first time in human history, these governments had to abide by rules based on liberal democratic principles and enshrined in constitutions. Admittedly, Greek cities, the Roman Republic, and, during the Renaissance, some parts of Italy, had experienced forms of ‘limited’ government; but only from the end of the seventeenth century did representative govern-

ment take root in a nation, England, which hitherto had been a traditional monarchy, while, almost a century later, American independence marked the emergence of the first liberal Republic.

England and America were indeed the originators of the two main forms of constitutional government which have existed to this day, the cabinet system, which was gradually adopted in almost every Western European country and in the Commonwealth, and the presidential system, which spread to Latin America, though with only limited success. Thus, at the beginning of the third millennium, approximately half the national governments could be described as being broadly responsive to their people. Meanwhile, in the rest of the world, including in many countries of Africa, of Asia and, at least in the past, especially up to the mid-1980s, of Latin America, governments have tended to be authoritarian. The Communist form was particularly harsh, both in what was the Soviet Union and in Eastern Europe up to the late 1980s, as well as in China and a number of other Asian countries: it was based on the dominance of a single party whose leaders formed the real government and controlled the formal executive. Independence from colonial rule which started in the 1950s and gathered momentum in the 1960s, especially in Africa, also led to many dictatorships, sometimes based on single-party arrangements but relying even more on the 'charismatic' figure of a president, often drawn from the military and having come to power through a coup. Admittedly, since the 1980s, a move towards responsive governments occurred in many countries hitherto ruled by dictatorships, in Latin America, in what had been the Communist block, in East and Southeast Asia and in parts of Africa; but authoritarian rule continued to characterize many countries of the contemporary world, sometimes on the basis of old-fashioned monarchical governments but more often on 'modern' types of dictatorships, in many cases of the military type.

3. The Organization of Governments: Cabinet and Presidential Systems. The Role of Leaders

Partly as a result of these differences, governments vary widely in organization, duration, composition, and powers. The distinction between the cabinet system and the presidential system remains the main one among liberal democratic governments, although there are variations within each of these two models and some cases are intermediate. The cabinet system is based on the division between a symbolic head of state, often a monarch, and a government, responsible to parliament, headed by a prime minister: the prime minister has to operate in the context of a collegial system in which those members of the executive who are part of the cabinet (normally referred to as 'ministers') are entitled to participate fully and col-

lectively in decision-making: in practice, this is far from being always the case and there are substantial differences from country to country and within each country. Some cabinets are controlled closely by their prime minister, while others are more egalitarian. By and large, the British government and the governments of many Commonwealth countries tend to be tilted toward prime ministerial dominance, while a number of continental European governments, at least those which are based on complex party coalitions, give more elbow room to ministers. Yet, even though cabinet members may be more or less collectively involved, there is a real sense in which the government constitutes a team. A variation on this model is provided by the French system, in which the picture is somewhat complicated by the fact that the president, elected directly by the people, plays a significant part in policy making, at least when the party of the president also dominates parliament: this model has been adopted by some of the post-Communist governments of Eastern Europe as well as in the developing world.

The presidential system, on the other hand, is based on the separation of powers principle, with both president and legislature elected directly by the people. The president, as head of the executive, has thus greater leeway to organize the government in the way he/she wishes: this is particularly true in the USA. In some Latin American countries, on the other hand, the president has to an extent to take into account the desire of party leaders to play a part in the appointment of members of the executive. However, by and large, presidential governments are both constituted more independently from the parties and in particular from the parties in the legislature. They are also more hierarchical in character: the members of the government (typically known as 'secretaries') are responsible to the president for the department of which they are in charge (although in some cases the legislature can also censure them individually). They are not collectively involved in decision-making.

Governments are fashioned by their leaders. The twentieth century was characterized by the emergence of 'charismatic' rulers, including and perhaps above all in Communist countries, despite the fact that the Communist ideology is based on the notion that the economic 'substructure' is held to be paramount. 'Charismatic' or personalized leadership can also be found in liberal regimes, whether of the cabinet or of the presidential type. This has been the case not just of wartime leaders, such as Churchill or de Gaulle, but of rulers who attracted the imagination of whole populations, such as Kennedy.

4. The Duration of Governments

The duration of governments varies. The longest are those of the few remaining traditional regimes, in effect absolute monarchies, and those of some of the

more 'modern' dictatorships: Communist governments, their leaders in particular, often remained in power for over a decade; on the other hand, most other leaders are only in office for a few years. The other government members often stay for even shorter periods: the average duration of ministers or secretaries is in the region of three to four years, with a substantial proportion of them being truly transient and remaining in government for a year or less.

5. Career Patterns

The composition of governments also varies. Some of these, primarily where there is a cabinet system, include almost exclusively professional politicians, men and women who have been active in a political party, have then been elected to parliament and have joined the ranks of the executive after a few years on the 'backbenches.' Other government members are 'technicians,' more often drawn from the public sector than from the private sector: their career in politics tends to last a few years only and they return to other public or private jobs after their stint in government. Moves of this kind occur less frequently in cabinet systems than in presidential systems, where, as we noted, the head of the state has more freedom to pick and choose whom he or she wants to appoint; they also occur in many authoritarian governments, though, in these, military men are often side by side with civilian 'managers.'

Variations in career patterns are associated with differences in social background. In general, except where social democratic parties are frequently in power, as well as in the past Communist systems, men and women with a working class background are rare: lawyers, teachers, and managers form the bulk of the civilian members of governments. Only in the last decades of the twentieth century and in some countries, mainly Northern European, have women had more than a token share of governmental posts, especially in such key departments as those of foreign, economic, or internal affairs.

6. The Impact of Governments: An Unresolved Question

There is debate about the extent to which leaders and governments affect the countries which they rule. As their careers are often short, it is suggested that members of the higher civil service, together with the lobbies which gravitate around them, play a key part in decision making. As a matter of fact, it has not been possible so far to determine precisely the impact of leaders and members of governments even when they stayed in power many years: environmental factors, both internal and external to the country concerned, must be taken into account. Yet some leaders at least,

whether 'harsh,' such as Lenin or Stalin, or 'good,' such as Roosevelt or Churchill, had a major influence on the destinies of the countries which they ruled and indeed of the world. What can only be said is that there are massive variations in impact, as there are major variations in the characteristics of governments and in particular in their responsiveness, organization, duration, composition, and complexity.

See also: Authoritarianism; Authority; Delegation; Leadership: Political; Liberalism; Liberalism: Historical Aspects; Parliamentary Government; Parliaments, History of; Power: Political; Presidential Government

Bibliography

- Barber J D 1977 *The Presidential Character*, 2nd edn. Prentice-Hall, Englewood Cliffs, NJ
- Bass E M 1981 *Stogdill's Handbook on Leadership*. Free Press, New York
- Blondel J 1982 *The Organisation of Governments*. Sage, London
- Blondel J 1987 *Political Leadership*. Sage, London
- Blondel J, Muller-Rommel F (eds.) 1993 *Governing Together*. Macmillan, London, UK
- Bunce V 1981 *Do New Leaders Make a Difference*. Princeton University Press, Princeton, NJ
- Burns J M 1978 *Leadership*. Harper & Row, New York
- Cammack P, Pool D, Tordoff W 1988 *Third World Politics*. Macmillan, London
- Castles F G (ed.) 1982 *The Impact of Parties*. Sage, London
- Castles F G, Wildenmann R (eds.) 1986 *Visions and Realities of Party Government*. De Gruyter, Berlin
- Cronin T E 1975 *The State of the Presidency*. Little, Brown, Boston
- Finer S E 1962 *The Man on Horseback*. Pall Mall, London
- Hecl H 1977 *A Government of Strangers*. Brookings Institution, Washington, DC
- Hook S 1955 *The Hero in History*. Beacon Press, Boston
- Kellerman B (ed.) 1984 *Leadership*. Prentice-Hall, Englewood Cliffs, NJ
- Neustadt R E 1960 *Presidential Power*. Wiley, New York
- Rejai M, Phillips K 1983 *World Revolutionary Leader*. Sage, London
- Shugart M S, Carey G M 1992 *Presidents and Assemblies*. Cambridge University Press, New York
- Strom K 1990 *Minority Government and Majority Rule*. Cambridge University Press, New York
- Weber M 1968 *Economy and Society*. Bedminster Press, New York, 3 Vols

J. Blondel

Grammar: Functional Approaches

The term *functional* or *functionalist* has a meaning in linguistics distinct from the use of the term found in anthropology or philosophy. In linguistics, a func-

tional analysis is an analysis of linguistic structure that puts the primary emphasis on explanation in terms of factors outside the structural form, such as meaning or the function of language in discourse. Functional analysis is in contrast to *formal* analysis, which is an analysis of linguistic structure that puts the primary emphasis of explanation on principles defined purely in terms of structural form. The term *functionalist* has been applied to any school of grammatical analysis which argues for the primacy or centrality of functional explanations over formal explanations.

1. Introduction

Such a broad characterization of functionalism in linguistics, of course, requires further elaboration. We begin by illustrating the commonest types of functional explanations, and then turn to different schools of linguistic theory that call themselves functionalist.

2. Types of Functional Explanations

There are three types of functional explanations that usually are distinguished in the functionalist linguistic literature: processing, cognitive, and discourse or information structure explanations.

2.1 Processing Explanations

Processing explanations make reference to models of the production and comprehension of utterances. A processing explanation argues that some aspect of grammatical structure is due to limitations on the human ability to produce or comprehend the structure.

Processing explanations typically are employed to account for two sorts of linguistic facts. One is the relative frequency of a particular grammatical structure across the world's languages. For example, it has been observed that suffixes are on the whole more frequent than prefixes (or more complex alterations of word forms) in the world's languages. A processing explanation has been proposed for the suffixing preference. It is hypothesized that speakers process word stems before their inflections. Psycholinguistic experiments imply that the beginning of a word is most salient, hence the word stem should ideally occupy the initial position; the ends of words are the next most salient, hence inflections should ideally occupy the final position.

The second sort of phenomenon for which processing explanations are employed is to account for degrees of acceptability of a construction in a single language. For example, in English, subordinate clauses embedded in the middle of a clause are relatively less acceptable than subordinate clauses

embedded at the end of a clause: compare *I know the boy that owned the cat that caught your parakeet* to *The parakeet that the cat that the boy I know caught died*. This fact has been explained as a consequence of the difficulty in interpreting the latter sentence's center-embedded clauses. In fact, it is grammatical in English to move a relative clause from the middle of a clause to the end under certain circumstances: *A woman contacted me who I had shared an apartment with during my last year at college*.

Processing explanations have been proposed for constraints on certain types of syntactic structures, constraints on patterns of morphemes within words, and on sound patterns. The latter explanations typically are described in terms of ease of articulation or ease in discrimination (in speech perception). Processing explanations for syntax and morphology, especially in explaining differences in cross-linguistic frequency of word order and morpheme order, are particularly associated with the work of J. A. Hawkins (e.g. Hawkins 1994).

2.2 Cognitive Explanations

Cognitive explanations generally make reference to the conventional meaning of a word, inflection or syntactic construction (e.g. Bybee 1985, Tomasello 1998). A cognitive (semantic) explanation of a grammatical phenomenon essentially hypothesizes that some or all aspects of the analysis of the grammatical phenomenon make reference to conventional semantic properties. For example, some languages have plural inflections for all count nouns, while other languages restrict plural inflections to only nouns referring to animate beings, still others to only nouns referring to humans. This variation is built on a general ranking or hierarchy of nouns according to the semantic class of their referents: human < animate < inanimate. The description of the constraint on the occurrence of the plural inflection in each language, and the cross-linguistic pattern of constraints, requires reference to the semantic classes of entities denoted by the noun, described in terms of animacy.

Cognitive explanations, like processing explanations have been used to describe both grammatical constructions in particular languages and also patterns of cross-linguistic variation. The example of the animacy hierarchy in the preceding paragraph illustrates both sorts of explanations. It hypothesizes that the distribution of the plural inflection in any particular language will make reference to the semantic class of the entity denoted by the noun. It also hypothesizes that the cross-linguistic variation in the distribution of the plural inflection is subject to the hierarchical ranking of animacy categories given in the preceding paragraph. In recent cross-linguistic research, explanations of cross-linguistic patterns have been offered in which the patterns of occurrence of a

construction are mapped onto a single semantic or conceptual space applicable to all languages.

Some cognitively oriented functionalists make a stronger claim about the relationship between grammatical form and semantic function, namely that certain grammatical categories are better described in semantic terms. This hypothesis will be discussed below on the contribution of cognitive linguistics to functionalism.

2.3 Discourse (Information Structure) Explanations

Discourse explanations make reference to the organization and presentation of information in discourse, that is, the structuring of information for the purpose of communication. Such explanations are called discourse explanations by some functionalist, and information structure explanations by other functionalists. For example, the English active and passive voice constructions (*The car hit the tree* vs. *The tree was hit by the car*) both describe the same event and participants, but differ in which participant is encoded as the subject of the construction. It has been argued (based on naturally occurring examples) that the participant that is more topical is encoded as subject. Another example is the choice of construction in response to a question such as *Wasn't that incredible when Mary called the boss a pig?* One can respond with *Yeah, it really shocked me that she called him that*, but not with **Yeah, what really shocked me was that she called him that*. The unacceptable response uses the pseudocleft construction, and the pseudocleft construction is acceptable only if the information 'something really shocked me' is already in the hearer's consciousness. It is not in this case, and hence is unacceptable.

As with processing and cognitive explanations, discourse functional explanations have been offered both to explain restrictions on the acceptability of specific constructions in particular languages, and to explain the frequency of distribution of particular linguistic features across languages. Examples of the former were given in the preceding paragraph. An example of the latter is certain patterns of word order. In the overwhelming number of languages, the subject precedes the object in the most common order of elements in a clause. It has been proposed that the reason for this asymmetry is that the subject tends to be the topic, and the most common discourse pattern is for the topic to precede the 'comment,' or at least other less topical participants.

3. Schools of Functionalist Analysis

One of the most fundamental issues that divides schools of linguistic theory is: what is the relationship between linguistic form and linguistic function? The

relationship is not at all straightforward (Croft 1995). Also, depending on the position that a linguist takes on the relationship between form and function, certain additional hypotheses have to be made in order to accommodate linguistic facts. These hypotheses have to do with aspects of language not directly connected to the form-function relationship, such as the organization of grammatical knowledge in a speaker's mind, the nature of semantics, and the role of cross-linguistic evidence. Since these additional hypotheses are closely associated with various schools of functionalist analysis, they will be discussed here.

3.1 Functionalism and Formalism

Functionalist schools generally set themselves up as contrasting with the formalist approach, associated most strongly with N. Chomsky, but also advocated by other linguists who reject some of Chomsky's more specific claims about syntax. A purely formal explanation of a grammatical phenomenon makes reference exclusively to properties of grammatical form or structure. With respect to grammatical structure, a formal explanation of some pattern of sentence structure will make reference only to syntactic categories and structural relationships between the elements of a sentence, and not to the meaning of the words, inflections or constructions of that sentence, their discourse function, or their production and comprehension in language use. With respect to word structure, a formal explanation will make reference only to morphological categories and relations between roots and affixes, and not to aspects of the meaning and use of word forms. With respect to sound structure, a formal explanation will make reference only to the structure of the sound system independent of its articulatory and auditory realization.

This is not to say that formalist analyses have no place for the meaning and use of linguistic constructions. The formalist hypothesis is that there is a significant number of properties of grammatical form which can be explained in a purely formal way. The concept of a formal explanation is closely associated with the theoretical position that the organization of grammatical knowledge in a speaker's mind is divided into components which separate a linguistic form from its conventional meaning and use in discourse. A typical model describes grammatical knowledge as being compartmentalized into phonological, syntactic and semantic components. The phonological component is separated from its phonetic realization, and the syntactic component is separated from its semantic interpretation. Linguistic form is linked up to its phonetic realization or semantic interpretation, eventually: formal models have rules linking syntax to semantics and phonology to phonetics. The formalist hypothesis does however imply that linguistic form is organized in a purely formal way to a high enough

degree that it can be described best as existing in separate components. That is, the formal analysis of grammatical structure is self-contained relative to functional factors.

This is also not to say that formalists reject all explanations that make reference to functional factors. Formalism accepts that some, perhaps many, linguistic phenomena require reference to functional factors. However, formalists argue that there is a large body of linguistic phenomena, those which most linguists would consider to be central to grammatical structure, that can be analyzed without reference to any functional factors. Another way of describing this position is that the phonological and syntactic components represent a substantial core of a speaker's grammatical knowledge of a language.

Functionalism approaches to language challenge or reject the formalist model of the relationship between form and function to varying degrees. However, all functionalists accept that not all aspects of grammatical structure can be accounted for by meaning and use. There will always be some degree of arbitrariness in language structure, where arbitrariness is intended to mean not wholly explainable in terms of language function. The question is, how much arbitrariness is there in language structure, and to what extent does the arbitrary aspect of language structure form a self-contained system? Different functionalist schools give different answers to this question.

3.2 Autonomous Functionalism

The closest approach to formalism which calls itself functionalist can be called *autonomous functionalism*. Autonomous functionalism argues that the number of properties of grammatical form that are explainable solely in grammatical terms are considerably fewer than most formalist theories assume. In particular, the restrictions on the use of particular constructions are sought in functional properties, whether cognitive, processing or discourse-functional. To take a simplified example in the *There*-construction (as in *There is a bird in the bushes*), a noun phrase with an indefinite article (*a bird*) is acceptable but one with a definite article usually is not (compare) **There is the bird in the bushes*). A plausible explanation for this restriction is that the *There*-construction is used to introduce a new referent into the discourse, but definite noun phrases describe referents already known to both speaker and hearer. This would be an autonomous functionalist analysis of the restriction.

However, even in an autonomous functionalist approach, the rules governing the formation of the construction itself, such as the fact that *There* is the first element followed by a form of *be*, are considered to be purely formal. Autonomist functionalists focus their research efforts on accounting for as many of the

constraints on the construction as possible in terms of functional properties, while leaving the basic structure of the construction for formalist analysis. The autonomist functionalist research program is found in the work of Kuno (1987), Prince (1988), Ward, and their students and associates. Their explanations generally appeal to properties of information structure rather than semantic properties. Because they accept the basic structural analyses of formalist approaches, autonomous functionalists generally do not reject the view that linguistic knowledge in a speaker's mind is organized into components. Instead, they have proposed that there is another component, the information structure component, which should be included along with the syntactic and semantic components.

In fact, many functional analyses of specific constructions in more radically functionalist approaches to language are actually autonomous functionalist analyses, that is, they account for the restrictions on syntactic constructions in terms of the meaning and use in discourse of the construction. We may call these autonomous functional analyses in whatever sort of functionalist approach they are found in, that is, autonomist functionalist analyses of particular constructions are not incompatible with more radical functionalist approaches.

3.3 Mixed Formalism/Functionalism

Another sort of approach, not clearly identifiable with any particular school of linguistic theory, but sometimes called functional, challenges the hypothesis of a sharp separation between the use of formal or functional properties to describe grammatical constructions. In these theories, the description of both the structure and the constraints on a construction can mix formal and functional properties. Two schools, *Functional Grammar* of S. Dik and co-workers, and *Role and Reference Grammar* of R. Van Valin and co-workers, both use formal and functional properties (typically semantics, but also information structure) to describe the grammatical structure of sentences as well as the constraints on them. In addition, many grammatical descriptions of languages informally describe both syntactic and semantic properties of constructions without implying any sharp division between the two.

3.4 Typological Functionalism

As was noted above, cognitive, processing and discourse explanations have been offered for patterns of cross-linguistic variation as well as in the analysis of the grammars of individual languages. A school of linguistic theory known as typology has arisen which

focuses on cross-linguistic variation and its limits (see, for instance, Greenberg 1990, Comrie 1989, Givón 1979, Hawkins 1988, Croft 1990). Much cross-linguistic variation appears to be built on the base of general functional patterns, as illustrated above for cognitive, processing, and discourse explanations. This observation has led some typologists to argue that all cross-linguistically verified universals of language are susceptible to functional explanation in this way. This approach to analysis is called typological functionalism or the functional-typological approach.

Typological functionalism makes a stronger claim than the schools described in the preceding sections. The autonomous functionalist approach assumes that there still remains a significant amount of syntactic structure that can be described in terms of a self-contained set of rules (which are further assumed to be in part language-universal). The mixed formal-functional theories do not assume that formal principles governing syntactic structure are self-contained, but those theories allow for the possibility that there are formal properties of languages that are universal for all languages. The typological functionalist approach hypothesizes that any universal generalizations about linguistic structure are ultimately explained in functional terms.

The typological functionalist approach does not, in fact cannot, assume that the structure of a particular language's grammar is entirely explainable in functional terms. For instance, one example given above was that the occurrence of a plural inflection on nouns could be accounted for by the hierarchy human < animate < inanimate. The semantic hierarchy is universal, but each language applies the hierarchy to the distribution of the plural inflection in its own way, that is, one cannot predict what semantic class a language will extend (or restrict) its plural inflection to. That is, the exact class of nouns allowing a plural inflection is partly arbitrary, although it is partly explainable in terms of the semantic hierarchy: if the language allows plural inflection for animate nouns, it will allow plural inflection for human nouns. This sort of explanation of a linguistic phenomenon is sometimes described as motivation, in contrast to a set of rules that exhaustively predicts the forms that would be found.

A typological functionalist analysis must allow for some degree of arbitrariness in accounting for cross-linguistic patterns. If there were no arbitrariness, all languages would be alike in their grammatical structure; and they are not. But the typological functionalist approach hypothesizes that the arbitrary aspects of language structure are language-specific, and that any universal generalizations about linguistic structure are ultimately explained in functional terms.

An important element of typological functional explanations is the hypothesis of iconicity or iconic motivation, in which grammatical structure is hypoth-

esized to reflect semantic or conceptual structure to some degree. For example, the closer grammatical relationship between verb and object as opposed to verb and subject in many languages is attributed to the closer semantic relationship between states of affairs and the participants usually encoded as the object. Iconic motivation can be used to account for grammatical structure itself, not just constraints on grammatical structures. In this respect, typological functionalism represents a more radical functionalist approach than autonomous functionalism.

3.5 Cognitive Linguistics and Functionalism

Cognitive linguistics (e.g. Lakoff 1987, Langacker 1987, Talmy 2000) is a broad school of linguistic theory that is usually called functionalist. Cognitive linguistics, as its name indicates, focuses on cognitive (semantic) explanations for grammatical structure. Cognitive linguistics has contributed a number of ideas to other functionalist approaches that are relevant to assessing functional approaches.

Cognitive linguistics has proposed a model of grammatical knowledge, construction grammar, that offers an alternative to the formalist model in which form and function (semantics and information structure) are separated in distinct components. Construction grammar hypothesizes that knowledge of syntax involves the knowledge of individual constructions which combine formal properties (syntactic structures, morphological inflections) and functional ones (semantics and discourse function). For example, the ditransitive construction [X Verb Z Y], as in *Marilyn sent Gina a book*, specifies not only the syntactic structure of the construction but also the semantic relationships among the participants, such as the fact that Z comes to possess Y. Constructions are organized in a network of grammatical knowledge in a speaker's mind. More broadly, grammatical knowledge is organized as a system of signs or symbols, consisting of form (the signifier) and function (the signified). A model organizing grammatical knowledge in terms of constructions/symbols/signs allows for a direct statement of functional properties that are relevant to particular grammatical constructions and other grammatical units.

Cognitive linguistics argues that semantics involves conceptualization or construal of an experience by a speaker for the purposes of linguistic communication. For example, an English count noun can be used in a mass noun grammatical context, as in *There was a huge Buick there; just acres of car* (attested example). The noun *car*, normally denoting an individuable object, is construed to denote an unbounded quality (size). These and other examples indicate that the English count noun-mass noun distinction has at least some semantic basis. Cognitive linguists also accept that some alternative construals are conventionalized,

such as *chicken* (meat, mass noun) vs. *chicken* (live bird; count noun). Nevertheless, the conventionalized grammatical distinction conforms to the semantics of count noun vs. mass noun: a live bird is an individuable entity, while meat is a unbounded substance.

Conventionalized construals are of course partly arbitrary (see below); but the semantic regularity of both conventional and novel construals has led cognitive linguists to argue that even grammatical categories such as 'noun' and 'subject' are susceptible to a cognitive semantic explanation. That is, a grammatical category such as 'noun' encodes a construal of an entity in a certain way, which can be discerned by the occurrence of particular word roots directly as nouns, and in the semantic changes that occur when a non-noun root is used as a noun, or can be morphologically derived as a noun.

Finally, cognitive linguistics has developed a model of categorization which has in turn been exploited by functionalist linguists. A classical view of meaning is that the meaning of a word, inflection or grammatical category must be a single definition that will cover all and only the appropriate uses of the word or category. Cognitive linguists have argued that many word and category meanings can be described in terms of polysemy, that is, distinct but semantically related meanings. For example, the meaning of *over* in *The bird flew over the hill* (path of motion) is distinct from its meaning in *The house is over the hill* (location), but they are related systematically. More specifically, cognitive linguists (drawing on research by psychologists) have suggested that many polysemous categories have one meaning that can be considered central or basic, called the category's prototype. Cognitive and other functionally oriented linguists have argued that many if not all grammatical categories are in fact polysemous, and many of these categories have a semantic/discourse-functional prototype which is found to be valid across languages (see *Linguistics: Prototype Theory*).

3.6 More Radically Functionalist Schools

Grammatical structure is commonly assumed to exist in speakers' minds. However, grammatical structure is also directly involved in social interaction in language use, and language use is central to accounting for language acquisition, language variation and language change. In the more dynamic process of language acquisition and language change, functional factors have been argued to play a role. In both language acquisition and language change, it has been argued that competing motivations among functional principles play a major role.

The concept of competing motivations is that (functional) principles may come into conflict such that there is no grammatical system that satisfies all of

the functional principles. As a result, change occurs over time in acquisition and in the history of a language, and the languages of the world exhibit structural diversity even though their speakers' linguistic behavior conforms to the same functional principles. (Competing motivation models do not presuppose that the competing principles are functional, and in fact competing motivation models are now used by many formalists; see below).

The most general and commonly offered example of competing motivations is that between economy and iconicity (see *Linguistics: Iconicity*). *Economy* is the principle that a speaker uses the least effort necessary to express himself or herself. For example, English leaves the singular number of nouns unexpressed: *book-φ* vs. *book-s*. Economy is considered to be a speaker-oriented functional principle: its influence on language is for the benefit of the speaker. *Iconicity* is, in part, the principle that all of the relevant parts of the meaning conveyed are in fact conveyed by grammatical elements in the utterance (words, inflections, etc.). For example, in some languages, both the singular and the plural of a noun are expressed by overt suffixes. This aspect of iconicity is considered to be hearer-oriented: any aspect of meaning left out by the speaker may not be recoverable by the hearer. Economy and iconicity compete with each other: a linguistic expression that is economical will not be iconic (since it leaves some elements of meaning unexpressed), and an expression that is iconic will not be economical (since certain elements of meaning are not left unexpressed). Hence, across languages, there is diversity in the expression of the category of number, and languages change from one form of expression of number to another over time.

Perhaps the best known competing motivations model in language acquisition is the competition model of E. Bates and B. MacWhinney (1989). In language change, a number of linguists have put forward competing motivations models, in particular J. Haiman (1985).

More recently, some functionalist linguists (P. Hopper, J. Haiman, and J. Bybee) have emphasized the dynamic character of language in ordinary use, and have argued that a speaker's grammatical knowledge should not be considered to be as static and immutable as is usually believed. They argue that a speaker's grammatical knowledge is not a tightly integrated system, but rather a more loosely structured inventory of conventionalized routines that have emerged through language use. The empirical research of the functionalist linguists who advocate this view has focused on the role of frequency of use on the entrenchment of grammatical knowledge, and on inductive models of abstracting grammatical knowledge from exposure to language use, both in acquisition and in adult usage. There have been a number of studies of inflectional paradigms of words that support the hypothesis that frequency of use influences

grammatical representation; however, studies of syntactic construction in this approach are in their infancy.

4. *Directions and Prospects for Functional Approaches to Grammar*

Taken as a whole, functional approaches to grammar are widely practiced in linguistics. Nevertheless, there are important divisions between different schools of functionalist theory and between functionalists and formalists.

A common misunderstanding (among formalists in particular) is the belief that functionalists do not accept the existence of *arbitrariness* in language, that is, that not all aspects of grammatical knowledge are explainable by functional principles. There is no functionalist that holds that view. However, most functionalist theories have not developed a model of the representation of grammatical form, instead focusing their attention on the analysis of functional domains and their role in accounting for grammatical structure. There are significant exceptions to this observation, notably the mixed formal/functional theories of Dik and Van Valin, the model of lexical and morphological representation of Bybee, and construction grammar in cognitive linguistics. Construction grammar appears to be gaining a wider audience, and some functionalists are beginning to develop more detailed models of grammatical structure that are compatible with functionalist theories.

A significant split in functionalism is between those who are more cognitively oriented, focusing their attention on cognitive explanations, and those who are more discourse oriented. These two functionalist approaches have largely gone their own ways, developing models of language meaning (cognitive linguists) and language in use (the discourse functionalists). However, there is some functionalist work which draws on both cognitive and discourse models (see, for example, Givón (1979) and Chafe (1994)). Some functionalists have argued for a rapprochement between the two approaches, since they are compatible: language must be understood from both a psychological and a social-interactional perspective.

Functionalist approaches to language generally have also neglected research in sociolinguistics (see *Sociolinguistics*). Sociolinguistics begins with the hypothesis that language is a dynamic, variable phenomenon, and argues that factors outside linguistic form, namely social parameters, account for variation and change in language. In both of these respects, sociolinguistics shares theoretical assumptions in common with functionalism. The increase of research on social-interactional factors and on relative frequency of use in functionalist research may lead to an integration of social with other functional factors

in a social-functional theory of linguistic variation and change.

A rapprochement between functionalism and formalism is less likely, because the two differ in fundamental assumptions about the nature of language and the relation between form and function. There are some respects in which practical research is converging among formalists and functionalists, however. For instance, competing motivations as a mechanism of analysis is now widely used in formalist theories under the name of *Optimality Theory*. The rise of computer-aided corpus research has led to a convergence in interests on naturally occurring language and the role of usage in grammatical structure. Some formal semanticists have made reference to prototypes and polysemy, and some semanticists working in a generative grammar framework (Jackendoff 1983) also give conceptualization a major role in analysis.

See also: Functionalism, History of; Functionalism in Anthropology; Functionalism in Sociology; Generative Grammar; Grammaticalization; Linguistic Typology

Bibliography

- Bates E, MacWhinney B 1989 Functionalism and the competition model. In: Bates E, MacWhinney B (eds.) *The Cross-linguistic Study of Sentence Processing*. Cambridge University Press, Cambridge, UK
- Birner B J, Ward G 1998 *Information Status and Noncanonical Word Order in English*. Benjamins, Amsterdam
- Bybee J L 1985 *Morphology: A Study into the Relation between Meaning and Form*. Benjamins, Amsterdam
- Chafe W 1994 *Discourse, Consciousness and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. University of Chicago Press, Chicago
- Comrie B 1989 *Language Universals and Linguistic Typology*, 2nd edn. University of Chicago Press, Chicago
- Croft W 1990 *Typology and Universals*. Cambridge University Press, Cambridge, UK
- Croft W 1995 Autonomy and functionalist linguistics. *Language* 71: 490–532
- Dik S C 1997 *The Theory of Functional Grammar*, 2nd rev. edn. 2 Vols. Mouton de Gruyter, Berlin
- Givón T 1979 *On Understanding Grammar*. Academic Press, New York
- Greenberg J H 1990 *On language: Selected writings of Joseph H. Greenberg* (ed. Denning K, Kemmer S). Stanford University Press, Stanford, CA
- Haiman J 1985 *Natural Syntax*. Cambridge University Press, Cambridge, UK
- Hawkins J A (ed.) 1988 *Explaining Language Universals*. Blackwell, Oxford, UK
- Hawkins J A 1994 *A Performance Theory of Order and Constituency*. Cambridge University Press, Cambridge, UK
- Jackendoff R 1983 *Semantics and Cognition*. MIT Press, Cambridge, MA
- Kuno S 1987 *Functional Syntax: Anaphora, Discourse and Empathy*. University of Chicago Press, Chicago

- Lakoff G 1987 *Women, Fire and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press, Chicago
- Langacker R W 1987/1991 *Foundations of Cognitive Grammar*, 2 Vols. Stanford University Press, Stanford, CA
- Prince E F 1988 Discourse analysis: A part of the study of linguistic competence. In: Newmeyer F J (ed.) *Linguistics: The Cambridge Survey, vol. II: Linguistic Theory: Extensions and Implications*. Cambridge University Press, Cambridge, UK
- Talmy L 2000 *Toward a Cognitive Semantics*, 2 Vols. MIT Press, Cambridge, MA
- Tomasello M (ed.) 1998 *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*. Erlbaum, Mahwah, NJ
- Van Valin R D Jr, LaPolla R J 1997 *Syntax: Structure, Meaning and Function*. Cambridge University Press, Cambridge, UK

W. Croft

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Grammatical Agreement

Grammatical agreement is a phenomenon in which word forms co-occurring in a clause are sensitive to each other. Inflected forms often agree in their values of number, gender, or person, as can be seen from the contrasting examples in (1a, b) to (4a, b)

- (1) French a. *La^{fsg} fille^{fsg} est^{3sg} belle^{fsg}.*
'The girl is beautiful'
b. *Les^{pl} garçons^{mpl} sont^{3pl} beaux^{mpl}.*
'The boys are handsome'
- (2) French a. *Je^{1sg} les^{3pl} ai^{1sg} vues^{fp1}.*
'I have seen them'
b. *Tu^{2sg} l'^{3sg} as^{2sg} vu^{msg}.*
'You have seen him'
- (3) Finnish a. *Taina löysi minun^{1sg} kirja-ni^{1sgPoss}.*
'Taina found my book'
b. *Taina löysi heidän^{3pl} kirja-nsa^{3Poss}.*
'Taina found their book'
- (4) German a. *Ein kleines^{nsg} Mädchenⁿ saß^{3sg} auf der Treppe, und es^{3nsg} lachte.*
'A little girl sat on the stairs, and she laughed'
b. *Ein kleiner^{msg} Junge^{msg} stand^{3sg} daneben, und er^{3msg} heulte.*
'A little boy stood beside, and he cried'

In (1), the definite article agrees with the following noun, and both the copula verb and the predicative adjective agree with the whole subject phrase. If any item of (1a) is exchanged with the corresponding item of (1b), the sentence becomes ungrammatical (e.g., **la fille est beaux*). Example (2) shows two agreement relations intertwined: the auxiliary agrees with the subject pronoun, while the participle agrees with the object clitic. Sentence (3) is an example of a possessed noun agreeing with the possessor in the genitive. In (4)

it is the pronoun in the second clause that agrees with the antecedent noun phrase in the first clause. Notice that *Mädchen* 'girl' is a neuter noun, triggering a neuter pronoun even though it refers to a female being; in contrast, the English pronouns in the translation are selected solely on semantic grounds.

Grammatical agreement can be seen as an interplay of morphological, syntactic and semantic aspects (Anderson 1988, Lapointe 1988, Corbett 1994, 1995, 1998, Pollard and Sag 1994). Morphosyntactically, agreement is a relation between syntactic constituents in virtue of the fact that the word forms they consist of bear similar information, either inherently or by means of morphological affixation (inflection). Semantically, agreement serves to keep record of discourse referents: only constituents that relate to the same referent may overtly agree with each other (Lehmann 1982). The sentences in (1) are about one referent (a single person or a group of persons), while those in (2) are about two referents in a seeing-relation.

Generally, agreement relations are found among the following elements:

- (a) NP- (or DP-) internally: determiner and noun, attribute and noun, possessor and possessed noun;
- (b) a predicate and its arguments: verb–subject, verb–object, preposition–object, predicative noun/adjective–argument; and
- (c) a pronoun or anaphor and its antecedent.

The categories that may appear in these relations are gender (e.g., feminine f, masculine m, animate, or neuter n), number (e.g., dual, plural pl, and singular sg), and person (1st, 2nd, and 3rd, or inclusive 1 + 2).

Nouns and pronouns (or pronominal affixes/clitics) function as the controller (or source) of an agreement relation in virtue of their features. Nouns are classified inherently for gender (or noun class), and sometimes also for number (cf. *scissors*, *trousers*, *brethren*), whereas pronouns are classified inherently for person and number, and sometimes also for gender. In some instances, this inherent classification is semantically based: as a default, female beings are represented by feminine nouns, but male beings by masculine nouns; aggregates are represented by plural nouns; addressees are represented by second person pronouns, but speakers by first person pronouns. In most other cases feature assignment is purely grammatical (consider the distinct genders in French *lune^f*, *soleil^m* vs. German *Mond^m*, *Sonne^f*, 'moon,' 'sun') and may even override the default (cf. German *Mädchenⁿ* 'girl') (Corbett 1991).

The grammatical elements that agree with the noun or the pronoun are said to be the agreement targets (or controllees); in (1), the article *la* is the target of *fille*, which is f, and both the copula *est* and the adjective *belle* are targets of the DP *la fille*, which is fsg. The agreement features of controller and target do not have to be identical, but they must be compatible with each other, as in *la fille^{fsg}* and *est^{3sg}*. In order for a DP to be the controller of an agreement relation, the

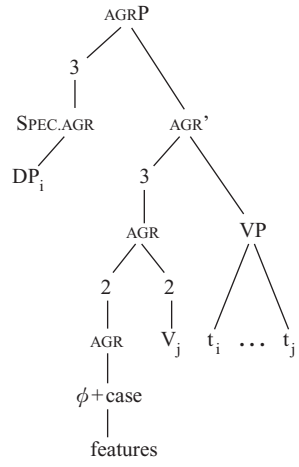
fsg ACC fsg ACC fsg ACC
 GEN GEN GEN

NOM ACC NOM ACC

j i i
 j j j

AGR AGR
 TENSE VOICE
 NOM
 PEC AGR AGR
 AGR
 AGR OSS
 PEC AGR AGR

(8)



AGR

ffi

AGR

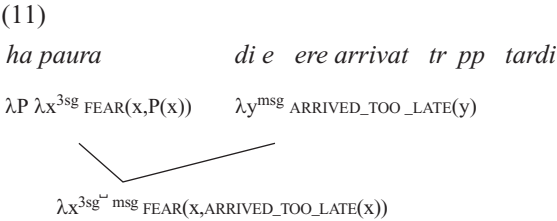
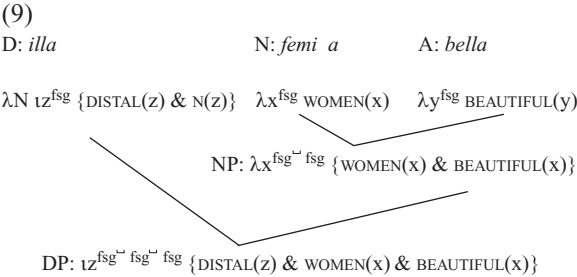
^{fsg} MO EST

fsg

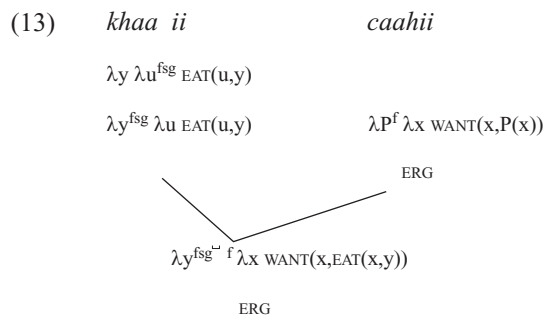
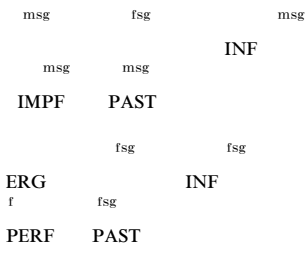
IN EX CONTENT AGR
IN EX

l

3msg 3sg msg

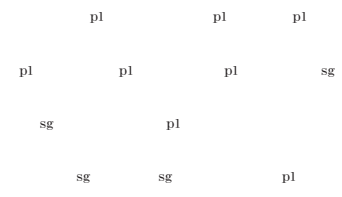
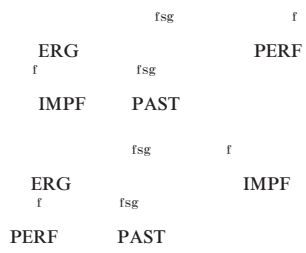


msg



ff

ffi



	sg	3sg	3pl		npl		mpl	3pl
ET					npl	mpl	mpl	
						ACC	ACC	

ff

		sg	
		pl	
	pl		pl
5	sg		
5	pl		

ff

pl	3pl

pl	3	3pl-anim
NOM		
pl	3pl-anim	3
NOM		

Grammatical Gender

Gender is a remarkable category, pervasive in some languages, for instance in the Indo-European and Dravidian families, and lacking in others, as in the Uralic family. We start with a brief consideration of terms, then examine how the number of genders in a given language can be established. The number varies considerably: two and three genders are often found, four and five are not unusual, while Fula (a Niger-Kordofanian language) has around 20. Most attention will be given to the problem of how the speaker assigns nouns to genders. Gender systems may have sex as a component, as in languages with masculine and feminine genders; but equally sex may be irrelevant, as in the Algonquian languages where the distinction is between animate and inanimate. Finally we consider prospects for investigating the category further.

1. Terms

The defining characteristic of gender is agreement; a language has a gender system only if noun phrases headed by nouns of different types control different agreements. The evidence that nouns have gender in a given language thus lies outside the nouns themselves. There is no substantive difference between 'genders' and 'noun classes'; the different terms are merely the products of different linguistic traditions. For instance, we find systems with three genders, to which nouns are assigned by similar rules, in both Kannada (a Dravidian language) and in Godoberi (a Nakh-Daghestanian language). By tradition the first is said to have three genders, and the second three noun classes. We shall use the term 'gender' in both instances.

2. The Number of Genders

The approach to gender sketched in Sect. 1 is based on Zaliznjak's (1964) notion of 'agreement class.' Two nouns are in the same agreement class only if they take the same agreements under all conditions, that is if we control for other grammatical categories such as case and number. Conversely, if two nouns differ in their agreements when factors such as case and number are held constant, then they belong to two different agreement classes and it will normally be the case that they belong to two different genders. While in many languages there is no dispute as to the number of genders, there are other languages where the question is far from straightforward. In the case of Romanian there has been a lengthy debate. The analytical problem of determining the number of genders and the tests for deciding the gender of a given noun depends on separating out the classes into which nouns are divided (the controller genders) from the number of different genders marked on agreement targets (the

target genders). Frequently the two match up, but in several languages they do not. A full treatment of the subject is not possible here; for a detailed account and extensive references see Corbett (1991).

3. Gender Assignment

We now ask how nouns are distributed over the genders of a given language, in other words, how does a native speaker 'know' what gender a noun belong to? Models of the mechanisms by which nouns are allotted to genders are called 'assignment systems.' Assignment may involve two sorts of information about the noun: its meaning and its form. We shall consider semantic systems first.

3.1 Strict Semantic Assignment Systems

This type is found in Dravidian languages like Kannada, spoken in the state of Karnataka, southern India (Sridhar 1990, p. 198). Nouns denoting male humans are masculine, those denoting female humans are feminine. There are also deities, demons, and heavenly bodies in these genders. All remaining nouns, including those denoting infants and animals, are neuter. Thus *appa* 'father,' and *candra* 'moon' are masculine, *amma* 'mother' is feminine, and *na:yi* 'dog' is neuter. Certain other Dravidian languages like Kolami have only two genders: nouns denoting male humans are masculine and all others belong in the nonmasculine gender (Emeneau 1955, p. 73). Elsewhere we also find the reverse. Thus in Diyari, a language of South Australia, there is a gender for nouns with female referents (such as women, girls, doe kangaroos), and the other is for all remaining nouns (Austin 1981). In Alamlak, a Sepik Hill language of Papua New Guinea, the masculine gender includes nouns denoting males and those denoting things like crocodiles, pythons, and arrows, which are long and thin, while the feminine is for nouns denoting females or short, squat items like turtles, frogs, and chairs (Bruce 1984, pp. 96–8, Foley 1986, pp. 80–81). An interesting system with some similar features is found in Taiap (Kulick and Stroud 1992).

3.2 Predominantly Semantic Assignment Systems

Many languages have semantic assignment rules, which, however, lack the coverage of those seen in Kannada. An example of what we shall call predominantly semantic assignment is found in Tsakhur, which is a member of the Lezgic group of the Nakh-Daghestanian family. Its four genders are shown in Table 1.

Assignment to genders I and II is straightforward: I is for male humans (also gods, angels, and so on) while II is for female humans (and female mythical beings). Most remaining animates are assigned to gender III,

Table 1
Gender assignment in Tsakhur

Criterion	Gender	Examples	Glosses
male rational	I (masculine)	dak _j gade	father boy
female rational	II (feminine)	jed _j sos	mother bride
animate (also some inanimates)	III (animate)	balkan	horse
other (including a few animates)	IV (neuter)	dadal (qow)	cock (nose)
		t'et'	flower
		nise (kabaj)	cheese (butterfly)

though a few belong to gender IV, along with some mythical beings. Inanimates are found in genders III and IV (Melnikov and Kurbanov 1964, Ibragimov 1990, pp. 54–6, and references therein, Kibrik 1999, pp. 48–9).

Languages of this general type are widespread. For some of them, researchers have proposed that abstract semantic criteria partly miss the point, and that if we can gain a better grasp of the world view of the speakers, we can then understand the assignment system. The most discussed case is that of Dyirbal, a language of North Queensland (Dixon 1972), which has four genders, primarily for: male humans and non-human animates (I), female humans (II), non-flesh food (III) and the residue (IV). There are numerous apparent exceptions: the moon is in the first, masculine gender and the sun is the second, feminine gender. However, in Dyirbal mythology, as indeed in much of Australia, the moon is the husband of the sun; in Dyirbal the mythological significance of referents is a key factor in assignment. Dixon's data and analysis continue to provoke debate, as shown by the recent paper by Polinsky (1998). There is similar debate as the role of world view in the gender assignment system of Ojibwa (and in other Algonquian language of southern Canada and northern USA).

In all the languages discussed so far, the meaning of the noun determines gender. In the strict assignment systems, the rules are immediately obvious; in the predominantly semantic systems, there may be a minority of exceptions, but these exceptions have been claimed to be largely apparent in some languages, once the cultural setting of the language is taken into account. If we ask which are the semantic criteria on which semantic systems can be based, we see recurring patterns. Quite often we find animate ~ inanimate, human ~ nonhuman, and male ~ female. Sometimes there is a gender for diminutives, as in various Bantu languages. There are also less usual genders, such as that for nonflesh food (Dyirbal) and the gender for insects (found in the Rikvani dialect of the Nakh-

Daghestanian language Andi). A criterion which is sufficient to define a gender in one language may be just one factor in the assignment to a gender in another. Thus the Bantu language Chichewa has a gender for diminutives, while in Dizi, diminutives together with nouns denoting females form a gender.

3.3 Formal Assignment Systems

In many languages, assignment by semantic rules would fail to assign many nouns to a gender. While in languages like Kannada the nouns not assigned by the semantic rules (the 'remainder' or 'semantic residue'), all belong to a single gender, in many languages they are spread over more than one gender. In such languages we find additional rules for assigning nouns to genders according to their form. Note the asymmetry: languages may use semantic rules, or semantic *and* formal rules, but not just formal assignment rules. In no language are nouns assigned to genders, as defined earlier, by purely formal rules. An example would be a language in which there were two agreement classes, and the nouns in the first all ended in a consonant cluster, and those in the second did not, and there was no semantic regularity for the distribution of nouns. I claim that this hypothetical type does not exist. Formal assignment rules may appeal to two types of information: phonological and morphological.

3.3.1 Phonological. A clear example of assignment depending on phonological information is provided by Qafar (Afar), an East Cushitic language spoken in north-eastern Ethiopia and in Djibouti (data from Parker and Hayward 1985). Qafar has rather standard semantic assignment rules, namely that for sex-differentiable nouns, those denoting males are masculine and those denoting females are feminine. It is the nouns which fall outside these rules, the residue,

Table 2
Russian nouns belonging to the semantic residue

Masculine	Feminine	Neuter
žurnal 'magazine'	gazeta 'newspaper'	pis'mo 'letter'
dom 'house'	izba 'hut'	zdanie 'building'
čaj 'tea'	voda 'water'	pivo 'beer'
dub 'oak'	sosna 'pine'	derevo 'tree'
ogon' 'fire'	peč' 'stove'	plamja 'flame'
avtomobil' 'car'	mašina 'car'	taksi 'taxi'
stul 'chair'	taburetko 'stool'	kreslo 'armchair'
den' 'day'	noč' 'night'	utro 'morning'
čas 'hour'	minuta 'minute'	vremja 'time'
nerv 'nerve'	kost' 'bone'	serdce 'heart'
nos 'nose'	ščeka 'check'	uxo 'ear'
lokot' 'elbow'	lodyžka 'ankle'	koleno 'knee'
flag 'flag'	emblema 'emblem'	znamja 'banner'

Table 3
Inflectional classes in Russian

	I	II	III	IV
NOMINATIVE	žurnal	gazeta	kost'	pis'mo
ACCUSATIVE	žurnal	gazetu	kost'	pis'mo
GENITIVE	žurnala	gazety	kosti	pis'ma
DATIVE	žurnalu	gazete	kosti	pis'mu
INSTRUMENTAL	žurnalom	gazetoj	kost'ju	pis'mom
LOCATIVE	žurnale	gazete	kosti	pis'me
gloss	'magazine'	'newspaper'	'bone'	'letter'

which are of interest. For them the phonological rules apply: nouns whose citation form ends in an accented vowel are feminine (for example, *karmà* 'autumn') while all others are masculine (for example, *gilàl* 'winter,' which does not end in a vowel, and *tàmu* 'taste,' which does end in a vowel, but not an accented one). These rules operate with few exceptions. Moreover, nouns denoting males and females typically fit them too (for example *bàqla* 'husband' and *barrà* 'woman, wife'). It might seem that we could dispense with semantic rules for Qafar, and treat it as being of a quite different type. However, while the phonological rule works for the vast majority of cases, we find the crucial example *abbà* 'father,' which is masculine, even though it ends in an accented vowel, which 'should' be an indicator of feminine gender. Conversely, *gabbixeèra* 'slender-waisted female' is feminine, though the accent is non-final. Qafar is remarkably for the simplicity of its phonological assignment rules, and for the fact that when the semantic and phonological rules are both applicable, they almost always give the same result. However, in cases of conflict, the semantic rules take precedence (as is the normal situation in gender assignment systems).

Other examples can occur in various parts of the world. Even French, often claimed to have no system to its gender assignment, has been demonstrated to have a phonological system, though not as simple as

that of Qafar. For example, of 938 nouns ending in *ê*, 99 percent are masculine (like *le pain* [pɛ̃] 'the bread'). And of 1,453 nouns in /ɔ/, 94.2 percent are masculine (like *le ménage* [mena:ʒ] 'the household' (for details see Tucker et al. 1977).

3.3.2 Morphological. In Russian, which is typical of many other Indo-European languages, nouns denoting male humans are masculine and those denoting female humans are feminine. But unlike the situation in languages like Kannada, it is not the case that the semantic residue is assigned to the neuter gender; in Russian the residue is shared between the three genders, with the neuter gender not even receiving the majority. Searching for additional semantic criteria is not at all promising, as the following data suggests (Table 2).

If we look instead at the morphology of the nouns, then we can make progress. There are four main inflectional classes in Russian, each with several thousands of nouns (for justification of this not uncontroversial view see Corbett 1982, pp. 202–11). There are six cases and two numbers (though no paradigm has 12 distinct forms because of various syncretisms). We give just the singular forms in Table 3.

The native speaker needs to know how a noun inflects, in order to produce grammatical utterances.

On the basis of that information, the assignment rules are straightforward. Nouns in class I are masculine, those in classes II and III are feminine, and those in IV are neuter. At first we might think that the semantic assignment rules are superfluous, since *mal'čik* 'boy' is in class I, while *devuška* 'girl' is in class II, and *mat'* 'mother' is in class III. In other words, many of the sex-differentiable nouns would be assigned to the appropriate gender by the morphological assignment rules. But there are also instances where this is not so, for instance, *djadja* 'uncle,' which denotes a male but is in class II, whose nouns are typically feminine. *Djadja* 'uncle' is masculine. Nouns like this show, once again, that we do not find languages where formal assignment rules are sufficient. Further rules are required in Russian for indeclinable nouns (like *taksi* 'taxi,' which is indeclinable and neuter). Morphological assignment systems are found in various other Indo-European languages. Looking further afield, in Arabic too, gender is assignable in the main according to morphology (Cowell 1964, pp. 372–5).

The different types of assignment criteria may overlap in ways which make it difficult to establish their relative weight in a given language. Thus there may be small clusters of nouns which can be accounted for by semantic criteria (apart from those covered by the main semantic rules) even within systems where formal rules have a major role (as shown for German by Zubin and Köpcke 1986, Köpcke and Zubin (1996). However, whenever gender languages are analyzed in sufficient detail, the gender of the vast majority of nouns turns out to be predictable.

4. Prospects

Studies on grammatical gender continue to appear regularly, including valuable studies of previously unresearched languages. Here we highlight two main areas where we can expect substantial new developments: lexicology and modeling.

4.1 Lexicology

Gender is a key category for lexicology, since it is crucial for understanding the nature of lexical entries. In a gender language the gender of a noun must be available. On the other hand, as we have seen, it does not normally need to be specified, since the information can be derived, by assignment rules, from other information which must be stored. We still do not know the full range of information which may be relevant (for instance, the possible semantic features which may determine gender assignment).

There are further surprises: it is possible too for certain nouns to be of different genders according to the sex of the speaker; this is found in Garifuna, a member of the Arawak family spoken in Belize, Honduras, Nicaragua, and Guatemala (Taylor 1977,

p. 60, Munro 1998). These two points raise interesting questions as to the shape of lexical entries, a general question which is of concern to lexicologists and increasingly to psycholinguists (see, for instance, Berkum (1995).

4.2 Modeling

When discussing some types of assignment system, the evidence is so clear-cut that alternative analyses are hard to imagine. However, with morphological assignment systems, particularly when the number of genders and the number of inflectional classes is similar in a given language, there are other possible analyses, in particular some would specify gender in the lexical entry and derive inflectional information from it. Arguments in favour of the approach taken in Sect. 3.3.2 include the following; first, and primarily, it is claimed that in languages like Russian the internal evidence shows that the gender from inflectional class approach is preferable (it proves impossible to predict inflectional class from gender but it is possible to predict gender from inflectional class); and second, there is the typological claim that one or other of the possible types of assignment system (or some combination) will always be in play in a gender language such that gender is predictable for the vast majority of nouns and so if the Russian type system were treated differently this would create an anomaly.

However, there is a further step which we can take, of a rather different nature. Since the Russian type languages are crucial, it worth investing additional time in clarifying them. Hence in Fraser and Corbett (1995) we give an account of the Russian gender and inflectional class relations in the Network Morphology framework. The analysis is implemented in the lexical knowledge representation language DATR (Evans and Gazdar (1996, Gazdar forthcoming), which means that the analyses can be verified by computer. Given minimal lexical entries for nouns our account does indeed predict the right genders for a wide range of nouns. An account like this which is both explicit and testable is not available for the alternative approach.

A more complex type is that in which gender is again assigned according to morphological class, but where the morphological class is itself predictable. One such case is Arapesh; a language of the Torricelli family, spoken on the north coast of Papua New Guinea. Genders and morphological classes form a substantial part of Fortune's (1942) grammar, which led to work based on it by Aronoff (1992, 1994, pp. 89–114). Once again, there are semantic assignment rules based on sex. But then there are morphological assignments rules, which determine 13 genders on the basis of 22 morphological classes (which are in turn predictable largely from phonological information). Again we have demonstrated how this system can be analyzed within the Network Morphology framework and have provided an implementation to demonstrate that our

claims are indeed valid (Fraser and Corbett (1997). There is a complex interaction of assignment to gender and to morphological class in Mayali, a non-Pama-Nyungan language of north Australia, which has four genders and five morphological classes for nouns. For the essential data see Evans (1997) and for details of a formal model of this system see Evans et al. (2001). Finally, there is interesting work on the development of assignment systems over time, including some modelling: see Comrie and Polinsky (1998), Polinsky and Everbroeck (1998), Polinsky and Jackson (1999).

5. Conclusion

Gender remains a puzzling category, apparently deeply embedded into the structure of some languages, and missing in other. This is found even in related languages: thus several Indo-Aryan languages have two or three genders, with substantial agreement, while another, Bengali, has no gender even in the personal pronouns. Where we find gender, the assignment rules can be simple, as in Kannada and Qafar, or complex as in French and German. It is an area in which gradual progress is being made and one in which there is still much to learn.

See also: Grammatical Agreement; Grammatical Number; Pronouns

Bibliography

- Aronoff Mark 1992 Noun classes in Arapesh. In: Booij G, van Marle J (eds.) *Yearbook of Morphology 1991*. Kluwer, Dordrecht, The Netherlands, pp. 21–32
- Aronoff M 1994 *Morphology By Itself: Stems and Inflectional Classes* (Linguistic Inquiry Monograph 22). MIT Press, Cambridge, MA
- Austin P 1981 *A Grammar of Diyari, South Australia*. Cambridge University Press, Cambridge, UK
- Berkum J J A van 1996 The psycholinguistics of grammatical gender: Studies in language comprehension and production. Ph.D. dissertation. Nijmegen University Press, Nijmegen, The Netherlands
- Bruce L 1984 *The Alambalak Language of Papua New Guinea (East Sepik)*. Department of Linguistics, Research School of Pacific Studies, Australian National University, Canberra, ACT
- Comrie B, Polinsky M 1998 Gender in a historical perspective: Radial categories meet language change. In: Justus C F, Polomé E C (eds.) *Language Change and Typological Variation: In Honor of Winfred P. Lehmann on the Occasion of His 83rd Birthday: II Grammatical Universals and Typology*. Institute for the Study of Man, Washington, DC, pp. 566–89
- Corbett G G 1982 Gender in Russian: An account of gender specification and its relationship to declension. *Russian Linguistics* 6(2): 197–232
- Corbett G G 1991 *Gender*. Cambridge University Press, Cambridge, UK
- Cowell M W 1964 *A Reference Grammar of Syrian Arabic (based on the dialect of Damascus) (Arabic series 7)*. Georgetown University Press, Washington, DC
- Dixon R M W 1972 *The Dyirbal Language of North Queensland*. Cambridge University Press, Cambridge, UK
- Emeneau M B 1955 *Kolami: A Dravidian Language*. University of California Press, Berkeley and Los Angeles
- Evans N 1997 Head classes and agreement classes in the Mayali dialect chain. In: Harvey M, Reid N (eds.) *Nominal Classification in Aboriginal Australia*. John Benjamins, Amsterdam, pp. 105–46
- Evans N, Brown D, Corbett G G 2001 The semantics of gender in Mayali: Partially parallel systems and formal implementation. *Language* 77
- Evans R, Gazdar G 1996 DATR: A language for lexical knowledge representation. *Computational Linguistics* 22: 167–216
- Foley W A 1986 *The Papuan Languages of New Guinea*. Cambridge University Press, Cambridge, UK
- Fortune R F 1942 *Arapesh*. J J Augustin, New York
- Fraser N M, Corbett G G 1995 Gender, animacy and declensional class assignment: A unified account for Russian. In: Booij G, van Marle J (eds.) *Yearbook of Morphology 1994*. Kluwer, Dordrecht, The Netherlands, pp. 123–50
- Fraser N M, Corbett G G 1997 Defaults in Arapesh. *Lingua* 103: 25–57
- Gazdar G forthcoming Ceteris paribus. In: Kamp J A W, Rohrer C (eds.) *Aspects of Computational Linguistics*. Springer, Berlin
- Ibragimov G X (ed.) 1999 *Caxurskij jazyk*. Nauka, Moscow
- Kibrik A E (ed.) 1999 *Elementy grammatiki caxurskogo jazyka v tipologičeskom osveščanii*. Nasledie Press, Moscow
- Köpcke K M, Zubin D 1996 Prinzipien für Genuszuweisung im Deutschen. In: Lang E, Zifonoun G (eds.) *Deutsch typologisch: Jahrbuch des Instituts für deutsche Sprache 1995*. de Gruyter, Berlin, pp. 473–91
- Kulick D, Stroud C 1992 The structure of the Taiap (Gapun) language. In: Dutton T, Ross M, Darrell T (eds.) *The Language Game: Papers in Memory of Donald C. Laycock*. Department of Linguistics, Research School of Pacific Studies, Australian National University, Canberra, ACT, pp. 203–26
- Meľnikov G P, Kurbanov A I 1964 Logičeskie osnovanija imennoj klassifikacii v caxurskom jazyke. In: Ivanov V V (ed.) *Voprosy struktury jazyka*. Nauka, Moscow, pp. 157–70
- Munro P 1998 The Garifuna gender system. In: Hill J M, Mistry P J, Campbell L (eds.) *The Life of Language: Papers in Linguistics in Honor of William Bright*. de Gruyter, Berlin, pp. 443–61
- Parker E M, Hayward R J 1985 *An Afar-English-French Dictionary (with Grammatical Notes in English)*. SOAS, University of London, London
- Polinsky M 1998 Women are not dangerous things: Grammatical gender and categorization. Paper presented at the Workshop 'Processing of Grammatical Gender,' Max Planck Institute of Cognitive Neuroscience and Centre of Cognitive Sciences, University of Leipzig, Leipzig, 24–25 July 1998
- Polinsky M, Everbroeck E van 2000 Noun classes: Language change and learning. Development of gender classifications: Modeling the historical change from Latin to French. Unpublished paper, UCSD
- Polinsky M, Jackson D 1999 Noun classes: Language change and learning. In: Fox B A, Jurafsky D, Michaelis L A (eds.) *Cognition and Function in Language*. CSLI, Stanford, pp. 29–50
- Sridhar S N 1990 *Kannada (Descriptive Grammars Series)*. Routledge, London

- Taylor D 1977 *Languages of the West Indies*. Johns Hopkins University Press, Baltimore, MD
- Tucker G R, Lambert W E, Rigault A A 1977 *The French Speaker's Skill with Grammatical Gender: An Example of Rule-governed Behavior*. Mouton, The Hague
- Zaliznjak A A 1964 K voprosu o grammatičeskix kategorijax roda i oduševlennosti v sovremennom russkom jazyke. *Voprosy jazykoznanija* 4: 25–40
- Zubin D, Köpcke K M 1986 Gender and folk taxonomy: The indexal relation between grammatical and lexical categorization. In: Craig C (ed.) *Noun Classes and Categorization: Proceedings of a Symposium on Categorization and Noun Classification*, Eugene, Oregon, October 1983. John Benjamins, Amsterdam, pp. 139–80

G. G. Corbett

Grammatical Number

Number is often taken to be a relatively straightforward category. It is often used as the textbook illustration of an inflectional category, and the situation in languages like English is believed by many to be the normal if not the only one. English makes a semantic distinction, and a formal distinction between, for instance, *cat* and *cats*, *child* and *children*. In fact the number category reveals many surprises, and the system found in English is a rather extreme one, when seen within the space of possibilities established from a survey of the languages of the world. We shall consider characteristics of the English system in turn, showing some of the ways in which it is typical or atypical. We shall give greatest prominence to the question of number values.

1. Nominal and Verbal Number

In English number is a nominal category, it involves the distinction between instances such as *cat* and *cats*. When it appears on verbs, this is by agreement with a subject noun phrase, as in *the children were playing happily*. The plural verb still indicates the number of children, and not the number of playing events. There are many languages which have verbal number, that is, where the number of the verb indicates the number of events (see Durie 1986 for examples).

2. Number as an Obligatory Category

Number is an obligatory category in English. If someone says, *there's a visitor waiting to see you*, when there are several, this is misleading and inappropriate. In many languages, however, the use of the plural is not obligatory, but is used when it is important to mark number. One example is the Cushitic language Bayso:

- (1) Bayso (Dick Hayward, personal communications, Corbett and Hayward 1987)
- liban* *foofe*
lion.GENERAL watched.1SG
'I watched lion' (it could be one, or more than that)

Another language, from many examples, is the Arawak family language Bare of Venezuela and Brazil (Aikhenvald 1995, pp. 18–19).

3. The Nominals Involved in the Number System

In English the majority of nominals can mark number, from the personal pronouns and nouns denoting persons (like *woman*) down to those denoting inanimate objects like *door*, and even to some denoting abstracts like *hour*. Many languages restrict the number opposition to fewer nominals, namely those which are high in animacy (Smith-Stark 1974). Thus in Slave (an Athapaskan language spoken in parts of the Northwest Territories, British Columbia, and Alberta, Canada) number may be marked on the noun phrase provided the noun denotes a human or a dog (Rice 1989, p. 247).

4. The Semantics of Number

As long as we stay with innocent examples like English *cat* and *cats*, the semantics of number look clear, since *cats* is appropriate for more than one instance of *cat*. In fact there are many concealed difficulties: a concise overview of work on number within formal semantics can be found in Link (1998), who shows how the study of plurals relates to different movements in linguistic semantics, and provides a substantial bibliography.

If we move from ordinary plurals to consider *I* vs. *we*, it becomes clear that here we have an 'associative' use of the plural; *we* is most often used for 'I and other(s) associated with me.' At the other end of the Animacy Hierarchy, nouns like *coffee* have a plural, provided they are recategorized: if *coffee* denotes a type, or a typical quantity of coffee, then there is an available plural, as in: *she always keeps three different coffees*, and *we've ordered two teas and three coffees*.

5. Number Values

A striking way in which English is limited in its number system is in terms of the number values. In English we can distinguish only between singular and plural. Many languages have more distinctions, and it is these on which we shall concentrate.

5.1 The Dual

A common system is one in which singular and plural are supplemented by a dual, to refer to two distinct

Table 1
Emphatic pronouns in Sursurunga

	<i>singular</i>	<i>dual</i>	<i>'trial'</i>	<i>'quadral'</i>	<i>plural</i>
1 exclusive	iau	giur	gimtul	gimhat	gim
1 inclusive	—	gitar	gittul	githat	git
2	iáu	gaur	gamtul	gamhat	gam
3	-i/on/ái	diar	ditul	dihat	di

Source: Hutchisson 1986 and personal communications
Note: á is used to indicate schwa (ə)

real world entities. Examples are widespread. We will take Central Alaskan Yup'ik, as spoken in the south-western Alaska (Corbett and Mithun 1996); the number forms are: singular *qayaq* 'kayak,' dual *qayak* 'two kayaks,' and plural *qayat* 'three or more kayaks.' The addition of the dual has an effect on the plural, which is now used for three or more real world entities, and more generally a change in system gives the plural a different meaning

5.2 The Trial

While the dual refers to two distinct real world entities, the trial is for three, and it occurs in systems with the number values: singular, dual, trial, and plural. This system is found in Larike, a Central Moluccan language spoken on the western tip of Ambon Island, Central Maluku, Indonesia. (Central Moluccan forms part of the Central Malayo-Polynesian subgroup of Austronesian.) This example is from Laidig and Laidig (1990, p. 7):

- (2) *Kalu iridu-ta-peu au-na-wela*
if 2.TRIAL-NEG-go 1.SG-IRR-go.home
'If you three don't want to go, I'm going home.'

In this language the dual and trial forms can be traced back to the numerals 'two' and 'three,' and the plural comes historically from 'four.' Such developments are fairly common in Austronesian languages. Sometimes the term 'trial' is used according to the form of the inflections (derived historically from the numeral three), even when the forms are used currently for small groups including those greater than three (and so are paucals) and sometimes the term is used according to meaning (for genuine trials). The Larike trial is a genuine trial, used for exactly three (Laidig and Laidig 1990, p. 92).

5.3 The Paucal

Paucals are used to refer to a small number of distinct real world entities, somewhat akin to the English quantifier 'a few' in meaning. There is no specific upper bound on the use of paucals (their lower bound will vary according to the system in which they are found). Let us return to the Cushitic language Bayso. Besides the general number forms (as in (1), where

number is not specified), there are also singular, paucal, and plural. Here is the paucal:

- (3) Bayso (Dick Hayward, personal communications, Corbett and Hayward 1987)
luban-jaa foofe
lion-PAUCAL watched.1.SG
'I watched a few lions'

The paucal is used in Bayso for reference to a small number of individuals, from two to about six. (Bayso has this system in nouns but not in pronouns.) The Bayso system is rare; normally the paucal is found together with a dual, so that the number values are singular, dual, paucal, and plural. This widespread system is found, for example, in Yimas, a Lower Sepik language spoken in the Sepik Basin of Papua New Guinea (Foley 1991, p. 216).

5.4 The Largest Systems

It is worth asking what is the largest system of number values. There are claims in the literature for languages with a quadral, for four distinct real world entities, thus giving a system of five number values. It is established that five number values can indeed be found, but the evidence specifically for a quadral is less clear. All the claims come from within the Austronesian family and the best documented case is Sursurunga (Hutchisson 1986, and personal communications), spoken in southern New Ireland. The forms labeled quadral are restricted to the personal pronouns, but are found with all of them, the first person (inclusive and exclusive), the second, and third (Table 1).

Here is an example of a quadral form in use (we will retain the traditional label 'quadral,' until we have justified replacing the term for Sursurunga):

- (4) Sursurunga (Hutchisson 1986 and personal communications)
gim-hat káwán
1.EXCL-QUADRAL maternal.uncle:nephew/
niece
'we four who are in an uncle-nephew/niece relationship'

The 'quadral' has two additional uses. First, in Sursurunga plural pronouns are never used with terms for dyads (kinship pairings like uncle-nephew/niece). Here the 'quadral' is used, for a minimum of four, and

not just for exactly four (Hutchisson 1986, p. 10). Second, in hortatory discourse, the speaker may use the first person inclusive 'quadral' in order to suggest joint action including the speaker, even though more than four persons are involved. These two special uses of the 'quadral' account for the majority of instances of its use. According to its meaning, the term 'quadral' is not appropriate. Rather the forms would be better designated 'paucal.'

Let us now consider the rest of the system of Sursurunga (data and judgments from Don Hutchisson, personal communications). The dual is used quite strictly for two people. It is also used for the singular when the referent is in a taboo relationship to the speaker. This is a special use; the dual's main use is as a regular dual. The trial will be used for three. However, it is also used for small groups, typically around three or four, and for nuclear families of any size. It is therefore not strictly a trial, rather it could be labeled a paucal (an appropriate gloss would be 'a few'). We noted earlier that this is a frequent development for the trial. The 'quadral,' as we have just seen, is used primarily in different ways with larger groups, of four or more (an appropriate gloss would be 'several'). This too would qualify as a paucal; we therefore have two paucals in Sursurunga, a paucal (traditionally 'trial') and a greater paucal (traditionally 'quadral'). The plural, as we would expect, is for numbers of entities larger than what is covered by the greater paucal, however, there is no strict dividing line (certainly not at the number five). Using semantic labels, we might represent the system in Sursurunga as having: singular, dual, paucal, greater paucal, and plural. Sursurunga has a well-documented five-value number category. There are certainly other languages with five number values; we do not have such detailed information as for Sursurunga and there is no certain case of a quadral: it seems that in all cases the highest value other than plural in such systems can be used as a paucal. Besides the split in the paucal, we may also find a split in the plural, with 'greater plurals' of different types, as for instance in Mokilese (Harrison 1976). Greater plurals may imply an excessive number or else all possible instances of the referent. They are as yet poorly researched, except in a few languages.

6. Conclusion

Number is indeed a much more complex category than is generally recognized, and the English system, though a common one, is at one extreme of the typological space. We noted that number may be a verbal category as well as a nominal one (but the possibilities are rather more restricted for verbal number). We also saw that number need not be obligatory, that the range of items which distinguish number varies dramatically from language to language, that the semantics of number are affected by

the particular item involved, and that the values of the number category can be just two or, in various combinations, up to five. More complex issues, which we have not been able to address here, include various interactions, notably between the forms of marking of number, the particular nominals which distinguish number in a given language, and the systems of number values. For these issues see Corbett (2000).

See also: Grammatical Gender; Number Systems, Evolution of; Pronouns

Bibliography

- Aikhenvald A 1995 *Bare* (Languages of the World/Materials 100). Lincom Europa, Munich, Germany
- Corbett G G 2000 *Number*. Cambridge University Press, Cambridge, UK
- Corbett G G, Hayward R J 1987 Gender and number in Bayso. *Lingua* 73: 1–28
- Corbett G G, Mithun M 1996 Associative forms in a typology of number systems: Evidence from Yup'ik. *Journal of Linguistics* 32: 1–17
- Durie M 1986 The grammaticization of number as a verbal category. In: Vassiliki N, VanClay M, Niepokuj M, Feder D (eds.) *Proc. of the Twelfth Annual Meeting of the Berkeley Linguistics Society: February 15–17, 1986*. University of California, Berkeley, CA, pp. 355–70
- Foley W A 1991 *The Yimas Language of New Guinea*. Stanford University Press, Stanford, CA
- Harrison S P, Albert S Y 1976 *Mokilese Reference Grammar*. University Press of Hawaii, Honolulu, HI
- Hutchisson D 1986 Sursurunga pronouns and the special uses of quadral number. In: Wiesemann U (ed.) *Pronominal Systems (Continuum 5)*. Narr, Tübingen, Germany, pp. 217–55
- Laidig W D, Laidig C J 1990 Larik pronouns: Duals and trials in a Central Moluccan language. *Oceanic Linguistics* 29: 87–109
- Link G 1998 Ten years of research on plurals—where do we stand? In: Hamm F, Hinrichs E (eds.) *Plurality and Quantification*. Kluwer, Dordrecht, The Netherlands, pp. 19–54
- Rice K 1989 *A Grammar of Slave*. Mouton de Gruyter, Berlin
- Smith-Stark T C 1974 The plurality split. In: La Galy M W, Fox R A, Bruck A (eds.) *Papers from the Tenth Regional Meeting, Chicago Linguistic Society, April 19–21, 1974*. CLS, Chicago, pp. 657–71

G. G. Corbett

Grammatical Relations

The linguistic concept 'grammatical relation' with its equivalents 'syntactic relation,' 'syntactic function' belongs to the most important concepts of traditional syntactic theory. By grammatical relations are understood structurally expressed characteristic syntactic relations between sentence constituents (see *Constituent Structure: Syntactic*), in particular between the

predicate and noun phrases (NPs). Thus, in the sentences (1) *Peter sent a message to Mary*, (2) *My friend sent a congratulatory telegram to his boss*, in addition to the predicate *sent* one can identify three NPs linked to the predicate by specific grammatical relations. One of the NPs (*Peter/my friend*) is the subject, the second (*a message/a congratulatory telegram*) is the direct object, while the third (*Mary/his boss*) is the indirect object. (Noun phrases that realize the same grammatical relation occupy the like-named position in the clause. Thus, the NP *Peter* in (1) occupies the subject position.) Each of the NPs in these sentences is structurally distinct from the other two and makes unique reference to a concrete participant in the events denoted by the given sentences. The subject and the direct object of this particular example are expressed positionally—the subject precedes the verb, the direct object follows the verb immediately—while the indirect object contains the preposition *to*.

Grammatical relations fulfill *distinguishing* and *characterizing* functions. First, they permit identification of the various participants in the event (what was sent, to whom and by whom it was sent), secondly they classify the NPs in the given sentence in terms of a restricted set of characteristic properties, allowing the identification of NPs in different sentences. Thus, in (3) *A message was sent to Mary (by Peter)*, (4) *Mary was sent a message (by Peter)*, both semantically related to sentence (1), the subjects will be, respectively, the noun phrases *a message* and *Mary*; in (5) *That he won is fine*, there is a clausal subject *that he won*; while in (6) *It is fine that he won* and (7) *It is dark*, the subject is the pronoun *it* (the so-called dummy subject). The concept ‘subject’ as a homogeneous syntactic relation unites different kinds of specific constituents. Each grammatical relation has a specific function (see *Grammar: Functional Approaches*) and specific linguistic means for its expression.

1. What Do Grammatical Relations Mean?

Despite the widespread use and obvious necessity of the terms subject and object, traditional grammar did not succeed in defining their meaning adequately. Definitions for ‘subject’ of the type ‘what the sentence is about’ are at first sight suitable for sentences (1)–(4), but they are dubious for (5), while the subjects in (6) and (7) clearly do not fall under a definition of this type.

In a series of investigations of the last decades (cf., for instance, the pioneering work of Keenan 1976) it has been established that in the most elementary sentences subject NPs correlate statistically with the meanings of various autonomous semantic domains.

First, subjects are linked to the semantic role properties of participants in the given situation. The subject usually encodes the agent, if there is one in the

sentence. If there is no agent in the sentence, then the prime candidate for subject position is the patient. In the presence of an agent the patient is the prime candidate for the position of direct object. Sentences with an agent and patient constitute the prototypical transitive clause with a Subject and direct object. Sentences with participants in other roles are assimilated to varying extents to the structure of transitive or intransitive clauses, or use the position oblique object. The participant with the role of recipient/beneficiary usually occupies the position of indirect object.

Second, there are statistical correlations with the discourse (communicative) domain (see *Linguistics: Information Structure*). Thus, subjects often communicate known (given) information, while objects communicate new information; subjects often coincide with the focus of empathy, i.e., with that participant from whose point of view the situation is described, and with the topic (theme), while objects often coincide with the comment (rheme).

Third, the ontological domain, connected with the properties of referents, influences the subject or object position of NPs. Nps denoting animate, and especially human participants tend towards the subject position, inanimates towards the object position (*Grammatical Gender*). Within the set of humans the greatest tendency towards subject position is shown by the speech act participants (concepts from the deictic domain, see *Deixis*).

In addition, concepts from the referential domain, which classify kinds of referents denoted by NPs and facilitate their correct identification by the addressee, also correlate with grammatical relations. Thus, the subject has a presumption of existence, usually has autonomous reference, and is definite and specific, while the direct object is often nonspecific.

However, it is not possible to assign any of the above-mentioned meanings as obligatory for the subject or direct object; in each given sentence these grammatical relations satisfy only some of these meanings.

What, then, is the invariant meaning of grammatical relations, if indeed one exists? There is no generally accepted answer to this question. However, the following can be offered as the most plausible answer. Grammatical relations order NPs in the sentence by rank in accordance with their overall communicative value in a concrete utterance. The ordering takes place by means of a system of rules which takes into account the concrete meanings of NPs according to all the relevant semantic domains, so that the first rank is assigned to the subject, the second to the direct object, then the indirect object and oblique objects, as shown in the so-called ‘relational hierarchy’:

$$\text{Su} > \text{DO} > \text{IO} > \text{OO}$$

where Su represents ‘subject,’ DO ‘direct object,’ IO ‘indirect object,’ and OO ‘oblique object.’ The subject

occupies a special position in this hierarchy. The subject is the most essential component of the sentence, and in some languages it is structurally necessary, in contrast to all other grammatical relations. The other grammatical relations are ranked primarily according to their degree of difference from the subject. In this way, grammatical relations have not an absolute, but rather a relative syntactic value in the sentence, and the relation of this value to meanings of various kinds is indirect in nature.

Grammatical relations are a means of cumulative expression of permitted combinations of more elementary meanings, primarily from the role and discourse domains. Given this, then, first, the same grammatical relation in different sentences of the same language can have a different set of meanings; and second, in different languages grammatical relations can be organized differently, since different combinations of these meanings may be grammaticalized (see *Grammaticalization*) into different grammatical relations.

It can easily be seen that the subject and direct object are furthest removed from the semantic role of a NP and that their basic characteristic is their rank position in the hierarchy. Together with the predicate they form the nucleus of the predicate–argument structure. Oblique objects, in contrast to subject and direct object, have a specific semantic role (e.g., place, time, cause, manner), and in a clause there can be more than one oblique object. The indirect object has an intermediate status. On the one hand, it is considerably more loaded, in terms of semantic role and ontological semantics, than are the subject and direct object, and in this it is like oblique objects, but on the other hand it is the easiest to incorporate into the nucleus, compare (1) and (8) *Peter sent Mary a message*. In (8), in contrast to (1), *Mary* occupies the position of direct object.

2. How are Grammatical Relations Encoded?

Languages use different means to encode grammatical relations. The best known morphological category specifically designated to encode grammatical relations is case (see *Morphological Case in Linguistics*). Case marking is present in many languages, e.g., in Latin and Greek, in the Slavic and Dravidian languages, and in the languages of Australia. Under case marking the grammatical relation is expressed by the morphological form of the head of the NP (i.e., of the noun, as in Latin and Russian) or by the forms of other structural constituents of the NP (e.g., forms of the article and noun in German). Alongside inflectional morphology many languages make use of grammatical words or particles (usually prepositions or postpositions), which combine with a NP to give analytical constructions that indicate the kind of grammatical relation (e.g., the combination of pre-

position *to* and NP, such as *to Mary* in (1), encodes the indirect object). The linear order of NPs in the sentence is the positional equivalent of inflectional or analytical marking (see *Word Order*) and is calculated according to the position of the NP relative to the verb, relative to the beginning or end of the sentence, or relative to other NPs. Morphological marking may be indicated not on NPs, but on the verb by means of agreement (see *Grammatical Agreement*): the form of the verb expresses information about the grammatical properties of that NP which stands in a particular grammatical relation to the verb. An example in English is agreement of the verb with a third-person subject in number, combining with the indication of grammatical relations through linear order. Some languages have polypersonal agreement of the verb with two or more NPs. Thus, in Abkhaz the verb has three agreement positions indicating agreement with specific syntactic types of NPs.

3. The Syntactic Functions of Grammatical Relations

The question of which grammatical relation a particular NP occupies in the sentence can by no means always be decided on the basis of the coding technique. The same means may be used to encode different grammatical relations and the same grammatical relation may use different coding techniques. Thus, in English an immediately postverbal prepositionless NP is not always a direct object: (9a) *He was sleeping all night*, cf. the equivalent Russian sentence with a time adverbial in the accusative: (9b) *On spal vsju noč'*, where the accusative is the typical case of the direct object. Conversely, in Russian the direct object can stand in the genitive case in the context of negation, e.g., (10a) *On kupil knig-u* (acc.) 'He bought a book' vs. (10b) *On ne kupil knig-i* (gen.) 'He didn't buy a book.' For this reason the question often arises of the criteria by which one can assign different grammatical relations to identically marked constituents or assign the same grammatical relation to differently marked constituents.

The decisive diagnostic role in identifying grammatical relations is played not by coding means, but by syntactic tests. The point is that the coding technique is merely the outer packaging that helps identify grammatical relations in the sentence, whereas the grammatical relations are really components of the syntactic structure of the sentence; their identity or nonidentity is determined in the last analysis on the syntactic level (see *Syntax*). In many languages the system of rules of syntactic organization operates to a considerable extent on the basis of units characterized in terms of grammatical relations, and thus the identity of syntactic behavior of constituents differing in form is an important criterion for uniting them as a single grammatical relation.

Within the clause, voice transformations are oriented to grammatical relations (see *Grammatical Voice*). Thus, the difference between active and passive voices is that different grammatical relations adhere to the same NPs in active and passive versions. Passivization is usually characterized as the promotion of the syntactic status of the direct object to that of subject and/or demotion of the syntactic status of the subject. In English both processes occur in passivization, cf. (1) and (2). The subject NP of the active loses its subject status (it is expressed by the prepositional phrase *by Peter*), while the direct object NP (*a message*) becomes subject (promotional passive). There are also languages in which passivization involves only the demotion of the status of the subject NP (demotional passive). However, in all cases passivization is defined in terms of grammatical relations, irrespective of the coding techniques used in the particular language.

Among nearly universal properties are constraints on reflexivization, operating in cases of coreference between two NPs in the same clause, e.g., (11) *Mary hates herself*, (12) *Mary does not think about herself*. Here, one of the NPs (*Mary*) must be in subject position and is a full identification of the referent (the controller of reflexivization). The other, nonsubject, coreferential NP is replaced by the reflexive pronoun (the target of reflexivization).

Under coordination there are also constraints on grammatical relations. Thus, in English coordination reduction is possible only with coreferential subjects. The second coreferential subject is omitted, e.g., (13) *John washed his hands and—[= John] began to eat*. Another construction formulated in terms of subjects is switch reference, attested in various languages: clauses are expressed differently according to whether their subjects coincide or not (usually by marking the verb of one of the clauses).

As a syntactic test one can use constructions with sentential complements with verbs like ‘want,’ ‘must,’ ‘begin.’ Sentential complements are usually expressed differently depending on the coreference versus non-coreference of the subject of the sentential complement and the subject of the main clause, cf. (14a) *Peter wants to congratulate Mary*, but (14b) *Peter wants Mary to congratulate him*. In (14a) the subject of the sentential complement is coreferential with the subject of the verb *want*, and is therefore not expressed overtly in the subordinate clause. In (14b) the coreferential NP is the direct object, therefore it cannot be omitted, but must be expressed by means of a pronoun.

The formation of relative clauses (see *Relative Clauses*) is constrained, in many languages, by the grammatical relation of the target of relativization. For instance, reduced (participial) relative clauses in English are restricted to subject targets of relatives: (15a) *the student buying a book*, (15b) *the book bought by the student*. In (15a) the target of relativization is in subject position (*The student buys a book*), therefore

the relative clause is formed directly from this structure. In (15b) the target of relativization is the direct object (*The student bought the book*), therefore in order to construct the participial version it is necessary first to passivize the relative clause, promoting the status of the target of relativization to subject position (*The book was bought by the student*). In some languages relativization is constrained also by the grammatical relation of the controller of relativization.

Not infrequently constraints on the formation of adverbial clauses (see *Adverbial Clauses*) also rely on grammatical relations. In Russian there is a set of subordinate clauses expressing background activity, the so-called converbal (gerundive) clauses. The subject of the converbal clause must be coreferential with that of the main clause (and is omitted in the converbal clause): (16) *Pokupaja ètu knigu, student potratil na poiski massu vremeni* ‘Buying this book, the student spent a lot of time in searching.’ In (16) the omitted subject of the converbal clause and the subject of the main clause is the NP *student*. Sentence (17) is ungrammatical: (17) **Pokupaja ètu knigu, u studenta propal košelek* ‘Buying this book, the student’s purse got lost,’ since *u studenta*, the NP in the main clause coreferential with the subject of the converbal clause, is not the subject (the subject of the main clause is *košelek* ‘purse’).

To the extent that many syntactic rules are operations in terms of grammatical relations, these rules can be used as tests to determine the syntactic status of a NP in morphologically ambivalent instances.

4. Initial vs. Derived (Primary vs. Secondary) Grammatical Relations

Certain syntactic processes change the grammatical relation of a particular NP or delete it. Such processes change the valency of a sentence (see *Valency and Argument Structure in Syntax*). In such instances it is important to distinguish initial and derived grammatical relations of NPs. Thus, in the passive the initial subject loses this grammatical relation, while the initial direct object becomes the subject of the passive sentence. It follows from this that syntactic processes are often formulated in terms of initial syntactic relations which are not present in the observable, derived syntactic structure. Their existence can nonetheless be reconstructed on the basis of the syntactic process that conditions the observed structure. In a series of instances properties of the initial grammatical relation of a NP are partially retained. For instance, in the examples (18a) *Uncle Tom built this hut for himself* and (18b) *This hut was built by Uncle Tom for himself*, the controller of the reflexive is in both instances the initial subject *Uncle Tom*, even

though in (18b) the derived, observable subject is *this hut*.

5. Nonuniversality of Grammatical Relations

In linguistics the aprioristic assumption of the universality of grammatical relations was long dominant. Many current syntactic theories derive from this assumption, in particular generative grammar (see *Linguistics: Incorporation*) and the theory of Relational Grammar. In the last decades, however, as a result of considerable progress in studying the nature of grammatical relations and thanks to the considerable broadening of empirical linguistic material this assumption is no longer self-apparent.

First of all, the functional and formal properties of grammatical relations characterized above (poly-functionality of content and cumulative expression) testify to their specific semiotic nature. Why should such a technique of identification and characterization of noun phrases be an obligatory property of natural language? Conversely, it is natural to allow that different languages might grammaticalize various concepts of different semantic domains and package them differently in linguistic form. What we identify as grammatical relations might be filled with different content in different languages and, moreover, one might find not a cumulative, but a separative technique of packaging elementary concepts. Finally, there is no *a priori* basis to assume that the concepts of some or even all the relevant domains should be grammaticalized in each language, given that so far not a single semantic domain has been found that would be grammaticalized in all languages (be it the category of gender, number, tense, aspect, mood etc., see *Grammatical Gender; Grammatical Number; Tense, Aspect, and Mood, Linguistics of*). Even as apparently as indispensable ontological category as part of speech (see *Word Classes and Parts of Speech*) has recently increasingly lost the assumption of universality.

Second, uncritical attempts to use relational concepts with respect to different languages often encounter serious empirical difficulties. One may start with the observation that the morphosyntactic means of encoding NPs in some languages is not only different from, but sometimes not even comparable with the encoding characteristic of Standard Average European languages. If one uses the widespread notation for the nuclear arguments with the symbols S, A, P (or O), where S is the nuclear argument of an intransitive clause, and A and P the agent-like and patient-like arguments, respectively, of a transitive clause, then in European languages the S/A arguments have identical encoding, grouping them together as the grammatical relation of subject, in opposition to the encoding of the P argument. Such a system is called accusative alignment. Alongside this possibility, many languages found in various continents have the ergative pattern, uniting the S/P arguments in opposition to the A argument. Here the S/P arguments are

usually morphologically unmarked, while the A argument is marked (ergative case or its analog). The ergative alignment of the nuclear arguments creates an empirical problem for the establishment of grammatical relations.

In the languages of the world other patterns are also found. One possibility is a pattern that distinguishes agent-like and patient-like arguments of both transitive and intransitive clauses (active alignment). Under active alignment one finds different expression of the single argument of verbs like 'run' and verbs like 'fall.' There is even the tripartite scheme, distinguishing all three arguments, S, A, and P. Finally, there is the neutral pattern, which fails to differentiate the S/A/P arguments by means of coding.

Languages characterized by 'unusual' non-accusative patterns are quite varied from a syntactic point of view. For instance, languages with ergative alignment may behave as syntactically accusative, i.e., in defiance of the coding technique S and A arguments are united in a single grammatical relation subject. Much more interesting is the situation where ergative alignment is also realized at the syntactic level: S and P arguments are united syntactically in the highest ranking grammatical relation, equivalent to the subject under accusative alignment. The A argument is also a nuclear argument, i.e., it occupies the second position in the Relational Hierarchy. Finally, there are ergative languages in which it is impossible to rank S, A, and P at the syntactic level. In languages of this type it is justifiable to speak of the absence of grammatical relations.

Among languages with active alignment are also found languages behaving accusatively at the syntactic level, as well as languages lacking syntactic relations. So far there is extremely little information available on languages with tripartite or neutral alignment. It is, however, quite clear that they too are quite varied. They may have or not have grammaticalized grammatical relations. In particular, there is evidence for the existence of languages with neutral alignment lacking grammatical relations. But one can assume the existence of a neutral alignment language opposing nuclear and nonnuclear arguments. In such a language the nuclear arguments would be equivalent to subject, and the nonnuclear arguments to oblique objects. If such a language is attested, then the thesis of the uniqueness of the subject will have been overturned.

Thus, grammatical relations both have a narrower sphere of use and are considerably more varied than is usually assumed.

It should be kept in mind that the thesis of the nonuniversality of syntactic relations is not universally held, and is even ignored in many works. However, the assumption of nonuniversality is preferable by far in empirical research of little-studied languages, since it does not impose an aprioristic standard on the description and helps to see the real state of affairs in each language under discussion.

6. Grammatical Relations and the Functional Approach to Language

The concept of grammatical relation has been known in linguistics since the time of the grammarians of antiquity. However, significant progress in understanding the nature of grammatical relations has been achieved only in the last decades thanks to the orientation towards the functional explanation of grammar (see *Grammar: Functional Approaches*), which has permitted establishing the polyfunctional character of the elementary categories which form the clusters known as grammatical relations. The majority of these categories were themselves discovered and investigated in the framework of the functional approach.

In this connection the question of the existence of different patterns of encoding S, A, and P arguments merits special attention. Proceeding from the now widely established assumption of the iconicity of the linguistic sign (see *Linguistics: Iconicity*), replacing the Saussurean assumption of the arbitrariness of the sign, one can allow for the existence of a functional motivation for the various coding patterns. These patterns unite arguments with different elementary roles in different ways. The basic prototype of all these groupings (see *Linguistics: Prototype Theory*) is given by the agent and patient of the transitive construction. From this viewpoint active alignment is a natural metonymic extension of the elementary roles of agent and patient, generalizing them to the hyperroles (macroroles) of actor (the participant in the event which performs, effects, instigates, or controls it) and undergoer (the participant in the event which does not perform, effect, instigate, or control it, but rather is affected by it in some way). Arguments with different elementary roles can be interpreted in terms of these hyperroles.

Accusative alignment is also motivated essentially not by grammatical relations, but by the hyperroles principal and patientive. The principal indicates the main participant, the 'hero' of the situation, who is primarily responsible for the fact that this situation takes place. In the transitive clause the principal is, of course, the agent-like argument, while in the intransitive clause the single nuclear argument has no competitors for the role of principal. The patientive is the most patient-like participant of a multi-participant event. This role unites, in addition to the patient, arguments of other types.

Ergative alignment articulates the S, A, P arguments by means of other hyperroles, absolutive and agentive. The absolutive is the metonymic extension of the patient to intransitive verbs, it is the immediate, nearest, most involved or affected participant in the situation. In the transitive clause the patient is most involved in the situation. In the intransitive clause there is no competition for the role of closest and most involved participant in the situation. The absolutive in

the transitive clause is in opposition to the agentive, the most agent-like participant of the multi-participant event.

Tripartite alignment distinguishes agentive and patientive of the transitive clause, distinguishing them from the single nuclear argument of the intransitive clause.

All these means of generalizing the elementary roles are in principle motivated, but a language has the option of grammaticalizing just one of them as its basic pattern. In addition, a language may have different patterns in different contexts, and examples of this kind have been attested.

7. Grammatical Relations and the Cognitive Approach to Language

The 'statistical' correlations between concepts of different semantic domains noted above and which are grammaticalized in grammatical relations and the means of their encoding are not chance, rather they have a cognitive nature, reflecting the human capacity for categorizing reality. Behind the limitless variety of whole grammatical systems lie the same stable cognitive structures of human thought and the pragmatic strategies (see *Pragmatics: Linguistic*) for their verbalization. The reconstruction of these structures and strategies on the basis of data from different languages is the task of cognitive typology (see *Linguistic Typology*).

Bibliography

- Aissen J 1999 Markedness and subject choice in optimality theory. *Natural Language and Linguistic Theory* 17: 673–711
- Cole P, Sadock J M (eds.) 1977 Grammatical relations. *Syntax and Semantics*. Academic Press, New York, Vol. 8
- Croft W 1991 *Syntactic Categories and Grammatical Relations*. The University of Chicago Press, Chicago
- Dik S 1997 *The Theory of Functional Grammar*. Mouton de Gruyter, Berlin
- Dixon R M W 1994 *Ergativity*. Cambridge University Press, Cambridge, UK
- Dixon R M W, Aikhenvald A Y (eds.) 2000 *Changing Valency: Case Studies in Transitivity*. Cambridge University Press, Cambridge, UK
- Dryer M 1986 Primary objects, secondary objects, and anti-datative. *Language* 62: 808–45
- Dryer M 1997 Are grammatical relations universal? In: Bybee J, Haiman J, Thompson S (eds.) *Studies in Language Function and Language Type*. Benjamins, Amsterdam, pp. 115–43
- Givon T (ed.) 1997 *Grammatical Relations: A Functional Perspective*. Benjamins, Amsterdam
- Keenan E 1976 Towards a universal definition of "subject". In: Li Ch N (ed.) *Subject and Topic*. Academic, New York, pp. 303–34
- Kibrik A E 1997 Beyond subject and object: Toward a comprehensive relational typology. *Linguistic Typology* 1: 279–346

- Li Ch N (ed.) 1976 *Subject and Topic*. Academic, New York
Perlmutter D M (ed.) 1983 *Studies in Relational Grammar*. The University of Chicago Press, Chicago
Plank F (ed.) 1979 *Ergativity: Toward a Theory of Grammatical Relations*. Academic, London
Van Valin R, LaPolla R J 1997 *Syntax: Structure, Meaning, and Function*. Cambridge University Press, Cambridge, UK

A. E. Kibrik

Grammatical Voice

1. Introduction

A typical sentence (clause) of a natural language consists of the predicate and its arguments. Each argument is associated with a certain participant role and is expressed by a certain grammatical function in a clause. Voice is the grammatical category that relates the participant roles associated with arguments on the one hand and the grammatical functions of the noun phrases (NPs) expressing these arguments on the other.

Predicates can be characterized by the number of arguments they take; a one-place predicate is intransitive, a predicate with two arguments is transitive. The ability of a predicate to subcategorize for semantic arguments does not necessarily mean that all these arguments are overtly expressed by NPs; some arguments may be omitted and some arguments may be understood (implicit) but never expressed.

In terms of semantic arguments, the description of grammatical voices relies crucially on the recognition of two types of core arguments of a predicate, agent (logical subject) and theme (logical object). An agent is a participant that has some or all of the following properties: it exists independently of the event named by the predicate, can cause a change of state in another participant, is capable of movement, is volitional and capable of intention (Gruber 1965, Fillmore 1968, Dowty 1991). A theme is a participant that has some or all of the following properties: it may not exist independently of the event, undergoes a change of state in the event, is causally affected by an agent.

Turning now to the grammatical functions subject and object, these functions can be manifested in all or some of the following formal properties: case-marking, linear position in a sentence, agreement, and ability to determine the form and/or linear position of other constituents (for example, the form of reflexives).

Based on how agent and theme are mapped into grammatical functions, linguistic theory recognizes the following main voices: active, passive, middle, and antipassive.

In the active voice, agent is mapped into the syntactic subject of a clause, while theme is mapped

into the syntactic object, as in the English example (1).

- (1) *Invisible God* [agent, subject]
created a visible world [theme, direct object]

In the passive voice, agent is mapped into an oblique phrase or adjunct, whereas theme may be mapped into the syntactic subject, as in (2).

- (2) *The visible world* [theme, subject]
was created by invisible God [agent, adjunct]

In the middle voice, the event denoted by the predicate is construed as developing without external causation; the main participant is both the source and the recipient (beneficiary) of the event. Thus, the main argument of the middle voice is construed as combining features of agent and theme. The contrast between the middle voice and the passive voice can be illustrated by the English (3) and (4); in (3), the ability to wash well is perceived as the property of cotton, and no external agent is permitted:

- (3) *Cotton* [theme, subject]
washes well (*by consumers)
(4) *Cotton* [theme, subject]
is washed (well) (by consumers)

In the antipassive voice, agent is mapped into subject and theme is mapped into an adjunct. Although English does not have a regular antipassive, the contrast between (5) and (6) is illustrative of the contrast between the active voice (theme = direct object) and antipassive (theme = adjunct):

- (5) *The soldiers* [agent, subject]
shot the pirates [theme, object]
(6) *The soldiers* [agent, subject]
shot at the pirates [theme, adjunct]

Alternations in grammatical voice that operate on the same argument structure consistently show information-structural differences. The major distinction in information structure is between topic ('what the utterance is about') and focus (the constituent that, given a well-formed question-answer pair, is the part of the answer that corresponds to the wh-word in the question). Topic-focus differences can be used to motivate the choice of one voice over another. It is erroneous however, to think that each and every language has voice alternations; for a given argument structure, some languages have just one way of mapping arguments to grammatical functions.

2. Active Voice

A verb in the active is often unmarked for any morphological indication of voice; however, this is not a universal characteristic. A number of languages, for instance, many Austronesian languages, mark active voice on the verb with a special morpheme. For

example, in Malagasy, a significant number of verbs form the active by prefixation on the verb stem (7a); for some of these verbs, the passive form is morphologically simplex (7b) (Keenan and Polinsky 1998):

- (7a) *sasa-* ‘wash’ (root)—*manasa* ‘wash’ (active, present)
 (7b) *hita* ‘see’ (passive)—*mihita* ‘see’ (active, present)

A common active voice structure is the one where the agent or most agent-like argument of the predicate is subject, while the theme, if there is one, is direct object. Agent appears in the subject case which, depending on a particular language, could be nominative, as in (8a, b), (11), absolutive (9a), ergative (9b), or dative (10). If the verb is transitive and has a theme argument, the case marking of this argument is typically that of direct object (accusative in an accusative language, absolutive in an ergative language); however, theme objects of some verbs can also be marked by a special case, inherent to a given verb, e.g., as in Russian (11).

- (8) Latin
 (a) *rex* *dēcessit*
 king.NOMINATIVE died
 (b) *rex* *omnes*
 king.NOMINATIVE (all
cōpias *castrīs* *ēducit*
 troops).ACCUSATIVE of.camp led
 ‘The king led all the troops out of the camp.’
 (9) Chukchi
 (a) *ətlʔa* *wiri-gʔi*
 mother.ABSOLUTIVE descend-AORIST:3sg
 ‘The mother came down.’
 (b) *ətlʔa-ta* *məčək-w-ən*
 mother-ERGATIVE shirt-ABSOLUTIVE
təni-nin
 sew-AORIST:3SG:3SG
 ‘The mother sewed the shirt.’
 (10) Hindi
mujʰ-e tum-se pyaar he
 I-DATIVE you-INSTRUMENTAL love AUXILIARY
 ‘I love you.’
 (11) Russian
Korolʹ upravljaet
 king. NOMINATIVE rules
stran-oj
 country-INSTRUMENTAL
 ‘The king rules the country.’

The determination of grammatical functions associated with agent and theme depends on language-internal factors and cannot be done on the basis of isolated examples from a given language.

If the verb in the active voice is intransitive, its single argument may correspond to either topic or focus; focused subjects of intransitives often serve to introduce new information. In a transitive active sentence, the subject/agent argument is more typically associated with topic than the object/theme, as shown

by a number of text counts (Mallinson and Blake 1981). However, this information-structural preference can be overridden by a number of contextual factors, and an argument other than agent or even an adjunct can be construed as topic.

3. Passive Voice

Natural languages use passives extensively and show a wide range of variation in the expression of passives. The variation can be captured by the following parameters: (a) difference in overt marking between the active and the passive predicate; (b) expression of the subject of the passive; (c) accessibility of arguments to passivization; (d) expression of the passive agent.

In terms of the encoding of the predicate in the passive construction, passives can be morphological (synthetic) or periphrastic (analytical). In English, the formation of the passive always requires an auxiliary (‘be’ or ‘get,’ with slight semantic differences depending on the choice of the auxiliary), thus the passive is periphrastic. In some languages, passives are always formed morphologically, for example, in Swahili, where the passive is marked by the suffix *-w-* (12b):

- (12) Swahili
 (a) *Juma a-li-vunj-a* *dirisha*
 Juma he-PAST-break-ASPECT window
 ‘Juma broke the window.’
 (b) *dirisha li-li-vunj-w-a* *(na Juma)*
 windowit-PAST-break-PASSIVE-ASPECT by Juma
 ‘The window was broken (by Juma)’

Although it is common for passives to be morphologically more complex than actives, this is not always the case; in a number of Austronesian languages, morphological passives are simpler than actives, for instance, for subsets of verbs in Toba Batak, Indonesian, or Malagasy; it is also possible for both active and passive to be formed by affixation from a verbal root (stem), as in Malagasy.

Some languages combine morphological and periphrastic passives, depending on the verb categories. For instance, in Latin or Russian, imperfective verbs form morphological passives, and perfectives use periphrastic passives (Latin *capio* ‘take’ (active), *capitur* ‘is taken’ (passive imperfect), *captus est* ‘was/has been taken’ (passive perfect). In addition to passives proper, some languages use the so-called reflexive passive, in which the active form of the verb combines with a reflexive clitic such as Romance *se*, *si*, German *sich*, Slavic *-sja*. Finally, some languages do not use any verb marking to distinguish active and passive predicates; for example, in Acehnese and in Mandarin, the only overt marking of the passive construction comes from the marking of the agent phrase and, optionally, from word order (Perlmutter and Postal 1983).

In terms of argument expression, the mapping of agent into an adjunct is the necessary and sufficient characteristic of passive. It is therefore possible that agent is ‘demoted’ to an adjunct and the subject position is filled by an ‘expletive’ (‘dummy’)—a meaningless element similar to the English ‘it’ in ‘It rains’ or ‘there’ in ‘There was a riot.’ Passives whose subject is an expletive are known as impersonal passives. Impersonal passives can be formed from intransitive verbs, (13b), or transitive verbs (14b). In the latter case, theme remains in the object position, and the subject of the sentence is expressed by an expletive; in languages where pronominal subjects are typically deleted, the expletive of an impersonal passive is a 3-person null pronoun (*pro*), whose presence is signaled by the agreement features on the verb (14b).

13 German

- (a) *Alle Kinder tanzen*
all children dance
(b) *Es wird (von allen Kindern)*
it is by all children
getanzt
danced.PERF.PARTICIPLE
‘There is dancing (as performed by all children).’

(14) Lithuanian

- (a) *čia lankutojai klausosi muzikos*
here visitors.NOMINATIVE listen music.ACCUSATIVE
‘Visitors listen to music here.’
(b) *čia pro lankutoju klausoma muzika*
here visitors.GENITIVE listen-PASSIVE-3 music.ACCUSATIVE
‘There occurs listening to the music by visitors here.’

In a personal passive, the subject position is occupied by a meaningful argument (as in English). In some languages, arguments other than direct object can also become subject of the passive. In some dialects of English, passivization is equally possible for both objects of the verbs ‘give’ or ‘send’:

- (15a) *The doctor gave the child new medicine*
(15b) *New medicine was given the child (by the doctor)*
(15c) *The child was given new medicine (by the doctor)*

In Kinyarwanda (Bantu), any prepositionless object can become subject of a passive (Kimenyi 1988, p. 365), thus:

- (16a) *umugóre aréerekeesha imashiini ábáana amashusho*
woman is showing machine children pictures
‘The woman is showing pictures to the children with a machine.’ (Active)

- (16b) *amashusho are’érekeeshwa imashiini ábáana*
pictures are being shown machine children (n’umugóre)
by woman
‘Pictures are shown to the children with machine (by the woman).’ (Theme passive)
(16c) *ábáana baréerekeeshwa imashiini amashusho*
children are being shown machine pictures (n’umugóre)
by woman
‘The children are shown pictures with a machine (by the woman).’ (Goal passive)
(16d) *imashiini iréerekeeshwa ábáana amashusho*
machine is being shown children pictures (n’umugóre)
by woman
‘The machine is being used to show pictures to the children (by the woman).’ (Instrument passive)

In languages that allow the passivization of multiple arguments, the accessibility to passivization is determined by the following principle: if a lower (more oblique) object can be passivized, then all objects higher than it should be able to passivize as well. This predicts that there should be no languages where a passive as in (16d) is possible whereas passives as in (16b, c) are ungrammatical.

In terms of the expressions of the agent, no language seems to require that the passive agent be expressed; however, there are languages where an overt by-phrase in the passive is ungrammatical (Arabic, Latvian, Seri).

There are several functional motivations for the use of passives, and although they do not need to co-occur within a single language, these motivations are cross-linguistically consistent. In terms of information structure, passive allows a speaker to indicate that theme, not agent is the main topic of an utterance. Concomitantly, passive may serve to indicate that the passive agent is focus or part of focus. Syntactically, passive is needed if a language has rules that operate only on subjects; in such a case, passive is a ‘way-station’ for an argument that needs to be made accessible to a given rule. For example, English uses passive in embedded infinitival clauses to put theme in the position of the silent subject argument of the infinitive (PRO), as in (17b), where John is subjected to the interview:

- (17a) *We persuaded John_i [PRO_i to interview the lawyer]*
(17b) *We persuaded John_i [PRO_i to be interviewed by the lawyer]*

Passive may differ from active in pragmatic connotation, for example, politeness or indirectness. Such pragmatic differences are often invoked as the main reason for the prevalence of passives in certain narrative styles and in requests (passive imperatives) in Austronesian languages. A recurrent interpretive

difference between the active and the passive is in that the passive emphasizes the change of state incurred by the referent of the theme. Because of this semantics of change of state and the attendant semantics of the resultant state, passive constructions are more common in the past and perfect; the coalescence of passive and perfect in the history of Indo-European languages is well documented (Benveniste 1960, Kuryłowicz 1964).

Linguistic theories differ in terms of whether the alternation between voices should be accounted for in syntax, in the lexicon or in both modules; for a good overview of grammatical approaches to passive, see Spencer (1991, Chap. 7).

4. Middle Voice

The description of the middle voice, based on the Greek grammatical tradition, is that of an event initiated within/by the main participant and having a principle effect on the same participant. For example, in Sanskrit:

- (18) Sanskrit
 (a) *so namati daṇḍam*
 he.NOMINATIVE bends.ACTIVE stick.ACCUSATIVE
 'He bends a stick.'
 (b) *daṇḍaḥ namate*
 stick.NOMINATIVE bend.MIDDLE
 'The stick bends.'

A common extension of that meaning is that the volitional and sentient main participant is acting in his own interests. For example in Sanskrit:

- (19) Sanskrit
 (a) *so kamaṁ karoti*
 he.NOMINATIVE mat.ACCUSATIVE makes.ACTIVE
 'He makes a mat.'
 (b) *so kamaṁ kurute*
 he.NOMINATIVE mat.ACCUSATIVE makes.MIDDLE
 'He makes a mat for himself.'

The middle voice typically appears in clauses referring to body actions, grooming, speech acts, cognition, emotive events, and meteorological or spontaneous events (Kemmer 1993). Since middles often refer to properties, not events, verbs in the middle voice commonly have a stative meaning, but this is not a necessary characterization of the middle voice. Cross-linguistically, middle voice often coincides or partially overlaps with the passive (hence such terms as medio-passive, e.g., in the history of Indo-European languages), as well as with the reflexive and reciprocal. This conflation can be explained by the commonality of semantics—in the middle, reflexive, and reciprocal, the same participant is perceived as both the cause (source) of an event and as its recipient, thus as agent and theme.

5. Antipassive Voice

The antipassive voice is less common than the passive; in those languages where it is found, variation is observed in the following ways: (a) the marking of the antipassive voice on the verb; (b) the encoding of the subject (in contrast with the subject of the active); and (c) the expression of the theme.

The verb marking, as in the other voices, can be analytical or synthetic. Analytical antipassives (formed with an auxiliary verb) are found in some languages of the Caucasus, for instance in Lak (20b); synthetic antipassives appear to be more common, for example, (21b) (in (21a), where the verb is active, it agrees with both ergative subject and absolutive object; in (21b), the verb in the antipassive form agrees only with the absolutive subject):

- (20) Lak
 (a) *ga-nal uṣṛcu*
 he-ERGATIVE children.ABSOLUTIVE
buvcu-n-i
 carry-AORIST-3SG
 'He has carried the children.'
 (b) *ga (uṣṛcu)*
 he.ABSOLUTIVE children.ABSOLUTIVE
buclej ur
 carrying is.3SG
 'He is carrying children; he is doing the carrying.'
 (21) Chukchi
 (a) *əlpa-ta məčək-w-ən*
 mother-ERGATIVE shirt-ABSOLUTIVE
təni-nin
 sew-AORIST:3SG:3SG
 'The mother sewed the shirt.'
 (b) *əlpa (məčək-w-ək)*
 mother.ABSOLUTIVE shirt.LOCATIVE
ine-nni-gʔi
 ANTIPASSIVE-sew-AOR:3SG
 'The mother sewed (at the shirt).'

As the contrast between (21a) and (21b) shows, in an ergative language, the subject in the active and the subject in the antipassive may appear in different cases; however, this is not a necessary condition on the antipassive, and ergative languages are attested where the subject of the antipassive remains in the ergative case (Australian languages Djarru, Walbiri, Kuniyanti). The theme in the antipassive construction is syntactically optional, thus it is an adjunct. It can be expressed by any of the oblique cases or adpositional phrases available in a given language (for example, in Chukchi, it can appear in the dative, locative, or instrumental case), and it can be freely omitted. However, as (20b) shows, in some languages the case marking of the theme does not have to change; what is crucial though is that the theme be an adjunct syntactically, which may be manifested though 'behavioral' properties of an NP, not case marking. Finally, in

some languages, antipassives require that the theme not be expressed (for example, in Tzotzil), in ‘mirror image’ to the non-expression of the passive agent in passives of some languages.

The sources of functional motivation of the antipassive are similar to the sources observed for the passive. Syntactically, the antipassive is used when a language has rules that refer to absolutes. For example, in Chukchi only absolute arguments are accessible to relativization, thus, when an ergative subject has to be relativized, the antipassive is used as a ‘way-station’ for relativization:

- (22) Chukchi
- (a) *ətlʃa-ta məčəkʷ-ən*
 mother-ERGATIVE shirt-ABSOLUTIVE
təni-nin
 sew-AORIST:3SG:3SG
 ‘The mother sewed the shirt.’
- (b) *[*məčəkʷ-ən təni-lʔən*] *ətlʔa*
 shirt-ABSOLUTIVE sew-PARTICIPLE mother
 (c) [*məčəkʷ-ək ine-nni-lʔən*]
 shirt-LOCATIVE ANTIPASSIVE-sew-PARTICIPLE
ətlʔa
 mother
 ‘the mother that sewed the shirt’

Interpretive differences between the active and the antipassive construction derive from the presence of an affected participant (theme) in the former and its absence in the latter. A common inference in the antipassive construction is that the theme is not affected or affected just partially. Since the theme participant does not even have to be specified, the event denoted by the antipassive predicate is often construed as generic or habitual, taking scope over a set of possible objects. This in turn explains why the antipassive is more commonly found in the imperfective aspect and in non-past tenses which do not limit an event or limit it indefinitely. In terms of information structure, the antipassive shares with the active that agent is the most likely topic; unlike the active construction, the antipassive makes it more likely that sentential focus takes scope over the predicate and adjunct (if any) or just the adjunct.

Unlike passive, antipassive is less common across languages, and even in those languages that have an antipassive, it is often considered to be a lexical, not a syntactic process, because of its restricted productivity. A syntactic account of the antipassive has been offered by Baker (1988) in terms of incorporation: the object position in the antipassive construction is occupied by an NP which is morphologically realized as a bound affix (including null allomorphs), and this affix incorporated into the verb stem.

6. Phenomena Related to Voice

A number of phenomena overlap with voice or are encoded by some of the same markers as are used for

marking voice, in particular the passive. These phenomena include reflexives (predicates whose two main arguments have the same referent, acting upon itself), reciprocals (predicates whose two main arguments refer to participants acting upon each other), inchoatives or anticausatives (intransitive counterparts of ‘cause X to P’, as in the English *burn*, *melt*) which receive the same marking as regular or reflexive passives (for example, in Romance languages). Inchoatives are sometimes assimilated to middle voice forms. Discussions of impersonal passives sometimes include active impersonal passives to these constructions is problematic.

- (23) French
on discipline ces enfants
 ‘One disciplines these children.’

The active-passive alternation is sometimes extended to include the so-called inverse systems attested mainly in Native American languages. In an inverse system, there are two constructions differing in whether Agent or Theme is higher on the hierarchy Human > Animate > Inanimate (or such similar hierarchy). If Agent outranks Theme on the hierarchy pertinent to a given language, the direct form of the verb is used (24a); if Agent is lower, the inverse form is used (24b).

- (24) Navaho
- (a) *hastin líí yi-z-tal*
 man horse 3SG + DIRECT-3SG-kicked
 ‘The man kicked the horse.’
- (b) *hastin líí yi-z-tal*
 man horse 3SG + INVERSE-3SG-kicked
 ‘The horse kicked the man.’

Aside from the presence of an actual alternation, there is no structural similarity between a voice system and an inverse system.

Early descriptions of ergative case-marking systems (see (9) above) assimilated the ergative to the passive. Although historically some ergative systems can be shown to go back to passives (e.g., in Indo-Iranian or Polynesian languages), synchronically the ergative construction is active, not passive.

Finally, most languages have classes of verbs whose morphology is indicative of one voice but whose syntactic behaviour does not support the morphological appearance. Such verbs are typically described as deponent verbs, a name rooted in the Latin grammatical tradition where the term was applied to verbs with passive morphology (*vox passiva*) which functioned as active verbs (*vis activa*), for example, *loquor* ‘speak’, where the inflection *-or* is that of 1 person singular passive.

See also: Morphological Case in Linguistics; Syntax; Syntax–Semantics Interface; Valency and Argument Structure in Syntax

Bibliography

- Baker M 1988 *Incorporation*. University of Chicago Press, Chicago
- Benveniste E 1960 'Etre' et 'avoir' dans leurs fonctions linguistiques. *Bulletin de la société de linguistique de Paris* 55: 113–34
- Dowty D 1991 Thematic proto-roles and argument selection. *Language* 67: 547–619
- Gruber J S 1965 Studies in lexical relations. Ph.D. dissertation, MIT
- Fillmore Ch 1968 The case for case. In: Bach E, Harms R (eds.) *Universals in Linguistic Theory*. Holt, Rinehard and Winston, New York, pp. 1–88
- Keenan E L, Polinsky M 1998 Malagasy. In: Spencer A, Zwicky A (eds.) *Handbook of Morphology*. Blackwell, Oxford, UK, pp. 563–623
- Kemmer S 1993 *The Middle Voice*. Benjamins, Amsterdam
- Kimenyi A 1988 Passives in Kinyarwanda. In: Shibatani M (ed.) *Passive and Voice*. Benjamins, Amsterdam, pp. 355–86
- Kurylowicz J 1964 *The Inflectional Categories of Indo-European*. Carl Winter, Heidelberg, Germany
- Mallinson G, Blake B 1981 *Language Typology: Cross-linguistic Studies in Syntax*. North-Holland, Amsterdam
- Perlmutter D, Postal P 1983 Toward a universal characterization of passivization. In: Perlmutter D (ed.) *Studies in Relational Grammar I*. University of Chicago Press, Chicago, pp. 3–29
- Spencer A 1991 *Morphological Theory*. Blackwell, Oxford, UK

M. Polinsky

Grammaticalization

1. Definition of Grammaticalization

Many definitions of grammaticalization have been suggested according to various theoretical approaches. One of the most diffused and clearest is Kurylowicz' definition: 'Grammaticalization consists in the increase of the range of a morpheme advancing from a lexical to a grammatical or from a less grammatical to a more grammatical status', Kurylowicz (1975). Some examples of grammaticalization are: Latin *habere* from a full verb meaning 'to grasp, seize' developed to an auxiliary in periphrastic verbal constructions for the expression of the future (*dicere habet* 'he/she has to say > shall say') or the past (*dictum habet* 'he/she has said'). The prepositional + relative phrase of the Old English *þa hwile þe* 'at the time that' evolved to the adversative and even concessive meaning of the Modern English conjunction *while* (see Traugott and König 1991). The Latin demonstrative *ille* 'that,' with strong deictic function (*ille homo* 'that man,' *homo ille* 'that notorious, famous man') gave rise to the obligatory article of the Romance languages (French *l'homme*, Spanish *el hombre*), where its function is simply to mark the noun phrase. Grammaticalization is thus a process leading from lexemes to grammatical formatives: 'a sign is grammaticalized to

Table 1

Six aspects of grammaticalization

	Paradigmatic level	Syntagmatic level
Weight	Integrity	Scope
Cohesion	Paradigmaticity	Bondedness
Variability	Paradigmatic variability	Syntagmatic variability

the extent that it is devoid of concrete lexical meaning and takes part in obligatory grammatical rules.' (Lehmann 1995, p. 6). Sometimes the terms 'grammaticization' or 'grammatization' have been used instead of grammaticalization without any substantial difference.

As is shown by the example of Latin *ille*, grammaticalization may lead to the emergence of new grammatical categories (in the case in point, the article which did not exist in Latin). In all the adduced examples the grammatical output of the process has passed through a syntactic use of the lexeme—though this need not always be the case.

Lehmann (1995, p. 123) has suggested three relevant aspects of grammaticalization, which realize six parameters both at the paradigmatic and syntagmatic level (see Table 1).

While grammaticalization is a process, the six parameters are not processes but properties that linguistic forms possess in increasing or decreasing measure. A linguistic form shows integrity inasmuch as it maintains its phonological and semantic identity. Phonological erosion and semantic bleaching reduce the integrity of a form (Latin *ille* > Italian *il*, Spanish *el*, etc.). The scope is the structural size of a construction: when a full verb is grammaticalized as auxiliary it makes up one phrase with the main verb and thereby reduces its scope to just the main verb (see the example of *habeo cultellum inventum* 'I have found a knife' in Sect. 3.1). By paradigmatic cohesion is meant the formal and semantic integration of a form into a paradigm: the suppletive forms *better*, *best* are less paradigm integrated than *longer*, *longest*. Bondedness refers to the syntagmatic cohesion: the more a form is grammaticalized the more it will be syntagmatically constrained; French *le*, Spanish *el*, etc., are fully grammaticalized noun determiners which cannot appear in isolation. Paradigmatic variability refers, according to Lehmann, to the freedom with which the language user chooses a sign—and, again, variability decreases with increasing grammaticalization, if cases are weakly grammaticalized the case suffixes may vary with relative freedom:

(1) Hungarian

a gyerek az asztal-on *játsz-ik*
The child the table-SUPERESSIVE CASE plays
'the child is playing on the table'

but *asztal-nál* (adessive) and *asztallal* (< *asztal-val*, instrumental-comitative) would also be possible.

Another example of paradigmatic variability is: the use of Latin *ille* was not compulsory in pre- or postnominal position and was semantically governed, but in Romance languages it is in most cases impossible to have a noun without the article. Finally, the syntagmatic variability is the ease with which a form may be shifted in the linear order of a sentence: the more a form has undergone a grammaticalization process the less variable it will be; **I have a knife found* or **I have a found knife* are impossible in Modern English.

It has to be noticed that the processes alluded to under the three labels of weight, cohesion, and variability cut across different levels of language structure; therefore the same items (e.g., *ille*) turn out to be appropriate examples of different parameters (integrity, scope, etc.). Grammaticalization is a pervasive phenomenon whose effects may be seen at all linguistic levels. Consequently, 'There is a growing interest in integrating grammaticalization with theoretical work in descriptive and historical linguistics' (Giacalone Ramat and Hopper 1998, p. 1).

2. A Historical Sketch of the Research

The classical examples of *dicere habet* and *l'homme* go back to Antoine Meillet (1916), who probably used first the term grammaticalization—though the concept is implicitly referred to by some previous linguists such as Wilhelm von Humboldt (*Über das Entstehen der grammatischen Formen und ihren Einfluß auf die Ideenentwicklung*, 1825), William Dwight Whitney (*The Life and Growth of Language: An Outline of Linguistic Science*, 1875), Georg von der Gabelentz (*Die Sprachwissenschaft: Ihre Aufgaben, Methoden und bisherigen Ergebnisse*, 1901) and others (see Lehmann 1995, Chap. 1, Heine et al. 1991, pp. 5–11).

As can be seen from the titles just quoted, grammaticalization has from the very beginning been considered in a large theoretical frame as one of the most characteristic developments languages may undergo. This holds also for the linguists of the next generations. Edward Sapir (1921, p. 102), for one, observed that 'it is possible for a concrete concept, represented by a simple word, to lose its material significance entirely and pass over directly into the relational sphere [i.e., the expression of abstract morphosyntactic relations, PR] without at the same time losing its independence as a word': forms live longer than their own conceptual content, which may be bleached and eventually fade completely. Sapir's examples are the Chinese and Cambodian verbs for 'give' used in a very abstract sense to express the 'indirect object' relation (see also Sect. 3).

Though not using the term grammaticalization, the French linguist Emile Benveniste (1968) took up again the case of *habere*, and called it a conservative mutation whereby a new (periphrastic) form takes

over the function of an older one in the grammatical paradigm (e.g., *habeo dictum* 'I have said' instead of *dixi* 'I said').

3. The Present State of the Discussion about Grammaticalization

3.1 The Continuum of the Grammaticalization Process

Already Whitney and Benveniste have underlined the fact that linguistic changes do not happen abruptly: 'Present possession often implies past action: *habeo cultellum inventum*, (...) *habeo digitum vulneratum* 'I possess my knife (recovered after loss) [> 'I have found the/my/a knife,' PR], (...) 'I have a wounded finger' [> 'I have wounded my finger,' PR].' Thus, *I have found the knife* became the indicator of 'a peculiar variety of past action contemplated as completed,' (Whitney 1875, p. 91, cf. Heine et al. 1991, p. 7f, Benveniste gives an insightful philological description of how the periphrastic forms with *habere* came first to be used and later spread out.

As a matter of fact, grammaticalization, like all terms ending in *-ization*, alludes to a gradual process, not a state. Grammaticalization is different from grammar. Even what has recently been alluded to as 'synchronic grammaticalization' implies a process whose final stage cannot be simultaneous with the beginning moment. This is not in keeping with the assumption by generative transformational grammar of rigidly separated grammatical categories. Actually, the 'new trend' of grammaticalization studies which has flourished in linguistics since the 1980s is basically oriented toward a functionalist approach, with very few exceptions. Grammatical categories are not seen as waterproof boxes; they permit, on the contrary, the gradual leaking of forms from one category to another.

Being functionally oriented means that attention is paid to the internal mechanisms, which govern linguistic structures, and the external psychological/cognitive causes of grammaticalization. Several linguists have hinted at the fact that grammaticalization very often originates in discourse strategies (i.e., in pragmatics), evolving to syntactic structures and later obligatory morphological markers (up to a possible total erosion of the markers):

Discourse < Syntax < Morphology < Morphophonemics < Zero (see Givón 1979, p. 209, cf. Lehmann 1995, p. 13, Traugott and Heine 1991, p. 3, Bybee et al. 1994, p. 40f).

Habeo cultellum inventum originates at the discourse level: the speaker wants to underline that he does really possess (*habet*) a knife which has previously been found (*inventum*) and uses a construct where every word keeps its own value. This pragmatic, effect-

Table 2

The grammaticalization–degrammaticalization continuum; α represents the evolution towards grammar (grammaticalization) and β the opposite process of degrammaticalization

α . lexicon	→delexicalization	→morphologization	→morphology
full words:			full morphemes:
<i>habet</i>		<i>di(ce)re (h)a(be)t</i>	<i>dir-a</i>
β . lexicon	←lexicalization	←demorphologization	←morphology
full words:			full morphemes:
<i>ade</i>		<i>lemon-ade</i>	– <i>ade</i> (<i>lemon</i> + <i>ade</i>)
<i>perhaps</i>			<i>per</i> (+ <i>haps</i>)

loaded expression evolves to a standard analytic construction of the syntax (with inversion and no longer with syntagmatic variability (see Sect. 1 above): *habeo inventum cultellum*) and finally enters the morphological paradigm of the verb *invenire* ‘to find’ as its periphrastic (agglutinated) past form (note that at this stage of the Romance languages it is no longer possible to have the word order *habeo cultellum inventum* (**fai le couteau trouvé*): the auxiliary has agglutinated to the main verb). As an example of morphophonemic evolution one can quote the French future *dirai*, *dir-a*, etc. ‘I shall say,’ ‘s/he shall say’ from periphrastic future forms with *habeo*: *dicere habeo*, *dicere habet*: the auxiliary has undergone phonetic attrition and has become an inflectional ending.

There is a continuum between the phases of the grammaticalization process: Heine et al. (1991, p. 1) quote the case of Èwe (Kwa branch of the Niger-Congo family):

- (2) *me-ná ga kofi*
1Sg-give money Kofi
‘I gave Kofi money.’
- (3) *me-ple βstrú ná kofi*
1Sg-buy door give Kofi
a ‘I bought a door and gave it to Kofi.’
b ‘I bought a door for Kofi.’
- (4) *me-wɔ dʒ vévié ná dodókɔ lá*
1SG-do work hard give exam DEF
‘I worked hard for the exam.’

In (2) *ná* has the function of a full verb, meaning ‘give’; in (4) *ná* can be interpreted only as a preposition (‘for’), whereas in (3) it is ambiguous and allows two different readings. Should one day the Èwe language admit for (3) only the second reading, then we would have a categorial shift (VERB > PREPOSITION), much in the sense of the evolution of Old English *hwil* ‘time’ (NOUN) to *while* (CONJUNCTION) (see above, Sect. 1). At the end of an uninterrupted process there is really a break and a category shift.

Shifts like VERB > PREPOSITION, NOUN > CONJUNCTION are examples of so-called *transcategorization*. Items belonging to a main (and open) category (VERB, NOUN) pass to a functional (and restricted) category (e.g., ADPOSITION PRE-OR PREPOSITION according to the structure of the language, concerned, CONJUNCTION):

these examples clearly hint at a move from a less grammatical to a more grammatical categorial status—it is however clear that the term *transcategorization* (or *recategorization*) is more appropriate than *decategor(ial)ization* used by Hopper (1991, p. 30f) and others.

3.2 The Dynamic of the Grammaticalization Processes

In such a dynamic typology of grammaticalization it is possible to account also for interrupted grammaticalization processes: in Old Italian it was possible to use movement verbs (*andare* ‘go’ and *venire* ‘come’) as auxiliaries in far more cases than in Modern Italian (cf. Giacalone Ramat and Hopper 1998, p. 2f, with further references). Further, Hopper has drawn attention to what he calls ‘emergent grammar,’ i.e., to the incipient stages of grammaticalization at the level of discourse (see above): the negative French adverb *pas* meant originally ‘step’; therefore it was used first with movement verbs (‘he doesn’t go a step’), but later it spread to verbs such as ‘eat,’ ‘see,’ etc. and became the general mark of negation, ousting other nouns that denoted a least quantity and were used, like *pas*, to reinforce the negation (e.g., *mie* ‘crumb,’ *gote* ‘drop,’ etc.): specialization principle. Thereby the original meaning of *pas* was completely lost (bleaching process).

As noted by Hopper himself, the principles he suggests (specialization, persistence, divergence, etc.) characterize linguistic change in general and are not distinctive for grammaticalization (Hopper 1991, p. 21). This has raised the problem of defining the limits of grammaticalization vs. other dynamic linguistic processes (cf. Giacalone Ramat and Hopper 1998).

3.3 Unidirectionality of Grammaticalization?

Unidirectionality is a much debated question in the literature on grammaticalization. Many linguists maintain that the evolution towards grammar cannot

be reversed (Lehmann 1995, p. 19, Traugott and Heine 1991, p. 4f, Heine et al. 1991, p. 31f). This is true if by this it is meant that once a lexeme has entered the grammar, it will hardly come back and assume again a lexical value. But there are many instances of items or morphemes leaving the realm of grammar and assuming an autonomous value as lexical entries: Hungarian, Spanish, and Italian have a large number of present participles used as nouns, e.g., Hungarian *költő* 'poet' (< *költ* 'to compose'), Spanish and Italian *cantante* 'singer' (< *cantar(e)* 'to sing'); nobody can fail to recognize in French *seigneur* (> English *sir*) the comparative of the Latin adjective *senex* 'old' or in English *elder* 'official in the Presbyterian church' a comparative from Old English *ald* 'old.' Both were originally forms belonging to the grammatical paradigm of the adjectives, but later they developed an autonomous lexical meaning. Even syntactic constructs may develop to new lexemes: Dutch *misschien* 'perhaps,' (< (*het*) *mach schien* literally 'it may happen'), Italian *forse* 'perhaps' (< Latin *Fors sit (an)* literally 'be the case'), English *perhaps* literally 'per chances (plural).' Grammaticalization and lexicalization are two complementary processes (Moreno Cabrera 1998, Ramat 1992).

Degrammaticalization processes are certainly less frequent than grammaticalization processes, but they exist and have to be explained. The causes for their existence are to be sought in the overall tendency to use labels as economic symbols: e.g., suffix morphemes like *-ade* (from *orangeade*, *lemonade*, or even *gatorade*) can be separately manipulated and become autonomous lexemes: the label *ade* with the general meaning of 'fruit juice' is more economic than the phrase *fruit juice*, though not as transparent (see *bus*, in *schoolbus*, *autobus*, etc., as hyperonym for 'public transportation vehicle'). Economic symbolism and noneconomic iconism are conflicting forces which are at work in every natural human language (cf. Ramat 1992, p. 557, Newmeyer 1994, p. 69). If we had just unidirectional evolution toward grammar we would expect that languages become more and more grammaticalized, which by all evidence is not the case (see Moreno Cabrera 1998, p. 224). On the contrary, we have a double evolution towards grammar as well as towards lexicon (see Kurylowicz 1975).

Phonogenesis is also a process which shows that loss of grammatical function may give rise to new lexemes (cf. Hopper 1990, 1992). As the German grammatical prefix *ge-*, used to indicate collectives (see *Ge-beine* 'limbs,' plural, *Ge-dränge* 'crowd'), was no longer felt as a prefix, it underwent phonetic erosion to *g-* and gave rise to nouns such as *Glück* 'luck,' *Glied* 'limb,' *Gnade* 'grace' which synchronically are no longer analyzable: phonetic junk, no longer recognizable as a functional morpheme, produced new lexemes. In the same vein of reasoning we have to consider also remnants of disappeared paradigms: this applies to dual forms which survived the loss of the dual

declension: cf. Latin *ambo* 'both' (perhaps also *octo* 'eight') with the old dual marker *-o*.

Summing up, we may represent (see Table 2) the grammaticalization and degrammaticalization processes along a continuum which moves into two opposite directions. At both poles of the scale we have either clear lexemes (full words like *habet*, *perhaps*) or morphemes (inflectional *-a* or derivational *-ade*), while it is difficult to state when *habet* was delexicalized and the construct *dicere habet* became the norm to express the future, or *ade* was reanalyzed as a second member of a compound: that is why *di(ce)re (h)a(be)t* and *lemon-ade* appear in the previous scheme midway between (de)lexicalization and (de)morphologization. Transcategorization is a gradual process along a continuum; its final output represents a break in the categorial status of the input (see Sect. 3.1).

Bibliography

- Benveniste E 1968 Mutations of linguistic categories. In: Lehmann W P, Malkiel Y (eds.) *Directions for Historical Linguistics: A Symposium*. University of Texas Press, Austin, TX, pp. 83–94
- Bybee J, Perkins R, Pagliuca W 1994 *The Evolution of Grammar, Tense, Aspect, and Modality in the Languages of the World*. University of Chicago Press, Chicago
- Giacalone Ramat A, Hopper P (eds.) 1998 *The Limits of Grammaticalization*. Benjamins, Amsterdam
- Givón T 1979 *On Understanding Grammar*. Academic Press, New York
- Heine B, Claudi U, Hünnemeier F 1991 *Grammaticalization. A Conceptual Framework*. University of Chicago Press, Chicago
- Hopper P 1987 Emergent grammar. *Berkeley Linguistic Society* 13: 139–57
- Hopper P 1991 On some principles of grammaticization. In: Traugott E, Heine B (eds.) *Approaches to Grammaticalization*. Benjamin, Amsterdam Vol. 1, pp. 17–35
- Hopper P 1992 Phonogenesis. In: Pagliuca P (ed.) *Perspectives on Grammaticalization*. Benjamins, Amsterdam, pp. 119–35
- Kurylowicz J 1975 The evolution of grammatical categories. In: J K (ed.) *Esquisses Linguistiques* Vol. 2. Fink, Munich, Germany, pp. 38–54
- Lehmann Chr 1995 *Thoughts on Grammaticalization. A Programmatic Sketch*. (Arbeiten des Kölner Universalien-Projekts Vol. 48). Lincom Europa, Munich, Germany
- Meillet A 1948 L'évolution des formes grammaticales. In: A M (ed.) *Linguistique Historique et Linguistique Générale*. Champion, Paris, pp. 130–48
- Moreno Cabrera J C 1998 On the relationships between grammaticalization and lexicalization. In: Giacalone Ramat A, Hopper P (eds.) *The Limits of Grammaticalization*. Benjamins, Amsterdam, pp. 211–27
- Newmeyer Fr J 1994 Competing motivations and synchronic analysis. *Sprachtypologie und Universalienforschung* 47: 67–77
- Ramat P 1991 Thoughts on degrammaticalization. *Linguistics* 30: 549–60
- Traugott E C, Heine B (eds.) 1991 *Approaches to Grammaticalization*. Benjamins, Amsterdam

Traugott E C, König E 1991 On the expression of spatio-temporal relations. In: Greenberg J, Ferguson Ch, Moravcsik E (eds.) *Universals of Human Language*. Stanford University Press, Stanford, CA, Vol. 3, pp. 369–400

P. Ramat

Gramsci, Antonio (1891–1937)

Antonio Gramsci (born in Ales, Sardinia on January 22, 1891 and died on April 27, 1937) started out as a freelance journalist writing assiduously for *Il Grido del Popolo* and *Avanti!*, newspapers of the Italian Socialist Party in whose ranks he had been active since 1914. In 1919 he founded *L'Ordine Nuovo* which became the daily paper of the Turin Communists, of whom Gramsci was to become undisputed leader. In 1921 he joined the Communist Party of Italy, created from the walkout during the XVIII Socialist Congress in Livorno. The following year he was in Moscow representing the Italian Communists in the Third International. Returning to Italy in 1924 he founded the newspaper *L'Unità*. Two years later he participated in the III Congress of the Italian Communists, held in Lyon, where he opposed Amadeo Bordiga and obtained an overwhelming majority. The same year, despite his parliamentary immunity, he was arrested by the Fascists. Given a summary trial, he was sentenced to 20 years of imprisonment by the special Court for the Defence of the State.

Taken to the prison of Turi in the province of Bari, he remained there until 1933 when he was moved to Formia due to poor health. Obtaining a conditional discharge, in 1935 he was admitted to the Quisisana hospital in Rome, where he died following a stroke. Thanks to his sister-in-law Tatiana, his *Quaderni del carcere* ('Prison Notebooks'), written between 1930 and 1935, were sent underground to Moscow and delivered to Palmiro Togliatti. Published after the collapse of the Fascist dictatorship, they revealed a leading theoretician who, albeit following the path of Marxist–Leninist tradition, had elaborated an original concept of society and history underpinned by the theory of hegemony.

Gramsci's political thinking was deeply influenced by the October Revolution. He immediately saw it as confirming his activist and voluntaristic reading of Marxism, induced by Giovanni Gentile's neo-idealism. For Gramsci, Marxism was something more than a simple, empirical interpretation of history, as Benedetto Croce had defined it; it was a complete philosophy containing, in nuce, a new civilization to be constructed over a long and dramatic process. The industrial proletariat or, more specifically, its conscious vanguard, the Communist party, would be its protagonists. So Gramsci enthusiastically greeted the Bolshevik October as 'the revolution against the

Capital.' The revolution against the positivistic and deterministic interpretation of Marxism which the parties of the Second International had adopted while awaiting the inevitable, catastrophic collapse of capitalism. Gramsci also considered that Lenin had been able to grasp the systematic link between Marxism and German classical philosophy, of which Hegelianism had been the speculative summit. Thus Gramsci was prompted to see Lenin not only as the greatest strategist of the international workers' movement but also its greatest theoretician. In his deeds, even before his thoughts, he had expressed a higher degree of consciousness. Therefore, the unity of the theory and practice incarnated by Lenin and by the movement he had created had to be emphasized. Cultural work was thus a priority, to demonstrate that Communism was the positive response to the world crisis of capitalism and the moral dissolution of the bourgeois civilization, precisely because it had, thanks to Marxism, reached a higher degree of theoretical awareness. It was no coincidence that Bolshevism arose from the claim that the workers' consciousness should prevail over their spontaneity and that without a revolutionary theory no really revolutionary action was even conceivable. For Gramsci, the Communist revolution had to be seen as a philosophy which was being put into practice or, more precisely, as an ethical-political will enlightened by a theory whose aim was the reconstruction of the world.

Up to the end of 1923, Gramsci's thinking, expressed in a myriad of short articles on the political situation, was nothing else but an intelligent adaptation of Leninism to the Italian situation. All the typical elements of the Bolshevik model were to be found in it: the supremacy of theory over spontaneity, the concept of the party as a nucleus of activists devoted heart and soul to the revolution and submitted to iron discipline, the intransigent defense of orthodoxy against any form of revisionism, the military vision of the class war, the revolution as a declaration of war on the existing, the regime of transition focused on the dictatorship of the proletariat exercised by its conscious vanguard, the edification of Socialism as a substitution of the anarchy of the market with a single production and distribution plan; one can even find the idea, already expressed by Lenin to justify the red terror, that the Communists' historic mission was to 'purify the social environment with iron and fire,' exterminating the lower and middle bourgeoisie, 'corrupt, dissolute and putrescent humanity.' But, drafting his program-letter of February 1924, Gramsci started to work out the revolutionary theory that, in his *Quaderni del carcere*, was to be called the 'war of position' in opposition to the 'war of movement.'

Gramsci was the first Communist leader to realize that the Bolshevik strategy had few possibilities of success in Western Europe, with its robust civil society, rich in counter-powers and institutions relatively independent of the state. The Third International was

founded on the hypothesis that the world capitalist system had entered a phase of rapid and uncontrollable disintegration and that, therefore, the Communists' task would be that of accelerating this disintegration, mobilizing the proletarian masses against the bourgeois state, and enabling the Bolsheviks to take power. But the Comintern leaders had underestimated the vitality of capitalism that, despite the great postwar crisis, had shown it was strong enough to suffocate any revolutionary attempt before it was born. The 'war of position' was Gramsci's idea for rescuing the Communist movement from the blind alley down which it had gone. He had realized that the dominance of the bourgeoisie was based more on the consensus of the masses than on control of the coercive machinery. The celebrated theory of hegemony, developed in his *Quaderni del carcere*, says that the 'supremacy of a social group is shown in two ways, as dominance and as intellectual and moral leadership.'

Which is like saying that a class is hegemonic only insofar as it is capable of increasing the material and spiritual patrimony of society and obtaining the consensus of the working classes. Therefore there is no hegemony that is not founded on a 'historical bloc,' that is on an organized system of social alliances cemented by a common ideology. The historic bloc is the social basis of hegemony. It is not only a political phenomenon, it is also, and in a certain sense above all, a cultural phenomenon. This led Gramsci to claim that the superiority of a hegemonic class lasts insofar as it is able to keep its system of alliances united and bring in other social groups by satisfying, albeit within precise limits, their material and moral interests. Then, and only then, will it become the 'driving force of a universal expansion, of a development of all the national energies.' Gramsci's approach to the problem of social order was typical of the sociologies of consensus, stamped by a certain Marxism as interesting ideological distortions of reality, and he grafted their fundamental theorems on the trunk of historical materialism. He also took over Vilfredo Pareto and Gaetano Mosca's elitist theories of the late nineteenth and early twentieth centuries. His central premise was that the historical existence of a society is, always and necessarily, concentrated on the minorities–masses dialectic.

In *Quaderni del carcere*, society is conceived along Marxist lines as a hierarchical structure of classes, differing according to their position and function in organizing the production and reproduction of material life. But, for a class to make history, it must distinguish itself and become independent; it has to acquire a critical knowledge of itself. This happens gradually, transforming a class, an immediately corporate reality, into an ethical–political force with a national character that imposes its leadership on society. This is precisely the stage in which ideologies confront each other and enter into battle until one of them, or a combination of them, takes the upper hand

and unites not only the economic and political, but also the intellectual and moral, aims of society. Vice versa, a dominant class starts to decline when it gradually loses its ability to interpret the needs of the dominated classes and to mobilize national energies. It is true that a dominant class can keep its power even when it has lost its ability to rule, thanks to its control of the state's coercive machinery; but in the long run it is destined to be replaced by a new ruling class.

This led Gramsci to a conclusion which is, basically, the reformulation of Pareto's famous theory of revolution as a violent and traumatic substitution of an old aristocracy with a new aristocracy. But where Pareto only saw ideologies as rhetorical stratagems through which an elite justified its desire to dominate, Gramsci ascribed an enormously important function to them. Ideologies are not mere derivations, as Pareto called them; they are global visions of reality, essential for constructing a civilization. There is no civilization without an ideology, and without this ideology being absorbed by the lower classes and transformed into popular culture. So it can be said that history is dominated by the clash between ideologies, between hegemonic principles competing for the consensus of the masses. And the theatre of this clash is civil society. Gramsci saw this differently from Marx, that is as the society of industry and general competition, material base of the state and the ideological superstructures. For Gramsci, civil society is the theatre of intellectual and moral relations, where a social group that aspires to dominance constructs its own hegemony by soliciting the consensus of the dominated classes. Civil society, therefore, is a cultural rather than an economic reality: it is where rival hegemonic principles confront each other until one of them eliminates the others and becomes the impersonal engine of historical development.

Now, if what is decisive in history is not economic but spiritual power, then to move the middle classes out of the city of command, you first of all and above all have to earn the active consensus of the lower classes. That is, you have to incite a critical attitude to the existing order and persuade the workers to support the new type of society implicit in Marxist philosophy. The second consequence drawn from Gramsci's theory of hegemony was that the Bolsheviks' Jacobin strategy was not applicable to the type of war that had to be waged in Western societies. It presumed the power vacuum which had effectively occurred in Russia in 1917; in other words, it presumed the disintegration of the ruling class and the general collapse of the political system. Lenin and Trotsky's Blitzkrieg tactics had not, therefore, any possibility of success if there was a sufficiently cohesive historic bloc to repel frontal assault by the armies of the revolutionary movement. It was necessary to change the line of action, to work out a long-term strategy for gaining power.

The Italian Communist leader's answer to this problem is to be found in his *Quaderni del carcere*.

Studying the history of Western civilisation, Gramsci found a perfect model in the victory of Christianity over the Greco-Roman culture. Christianity had been, in its way, a revolutionary movement that had introduced a spiritual scission in ancient society and had triumphed thanks to its intellectual and moral superiority, to the indefatigable teaching of its apologists and to its organization. If it was impossible to conquer the bourgeois citadel *manu militati*, it was possible, as the history of the expansion of Christianity indicated, to conquer it by wearing away the ideological defenses of the middle classes. Thus the Communists had to imitate the early Christians by eroding from within, by a methodical, silent and all-embracing conquest of the ideological superstructures and the existing order until the working classes stopped supporting the ruling class. One of the fundamental theorems of Gramsci's sociological theory was that a class which wants to win power has first of all to be hegemonic. That is, it has to be able to exercise a firm and united, intellectual and moral leadership over the working masses and then to dominate them. It has methodically to take possession of the 'fortresses' and 'casemates' of civil society and transform them into the operational bases of its march towards the city of command. Thus, whereas Leninism had theorized the occupation of society through the violent conquest of the state, Gramsci proposes an inverse procedure: the conquest of the state through the cultural occupation of society. To this end, it was all-important to penetrate the mass-media, schools, universities and publishing houses and convert them into agencies for the socialization of Marxism.

In Gramsci's model for the conquest of power, violence was relegated to the background. Nonetheless, the strategic goal remained that of the Bolshevik Revolution: the institutionalization of the mono-party dictatorship as an indispensable condition for totally remodeling the social body. On this point Gramsci remained faithful to Lenin's teaching. The party, and only the party, had to be the leading light of the intellectual and moral reform for creating the new civilization whose institutional foundations had been laid in the Soviet Union. So that Gramsci, after having indicated the Communist party as the only institution which could be compared to the Catholic church, did not hesitate to define Marxism as the 'religion which could kill Christianity' by building the City of Man. And he saw the modern prince (the party of the professional revolutionaries, conscious vanguard of the working classes) as the historical agent. He who, overthrowing the entire intellectual and moral system of liberal civilization with his impassioned teaching, would have 'to take, in people's minds, the place of the divinity and the categorical imperative' (*Quaderni del carcere*).

Gramsci's is a rigorous theorization of the charismatic nature of the revolutionary party, not unlike Gyorgy Lukacs' in *Geschichte und Klassenbewusstsein*, where the Communist party is described as the

cosmic-historic agent which, in virtue of its superior theoretical knowledge, will have to achieve the 'salvation of humanity.' Gramsci defines the modern prince as the 'focus of the faith' and the 'custodian of the doctrine of Socialism' destined to introduce an epochal caesura in universal history. This prince has to ensure that he becomes the source of all the satisfaction his militants had previously found in a multiplicity of organizations. He must also destroy all these other organizations or incorporate them into a system in which this modern prince will be the 'only regulator.' Hence the distinction, spelt out in *Quaderni del carcere*, between two types of totalitarianism: the regressive kind, with a party opposing the drive of history, and the progressive one, which is precisely that of the modern prince bringing a new and superior culture.

This led Gramsci to affirm peremptorily that 'everything is politics,' and that, consequently, it is not possible to separate politics from philosophy, since the latter is the driving force of history in the making. A history that is always 'between two hegemonic principles, between two religions,' each of which aspires to regulate human life in all its material and spiritual manifestations. The idea of a competitive coexistence between a plurality of social forces and cultural models, so typical of liberal civilization, is not only entirely extraneous to Gramsci's political philosophy but even explicitly rejected. The final aim of his philosophy is the construction of the future city, conceived by Marxism as a 'harmonious society.'

The role of the intellectuals is of fundamental importance within this vision of the Communist revolution. They are the protagonists of history, since history is nothing else but the dramatic process through which a determinate philosophy takes over the world. Gramsci sees the revolutionary party as a 'collective intellectual' and gives the 'organized intellectuals' framed within the rigid discipline of Leninist centralism the function of 'permanent persuaders.' This explains why Gramsci's philosophy, in the decades following the publication of the Khrushchev Report (1956), had such a powerful influence on the Communist intelligentsia. The latter found in the *Quaderni del carcere* what they needed to liberate Marxism from Stalinist dogmatism and to conceive the fight for socialism as a magnificent pedagogical operation to conquer spiritual hegemony by changing how the subordinate classes felt and thought. But after the crisis of Marxism in the 1980s, Gramsci's theory of the peaceful conquest of power rapidly lost its fascination, despite the fact that some basic concepts of this theory—and first of all the concept of hegemony—have entered the lexicon of contemporary sociology.

See also: Bourgeoisie/Middle Classes, History of; Communism, History of; Hegemony: Cultural; Historiography and Historical Thought: Modern History

(Since the Eighteenth Century); History: Forms of Presentation, Discourses, and Functions; Marxism in Contemporary Sociology; Marxism/Leninism; Marxist Social Thought, History of; Power in Society; Power: Political; Russian Revolution, The; Social History; State, History of; Subaltern History; Violence, History of; Working Classes, History of

Bibliography

- Anderson P 1977 *The Antinomies of Antonio Gramsci*. London
 Buci-Glucksmann C 1975 *Gramsci et l'Etat*. Fayard, Paris
 Cammet J M 1967 *Antonio Gramsci and the Origins of Italian Communism*. Stanford University Press, Stanford, CA
 Femia J V 1981 *Gramsci's Political Thought*. Oxford University Press, Oxford, UK
 Finocchiaro M 1987 *Gramsci and the Methodology of Criticism*. Las Vegas
 Finocchiaro M 1999 *Beyond Right and Left*. New Haven
 Hobsbawm E J (ed.) 1995 *Gramsci in Europe and in America*. Bari
 Jocteau G C 1975 *Leggere Gramsci*. Feltrinelli, Milan
 Pellicani L 1981 *Gramsci: An Alternative Communism?* Hoover Institution Press, Stanford
 Pellicani L 1991 *Gramsci, Togliatti e il Pci*. Rome
 Portelli H 1973 *Gramsci et le Bloc Historique*, 1st edn. Presses universitaires de France, Paris
 Tamburrano G 1963 *Antonio Gramsci*. Lacaita, Manduria

L. Pellicani

Graphical Methods: Analysis

1. Introduction

Everyone is accustomed to seeing graphical displays used for presentation purposes: in scientific articles, newspapers, and magazines. Yet graphical displays may be used for exploratory purposes too, for investigating datasets and for uncovering the information hidden in them. Exploratory graphics are different. The aim is not to present information, but to use the graphics to uncover information. Static printed graphics can be useful for exploratory work, but the real gain lies in using interactive graphics on a computer.

In all the following examples it is assumed that exploratory analyses are being carried out, that each graphic is part of a broader analysis, and that they are drawn to aid the analyst, not to present the information to others. This is why there are few scales given. The analyst will know the approximate scaling and can always query the graphic on screen to get exact values, not only along axes, but for any point or object. In general a single presentation graphic is drawn to convey information to a potentially large

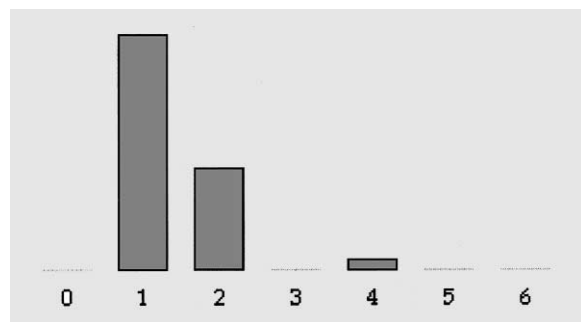


Figure 1

Bar chart of gender of car insurance policyholders

and varied audience, whereas a large and varied collection of exploratory graphics are drawn to extract information for a single person, the analyst.

Graphics are useful for checking basic assumptions and for getting a feel for data. Figure 1 shows a bar chart of gender from an insurance data set. Unusually there are seven codes for gender! Three made sense (firms are neither male nor female) and the rest were errors.

The histogram for policyholders' birth dates for the same data set was also revealing (Fig. 2). The tall white column to the left represents missing values for the variable (mainly firms again). There is a small isolated column at 1,900 (missings originally coded as '00?') and there are some very recent dates of birth, surely errors for car-driver insurance data. This is a large dataset (over 50,000 cases) and some errors may be expected, but it is extraordinary how often datasets which have been purportedly quality checked still have errors in them. Incidentally, the dip in the histogram is due to low birth rates during the war.

Assuming major errors have been corrected or removed, graphics are valuable both for checking old ideas and for generating new ones. After the historic 1990 Irish Presidential election it was hypothesized that Mary Robinson had won because she got the women's vote. Data from the last MRBI survey before the election suggests otherwise. Figure 3 shows a bar chart for the preferences for the three candidates linked to a spine plot for gender (in a spine plot all column heights are the same and column widths represent the counts so that highlighted proportions are directly comparable). The proportion of women intending to vote for Mary Robinson was only slightly higher than the proportion of men. Linking to another spine plot shows that urban voters were more important for Mary Robinson's win.

2. Background

Presentation graphics (see *Graphical Methods: Presentation*) have been used for a long time. Exploratory

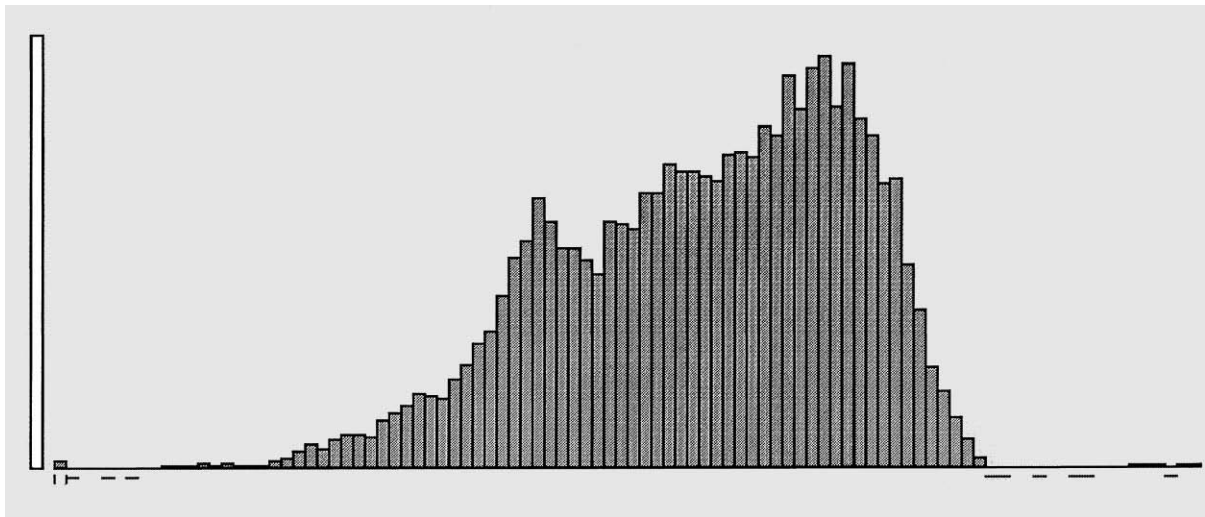


Figure 2
Histogram of years of birth of car insurance policyholders

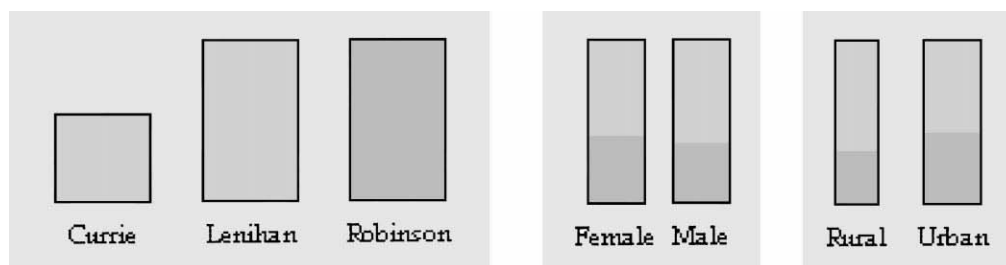


Figure 3
Barchart of candidates' support linked to spineplots for gender and location. (Results from the last MRBI survey before the 1990 Irish Presidential Election. Thanks to Jack Jones and the Market Research Bureau of Ireland for permission to use the data)

graphics are newer, not because people did not appreciate their value, but because without a computer it was not efficient to draw many variations of many different displays. John Tukey led the way with his innovative and provocative book, *Exploratory Data Analysis* (Tukey 1977; see *Exploratory Data Analysis: Univariate Methods*) although, curiously enough, the book was written slightly too early, before personal computers were becoming commonplace. Thus he emphasizes pencil and paper methods like stem and leaf plots for continuous data. These are like histograms laid on their side with each bin having a width of a power of 10. The advantages are the simple construction by hand, and the listing of each value individually. The disadvantages are the dependence on a very limited range of bin sizes, and the unsuitability

of the method for anything other than small datasets. Imagine a stem and leaf plot for the car insurance data set of Fig. 2.

More recent computer-oriented references include Cleveland and McGill (1988), Eick and Wills (1995), and Unwin (1999). There has also been much related work outside statistics for which Spence (2001) is a good source. In none of these works is a substantial formalization of exploratory graphics attempted. Considering that presentation graphics have been around much longer and yet Wilkinson's groundbreaking work (Wilkinson 1999) only appeared in 1999, this is not surprising. Furthermore, exploratory analyses are by their nature interactive and dynamic and so the process is difficult to explain on the printed page. Some of the best sources of information and examples

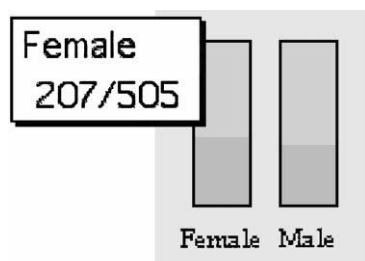


Figure 4

Direct querying of a bar. How many women supported Mary Robinson?

can be found on the web using keywords in this article and/or author names as search parameters.

3. Interacting with Displays

3.1 Interactive Features

Graphics may mislead. For instance, bar charts will look similar whether they are based on 50 or on 5,000 cases. (The survey in Figs. 3 and 4 was based on 1,000 interviews.) It is important to be able to check any impressions gained. Sometimes this might mean carrying out the corresponding statistical test (e.g., a chi-square test for comparing bar chart proportions), but a first step is to directly query the graphic, as in Fig. 4.

Direct querying is useful in all displays, especially when it is context sensitive. The same action may be used for each part of each plot and yet provide different information according to which part of which plot is queried. Other steps that may be carried out to gain more insight are to rescale the graphic (e.g., change the bin size in a histogram), to reformat the display (e.g., switch the axes of a scatter plot), to reorder the display (e.g., order the categories in a bar chart by count), to zoom in (e.g., concentrate on an outlying cluster of points in a scatter plot), or to select various groups of points and link them to their representation in other displays (see Unwin 1999).

Histogram bin sizes which are small pick out gaps and identify preferred values, while large bin sizes

emphasize overall trends. Figure 5 shows two histograms of maths marks for 126 Irish schoolchildren. Notice how one version picks out the typical gap below the pass mark of 40 and the other does not.

Scatter plots can be extremely informative tools for revealing the relationship between two variables, but the scales, the aspect ratio of the axes, the size and color of the points can all affect how they look. For large datasets with extensive overplotting (i.e., many points are drawn at the same location on the plot), some form of density estimation (see *Probability Density Estimation*) is necessary or there is a risk the dataset may be misunderstood. Figure 6 shows a scatter plot of the birth weights of 4,000 babies plotted against the estimated duration of pregnancy. A coarse density estimation has been used whereby black dots represent single cases and the brightest white ones at least 30 cases. It is clear where the majority of the data is concentrated.

Neither drawing sunflowers at each point with the numbers of petals representing the numbers of points nor jittering the data (adding small random components to each point to reduce the overlapping) are serious solutions for large datasets like this (although both suggestions are to be found in the literature). Sounder solutions have been suggested by several authors including Carr et al. (1987) and Hyndman (1996), but should be implemented interactively. Single displays optimized on a particular criterion cannot reveal all the structure in the data. An interesting problem for interactive scatter plots is how to display both highlighted and unhighlighted distributions informatively.

Variations of scatter plots may be used to display subgroup distributions in parallel, e.g., income of those surveyed by job category. The resulting parallel dot plots are a valuable initial exploration tool for smaller datasets, but parallel box plots are more effective for large datasets as in Fig. 7.

Bar charts are good for comparing absolute numbers, but are not good for displaying proportions. Simple pie charts are good for emphasizing shares, though it is difficult to compare the sizes of individual sectors and too often they are drawn with irrelevant three-dimensional features and other distracting distortions. Spine plots are best for comparing proportions based on selections, but can only deal with one

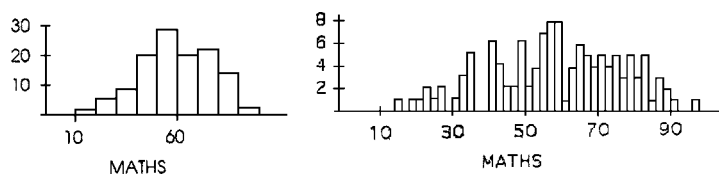


Figure 5

Histograms of maths marks out of 100 for a group of Irish schoolchildren

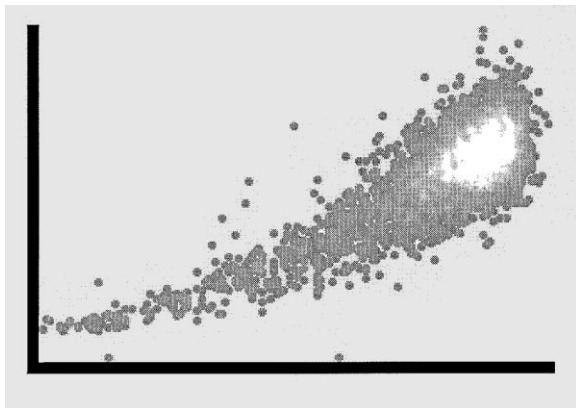


Figure 6
Babies' birthweights plotted against duration of pregnancy

selection at a time (see Fig. 3). Switching between bar chart and spine plot representations interactively is one solution.

3.2 Linking

Linking means that points selected in one display are highlighted in all other displays. Selecting a subgroup

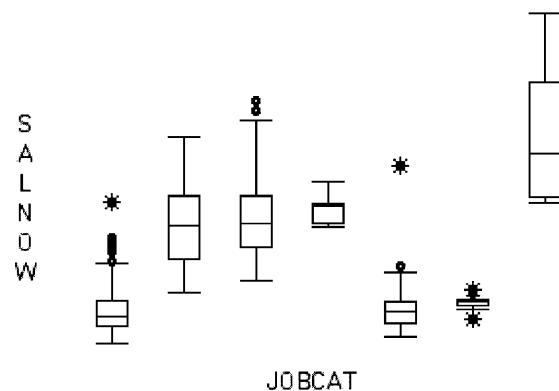


Figure 7
Boxplots of salaries by job category in a bank

may identify an outlier within the group in another plot. There may be patterns within subgroups which mask each other when the data for all groups are shown together. Highlighting area-based plots is straightforward; an area corresponding to the size of the selected group is colored. Highlighting point displays is complicated by overplotting, as mentioned above in relation to scatter plots. The best solution for box plots seems to be the idea due to Wills of drawing a box plot for all the data in the background and a box

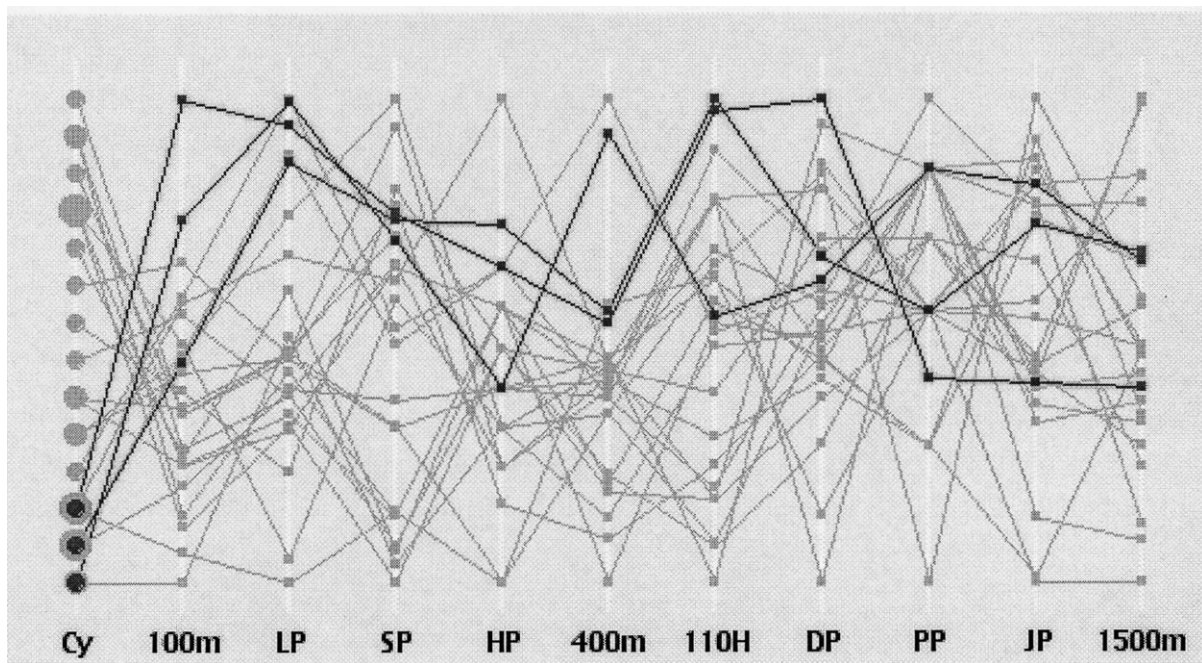


Figure 8
A parallel coordinate plot of the points by event for the mens' decathlon at the Sydney Olympics 2000. The medal winners have been highlighted

plot for the highlighted data on top (Eick and Wills 1995).

Linking is central to successful interactive graphics in the way it provides insight into multidimensional structure. Not just points in graphics may be linked, but cases, scales, orderings, and colors may be linked in any kind of window. In this sense querying is just a form of linking.

3.3 Interfaces

All of these tools are exploratory in nature and only work effectively when they are interactive. If you have to wait for the screen to clear and a graph to redraw, you will have forgotten why you wanted the change in the first place. Interactivity implies an intuitive interface, flexible commands, and a fast response. Command-line interfaces are valuable for the precise control they offer and the ease with which they can be recorded. For exploratory work it is direct manipulation of the objects on screen which is important and it is rarely of interest to know how a conclusion was first discovered, as it is common to find much better ways of showing results after you have first discovered them. All components of a graphic display (bars, axes, points, legend, etc.) may ideally be queried, moved, rescaled, and reformatted. To avoid clutter on screen, it is helpful to use cues to hint at what is possible rather than to fill the screen with visible controls. To protect the user from possible unwanted side effects of their actions, warnings, and undos are beneficial.

3.4 Dynamic Graphics

The term 'dynamic graphics' refers to graphical displays that move such as rotating plots, grand tour displays, or animations controlled by sliders. (A slider is a control which allows the user to change the value of a parameter of a model in real time and see the affected displays update continuously (see Data Desk or Tierney 1990). While the resulting dynamic displays may be informative, it is the interactive control in the hands of the user which is most useful and in this sense they are a specialist subset of interactive exploratory graphics.

4. Multivariate Displays

4.1 Continuous Variables

Modern graphical tools are very good for exploring many variables at once. Scatter plot matrices (a matrix display of scatter plots of all possible pairs of variables) have often been suggested for continuous variables, though they are inefficient in their use of space. For up to about eight variables (the number depending on the

screen's resolution and the analyst's eyesight) they can be used for exploratory purposes with linking.

Parallel coordinate displays are primarily for large numbers of continuous variables. Each variable is represented by a vertical axis and individual cases are joined by lines. There is an attractive geometric duality, pointed out by Inselberg (1998), that a point in a scatter plot is represented by a line in a parallel coordinate display and vice versa. Interactive tools are essential for these plots for two reasons: (a) the order of the axes affects what can be seen, so it has to be possible to change the order easily; (b) the large numbers of lines for a dataset of even moderate size makes it essential to be able to temporarily hide points and lines. Categorical variables can also be displayed on their own axes and Wills has suggested using circles whose areas reflect the numbers in a category. Since there is no natural order of categories for most such variables, being able to reorder categories interactively is also valuable. While there is in principle no limit to the number of axes which may be displayed in a scrollable window, it is clear that these displays provide only limited insight to the full multidimensional structure.

4.2 Categorical Variables

Interest in investigating multivariate categorical variables has grown with the rise of data mining. The large datasets which banks, retailers, and other commercial organizations want to study often include many categorical variables. Survey datasets commonly do too, but it is rare to find visualizations of such data. Mosaic plots are one way of uncovering the multivariate structures. Although first discussed in general by Hartigan and Kleiner in 1981, it is only in the late 1990s with the addition of interactive features that they have revealed their true potential (Hofmann 2000). A mosaic plot in its original form is constructed by dividing up the horizontal and vertical axes in turn by each new variable. Figure 9 shows a diagram for the Titanic data set (Dawson 1995). Initially the horizontal axis was divided up according to the numbers in each class, then the individual columns were divided vertically by age, then the resulting cells were each divided horizontally by gender. Finally, the numbers surviving in each group have been highlighted.

This mosaic plot allows the comparison of survival rates by gender within class and age group. The higher survival rates of the female adults compared to the males are obvious. Redrawing with a different order of variables (gender, age, class) permits direct comparisons of survival rates by class within each gender and age group. There are gaps drawn between the cells for two reasons: the display is easier to interpret, and empty cells can be shown (for instance, there were no children in the crew of the Titanic).

Alternative versions of mosaic plots include: (a) the same bin size display, where each cell has the same size;

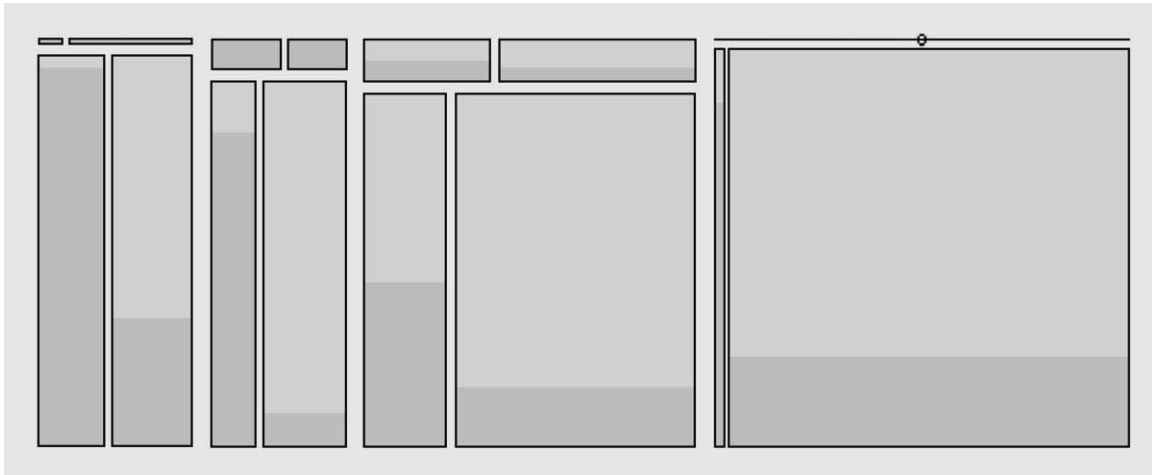


Figure 9

Mosaic plot of the Titanic data in the order class, age, gender with survivors highlighted

this is good for identifying patterns of empty combinations and for comparing highlighted rates for all combinations; (b) fluctuation diagrams, where each cell's size is proportional to the count of the combinations which it represents, but the cell is drawn at a fixed position on the grid. Fluctuation diagrams have the advantage of maintaining a clearer structure, but can be very inefficient in their use of space in comparison to the original format. The alternating of horizontal and vertical axes is an arbitrary restriction and Hofmann has introduced double-decker plots, where only the horizontal axis is divided, as a particularly insightful way of looking at association rules.

One of the most useful approaches for analyzing multivariate data, especially with large datasets, is to study subsets of the data in parallel as in Fig. 7. Trellis displays (Becker et al. 1996) are a more sophisticated version and draw individual displays for each subset in a structured panel display. As in a mosaic plot, the panels are defined by a hierarchical succession of variable splits. Unlike mosaic plots the panels are all of the same size. Each panel may display any one of a number of basic statistical graphics, for example, a box plot or a scatter plot. The art in a successful trellis display lies in the choices of panel plot and of how to split the data set into subgroups. For this reason it would be useful for them to be interactive, but at the time of writing trellis displays are only available in a static form.

4.3 Multivariate Projective Displays

The multivariate displays mentioned thus far show all the raw data. An alternative is to concentrate on projections of the data onto specially chosen lower dimensional spaces. Good projections reveal almost

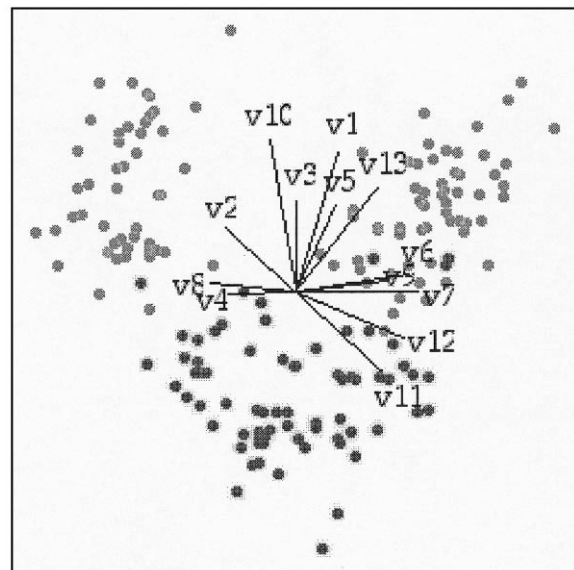


Figure 10

A biplot of the 13 variables and 178 cases of the UCI Wine data set

all the information in the data, albeit in the transformed space. Biplots are so called because they show both the projected points and also the axes of the variables (for continuous variables) or the category level points (for categorical variables) in the projection space (Gower and Hand 1996). Figure 10 shows the UCI wine dataset with the second group highlighted. This projection separates the groups fairly well. Long axes show that the corresponding variables are well represented in the projection space. Biplots have many

attractive properties, but remain two-dimensional representations. The larger the number of variables, the less likely it is that a biplot can reveal the major features of a dataset, but the addition of interactive tools enhances the capabilities of biplots substantially (Hofmann 1998).

The first major innovation in exploratory graphics was none of these tools: it was the rotating plot (Fisher et al. 1974 in Cleveland and McGill 1988), in which a succession of two-dimensional images gives the impression of rotation in three dimensions. This is a natural extension of a scatter plot, is excellent for three-dimensional continuous data, and is fun to use, but for higher-dimensional datasets it is not so practical. In that case it can be used in a different way, displaying three-dimensional projections of the data. Many methods of identifying 'interesting' projections have been discussed and the most effective implementation seems to be a directed grand tour. The grand tour was designed to rotate through all possible three-dimensional projections of a dataset of continuous variables and the analyst would decide which of the resulting projections gave him or her insight into the data. Directed grand tours make use of projection pursuit indices to guide the user towards revealing projections (Swayne et al. 1998).

5. Specialist Displays

5.1 Missing Values

In the histogram in Fig. 2 there was a special column for recording the numbers of cases with no values for that variable. Missing values are a problem for large datasets and any initial overview of the data needs to check for numbers of missings and for any structure among them. Some statistical methods cannot handle cases with missings. A simple approach is the missing value plot as in Fig. 11. There is a bar for each variable. The solid part of the bar on the left represents the cases with values and the blank section on the right the missing values. Bars can be queried and selected to give a rough idea of associations. In the figure it is clear that many cases are missing on a large number of variables.

This kind of display cannot reveal more complex structures and one idea is to use a fluctuation diagram of the corresponding binary variables (missing: yes/no), concentrating on the 10 variables which still have noticeable numbers missing.

5.2 Weighted Displays

In datasets arising from surveys or based on geographic areas, it can often be the case that individual

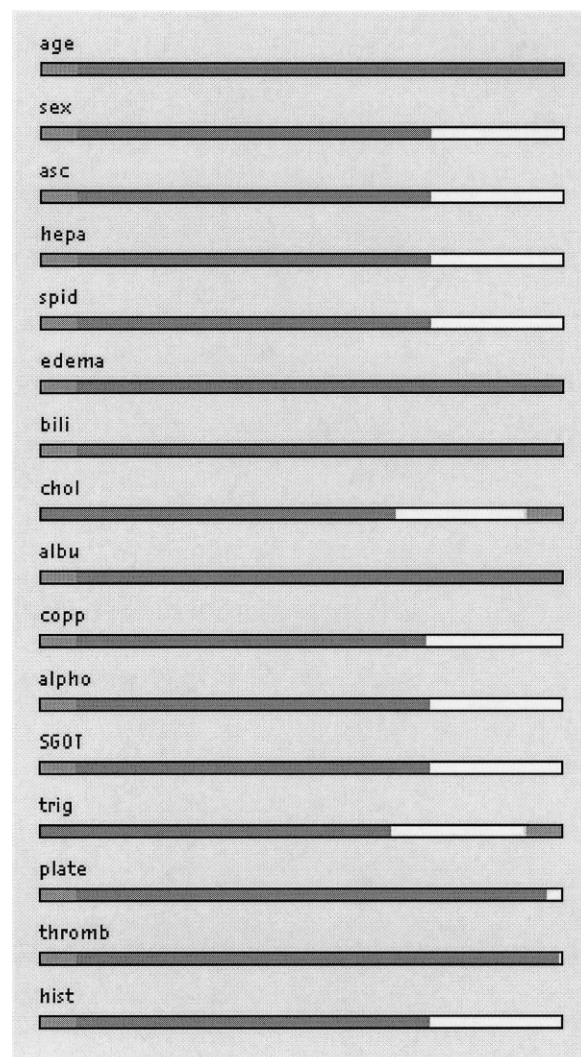


Figure 11

A missing value plot for variables from a patient survival study. A subgroup missing on only two variables is highlighted

data points are of differing importance: the unemployment rate in London should have a greater weight than the unemployment rate in Stratford-upon-Avon. Some authors have suggested using bubbles instead of points in dot plots and scatter plots, where the size of the bubble is proportional to the weighting variable. This can be effective for small numbers of cases, but generally leads to overplotting problems. Area-based displays such as histograms, bar charts, or mosaic plots have no such problems and counts are simply replaced by sums of weights. This can lead to very striking and informative differences in displays. The

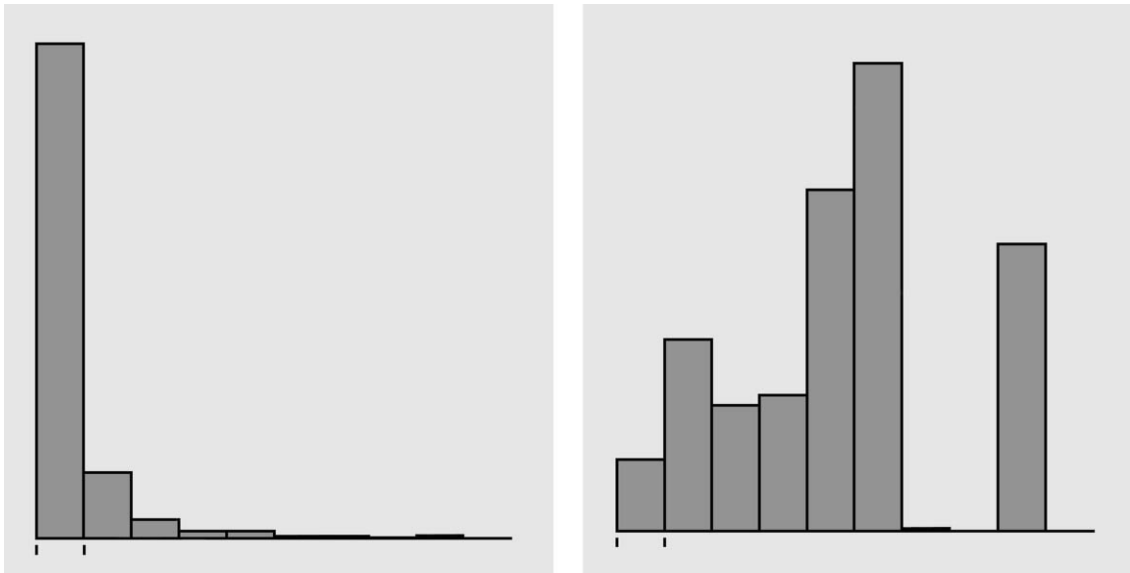


Figure 12

Histograms of black population percentage by US MidWest county, unweighted (left) and weighted by black population (right)

left-hand histogram in Fig. 12 just about shows a single county with the highest percentage of blacks among the over 400 counties in the five US mid-west States. Most counties have low percentages of blacks. The right-hand histogram has the same horizontal scaling, but the height of the bins is now proportional not to the number of counties but to the aggregate black population in the bins. Most blacks live in counties with relatively higher percentages of blacks. It is, of course, necessary that the weighting variable is additive! Population is fine, population density is not (see Unwin and Hofmann 1998).

6. Statistical Tools and Plots

Surprisingly, and disappointingly, statistical graphics rarely make use of statistical tools. There are a few exceptions. Some packages offer the option of displaying box plots with confidence intervals around the medians. Some packages offer the chance of plotting error bars round means. SAS have software (*JMP*) where circles are drawn round subgroups as a means to enable graphical detection of whether differences are significant or not. However, the last example illustrates one of the problems. *JMP* offers several different options because there is no agreed optimal way of comparing several means simultaneously. The same

problem would arise with bar charts and more seriously with mosaic plots; there are so many possible statistical tests that it is not easy to specify a universally acceptable default. The solution will be to offer the various options interactively.

7. The Future

Advances in statistical graphics have not been due to improvements in screen resolution. The best exploratory graphics are simple displays. What makes them powerful is the speed and flexibility with which they can be manipulated and combined. It is the increase in computing power which has been crucial for modern interactive graphics. More powerful computing has also led to the development of computer-intensive analytic procedures and these have sometimes been combined successfully with graphics. Evaluation of projection pursuit indices and rotating plots is one example. Smoothing methods such as Lowess are another.

A major requirement for all statistical methods is that they can be scaled up to deal with huge datasets. Area-based graphical tools such as histograms and bar charts can in principle be used for datasets of any size, but interaction is necessary to catch details and explore links between variables. Much more needs to be done to cope with displays for large numbers

of variables, and interactive capabilities will be essential.

Statistical graphics has grown in fits and starts. There are many attractive ideas, some general principles, but little theory. Real progress will only come with a successful formalization, so that the future can build on the successes of the past. This is something to look forward to in all senses of the phrase.

8. Software

The displays in this article have been taken from screenshots. There has been no attempt to pretty them up; they illustrate exploratory analyses in progress. Most are from MANET (Unwin et al. 1996 and www1.math.uni-augsburg.de/Manet/), a couple are from Data Desk (www.datadesk.com). Other software also offers interactive capabilities, but the degree of interactivity varies considerably. For up-to-date information on software, check statistical websites.

Bibliography

- Becker R, Cleveland W S, Shyu M-J 1996 The visual design and control of trellis display. *Journal of Computational and Graphical Statistics* **5**: 123–55
- Carr D B, Littlefield R J, Nicholson W L, Littlefield J S 1987 Scatterplot matrix techniques for large N . *Journal of the American Statistical Association* **82**(398): 424–36
- Cleveland W S, McGill M E (eds.) 1988 *Dynamic Graphics for Statistics*. Wadsworth & Brooks/Cole, Pacific Grove, CA
- Dawson R J M 1995 The ‘unusual episode’ data revisited. *Journal of Statistics Education (Online)* **3**(3):
- Eick S G, Wills G J 1995 High interaction graphics. *European Journal of Operational Research* 445–59
- Gower J C, Hand D J 1996 *Biplots*. Chapman & Hall, New York
- Hofmann H 1998 Interactive biplots. In: *NTTS 98* 1. Eurostat, Sorrento, Italy, pp. 127–36
- Hofmann H 2000 Exploring categorical data: Interactive mosaic plots. *Metrika* **51**(1): 11–26
- Hyndman R J 1996 Computing and graphing highest density regions. *American Statistician* **50**: 120–26
- Inselberg A 1998 Visual data mining with parallel coordinates. *Computational Statistics* **13**(1): 47–63
- Spence R 2001 *Information Visualization*. Addison-Wesley, New York
- Swayne D F, Cook D, Buja A 1998 XGobi: Interactive dynamic data visualisation in the X window system. *Journal of Computational and Graphical Statistics* **7**(1): 113–30
- Tierney L 1990 *Lisp-Stat*. Wiley, New York
- Tukey J W 1977 *Exploratory Data Analysis*. Addison-Wesley, London
- Unwin A R 1999 Requirements for interactive graphics software for exploratory data analysis. *Computational Statistics* **14**: 7–22
- Unwin A R, Hawkins G, Hofmann H, Siegl B 1996 Interactive graphics for data sets with missing values—MANET. *Journal of Computational and Graphical Statistics* **5**(2): 113–22
- Unwin A R, Hofmann H 1998 New interactive graphics tools for exploratory analysis of spatial data. In: Carver S (ed.) *Innovations in GIS 5*. Taylor & Francis, London, pp. 46–55

Wilkinson L 1999 *The Grammar of Graphics*. Springer, New York

A. Unwin

Graphical Methods: Presentation

Figure 1 shows a graphic of death rates against birth rates for 27 selected countries in a 1990 UN databank (Wilkinson 1999b). The contours in Fig. 1 show two concentrations of countries. One, to the left, has relatively lower death rates and small-to-moderate birth rates. The second, toward the upper right, has high death rates and extraordinarily high birth rates. The curve in the middle of the contours shows that the overall relation between death and birth rates is curvilinear.

The single plot in Fig. 1 contains three graphs: (a) a point graph (collection of points) whose labels show country names; (b) a set of contours based on a kernel density estimate (Silverman 1986) that represents the concentration of the countries; and (c) a LOESS smoother (Cleveland and Devlin 1988). The graphic also includes three guides for our understanding. The first is a general geometric object called a form that is, in this instance, a line delineating zero population growth. Countries to the left of this line tend to lose population, and countries to the right tend to gain. The other two are guides that demarcate axes for the represented space. Other examples of guides are legends and titles.

Each of the constituent graphical elements in Fig. 1 has a long lineage in the history of statistical charts. Scatterplots of points and contours are several centuries old, and regression smoothers are at least a century old. The use of a rectangular coordinate system to locate graphs within a frame demarcated by axes predates by centuries Descartes’ eponymous Cartesian coordinate system. We shall review briefly the history of statistical graphics and charts.

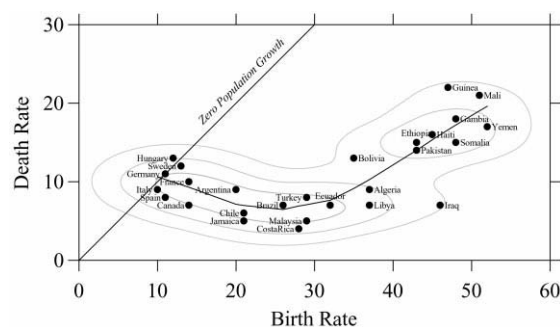


Figure 1
Plot of 1990 death rates against birth rates for selected countries

The lines, colors, typography, proportions, and other aspects of Fig. 1 have been chosen to facilitate communication of the ideas and models underlying the chart. We do not need to resort to platitudes such as, 'A picture is worth a thousand words' to understand that this graphic comprises a relatively substantial analysis of the data. We shall review briefly the topic of graphics principles which can help us devise effective ways to display analyzed data.

The world of charts is enormous. Principles can guide us through that world, but there are many alternatives for representing one or more variables in a graphic. In the final section, we shall present some of these choices and use them to derive additional guidelines for developing effective presentations.

1. History

The contours, or level curves, in Fig. 1 represent a map. This representation is based on a statistical function that maps birth rate and death rate onto a density. Any value on this density can be derived from a corresponding pair of birth rate and death rate values taken from the domain of values within the frame bounded by the two displayed scales. The geometric space in which this mapping is embedded is three-dimensional. The first two dimensions are birth rate and death rate, and the third dimension is density.

The technical mathematical terms in the previous paragraph (contours, level curves, map, function, onto, density, value, domain, frame, bounded, scales, geometric space, and dimensions) are derived from ordinary language, often ancient, describing the physical world. This correspondence between the world of charts and the physical world is not coincidental. Charts are maps of abstract worlds. The word chart and the word cartography have the same root (Latin *charta*, a piece of paper or papyrus). The Greek word for geometry means 'land measurement'. A scale (in Latin) is a ladder. We map in order to organize our world in our mind. However, our world is larger than the domain of geography. As Pinker (1997) has argued, abstract reasoning is built on metaphors for reality. We manipulate abstractions by making them analogous to concrete objects.

Statistical charts are maps, but it does not suffice to simply describe them as abstract geography. The history of statistical charts contains remarkable leaps from a world of continuous rectangular (or spherical) space to other worlds that are categorical (worlds with gaps), multidimensional (worlds beyond three dimensions), and topological (worlds in abstract coordinate systems). Funkhouser (1937), Tilling (1975), Beniger and Robyn (1978), Fienberg (1979), Robinson (1982), Stigler (1983), Tufte (1983, 1990, 1997), Collins (1993), Wainer (1997), and Wainer and Velleman (2001) document these leaps. These references also contain numerous examples of graphics, ancient and modern,

that represent significant landmarks in this history. The remainder of this section will highlight a few.

Physical coordinates were used by Egyptian surveyors as early as the third millennium BCE for locating points on land (Beniger and Robyn 1978). By the beginning of the second millennium CE coordinates were used to locate nongeographic entities. Funkhouser (1936) shows an example of a line graphic of planetary movements on a space-time grid produced by an unknown astronomer in the tenth or eleventh century. Not surprisingly, one of the 'planets' is the Sun.

Cartesian coordinates were devised by Descartes for locating points in abstract space, particularly for representing equations underlying his analytic geometry (Funkhouser 1937). However, relatively few seventeenth or eighteenth century embeddings of observed data points in Cartesian coordinates have been found. (Funkhouser 1937, Tilling 1975). Scientists of the Enlightenment sometimes regarded graphics as dilettantish, and if they used them at all, it was for geometrical diagrams or maps of algebraic functions rather than for plots of raw data (Beniger and Robyn 1978). Social scientists of this period focused on collecting tables of state statistics that could be perused in detail without resorting to graphical approximation (Beniger and Robyn 1978, although see Stigler 1983). (Beniger and Robyn (1978) and Wainer and Velleman (2001) discuss notable exceptions to these observations (graphics by Wren, Halley, Huygens, Plot, Lister, Priestley) but conclude that the important varieties of what we now call the statistical chart did not arise until the work of Playfair in the late eighteenth century.

Playfair graphed the data of the state, the root of the word statistics. He invented the pie chart, the circle chart, the categorical bar chart, the area chart, and he adapted other geometric forms in novel ways to graph data over time and space. Wainer and Spence (1997) provide a brief biography of Playfair's fascinating life and the charts he created.

Other widely used graphical forms were developed in the nineteenth and early twentieth centuries. Beniger and Robyn (1978) and Stigler (1983) summarize a number of these. Density bars were used by Guerry in 1833 to display crime by age groupings. Pearson adopted this form for representing a frequency distribution and named it a histogram. Minard developed a number of schematic maps overlaying statistical data on geography. Florence Nightingale introduced polar area charts. The age-sex pyramid (a dual histogram) was introduced by F. A. Walker, Superintendent of the US Census in 1874. The sample cumulative distribution plot was introduced by Galton in 1875.

If charts are maps of abstract worlds, we might conclude that guiding principles for graphics usage could be derived from the psychology of perception and cognition in the physical world. Indeed, recent

psychological research has tended to support the simple statement that the more graphics adhere to the rules of the physical world, the more likely we are to extract and process the information in them, reliably and accurately. Section 3 covers this area.

2. Principles

Over the centuries of their history, various charts have emphasized communication (e.g., Snow's map of cholera deaths in London, reproduced in Tufte 1983), persuasion (Fletcher's map of the distribution of ignorance in nineteenth century England and Wales, reproduced in Wainer 1997), and graphic design (e.g., Playfair's chart of imports and exports of England to and from North America, reproduced in Tufte 1983). These criteria are, of course, not exclusive. Some charts incorporate all three (e.g., Minard's compound chart of Napoleon's troop losses in the Russian campaign, reproduced in Tufte (1983).

Not surprisingly, modern critics of charts have tended to align themselves along these same dimensions. Communicators (e.g., Cleveland, Kosslyn) have conducted experiments to determine rules that improve the transfer of information from chart-maker to chart-viewer. Persuaders (e.g., Holmes 1991) have evaluated charts on the basis of their psychological impact. Designers (e.g., Herdeg 1981) have applied aesthetic criteria. Others (e.g., Tufte 1983) have adopted the criteria of the communicator, and the methods of the designer. None of these approaches can claim exclusive validity. Since scientists are more interested in rules that foster accurate communication of replicable results, however, this section will focus on communication.

Evaluating graphical communication has been the province of psychology, in the subdisciplines of human factors, ergonomics, and applied cognitive science. For almost a century, human factor psychologists have developed methods and criteria for evaluating graphics. Military psychologists, for example, have conducted randomized experiments to test the effectiveness of aircraft cockpit displays.

Psychologists gave much less attention to the perception of statistical graphics until a statistician, William Cleveland, published a series of experiments designed to identify aspects of charts that helped or hindered accurate decoding of quantitative data (Cleveland and McGill 1984). (Beniger and Robyn 1978, cite a number of statisticians' informal studies of the effectiveness of popular charts in the 1920s, but these did not induce many psychologists to study the topic more formally.)

Cleveland used paper-and-pencil tests containing simple graphical elements—points, lines, angles, areas, colors—generated randomly by a computer. His subjects gave magnitude estimates of the values generating the instances of each element. From these estimates, he

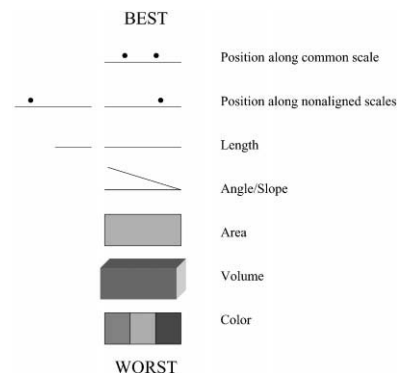


Figure 2

Cleveland (1994) graphical features hierarchy

derived and analyzed error rates. Figure 2 summarizes the main result of these experiments. The top of the figure (BEST) represents elements that resulted in fewer errors and the bottom (WORST) represents those with more.

Cleveland's hierarchy gives us a start in designing effective displays. All other things being equal, we might choose to use a bar chart instead of a pie chart because judgments of position on a common scale and judgments of length are both more accurate than judgments of angle. There are exceptions to these rules, however (see the following discussion of Fig. 3b). A chart is more than the sum of its parts; elements interact in the context of a chart to suppress or enhance errors. Kosslyn (1994) summarizes relevant psychological studies in the extensive notes at the end of his book.

Others have provided numerous general rules for effective graphical communication. The following guidelines are derived from Bertin (1981), Tufte (1983), Cleveland (1994), and Kosslyn (1994).

Avoid clutter. Especially, do not clutter the data region inside the plotting frame demarcated by axes. Tufte constructs a data/ink ratio and urges us to maximize it by removing irrelevant detail. This exhortation has value as long as we do not remove redundant features that can reinforce an accurate perception. Most writers agree that one should avoid jazzy textures (especially stripes), gratuitous colors, excessive tick marks and grid lines, ornate fonts, and unnecessary embellishment. Avoid visual illusions due to pseudo-3D and other gratuitous use of color, angles, area, or volume in a 2D plotting world.

Seek simple geometric forms—straight lines, circles, triangles, squares. When it helps to simplify a relationship, use transformations (logs, square roots) to convert curves to straight lines, for example. Use polar coordinates when variables are cyclical (astronomical perceptual, directional).

Sort and organize. This is especially important for complex graphics. Usually, one should sort the values

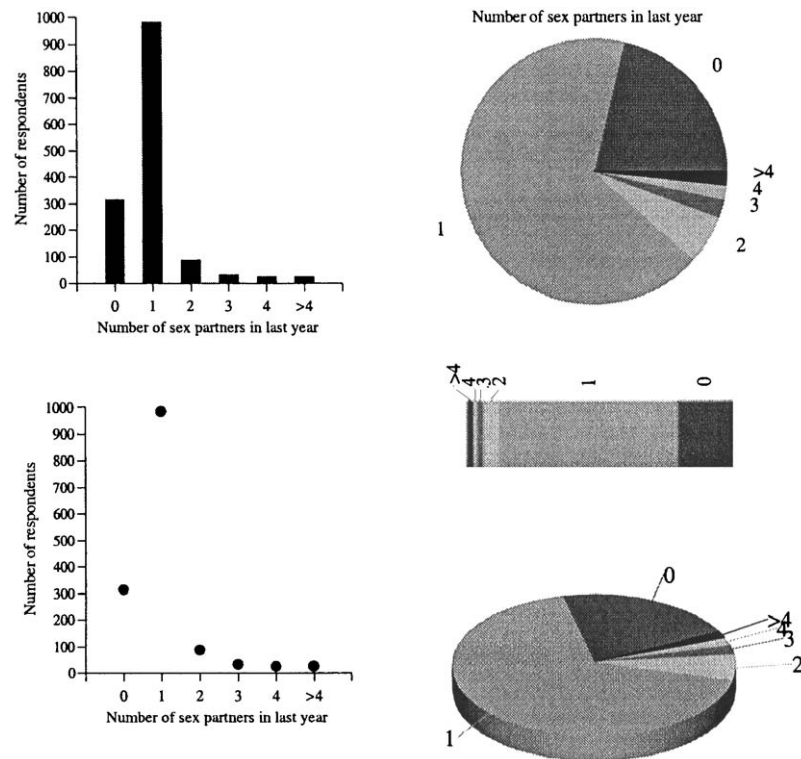


Figure 3
Reported number of sexual partners (General Social Survey, 1993)

of a dependent variable (the range of the graph), but other sorts can help viewers navigate and perceive structure. Bertin (1981) has numerous examples showing the value of simple and multivariate data sorts. Figure 12 shows an example.

Annotate extensively. Make legends comprehensive and informative. Figure captions should describe the source of the data and explain the relation of the data to the graphic. Note exceptions (see Fig. 6).

3. Choices

Rules guide, but examples instruct. We need to consider specific choices for specific problems to understand exceptions to rules. The graphic representation problem is often expressed as one of choosing the proper chart type for a given set of data. This is an unhelpful way to express the problem, however. As Wilkinson (1999c) shows, different chart types with their own popular names are often mirror images or simple transformations of each other. Instead, we will examine several basic problems confronting the designer of a scientific graphic: geometry (choosing the type of graphical element for a chart), coordinates (choosing a coordinate system for the chart), and uncertainty (representing random error).

3.1 Geometry

The following examples were selected to illustrate principles, not exhaust possibilities. It is organized by the configuration and number of variables, categorical and continuous, that we wish to graph.

3.1.1 One variable. Figure 3 shows five different ways of representing values on one categorical variable. The data are from 1606 respondents to the 1993 General Social Survey (Davis et al. 1993). Respondents were asked, 'How many sex partners have you had in the last 12 months?' Those reporting more than four partners (some reported up to 100) were consolidated into the last category. There were 1,466 responses in the resulting six categories.

The upper left graphic is a bar chart of the frequencies of occurrence of each category. Bars are most suited for displaying either magnitudes referenced against zero (anchored bars, as in this figure) or a range of values on a continuous scale (range bars, as in Fig. 7).

Anchoring bars at zero can sacrifice resolution when no data values exist near zero; this creates substantial white space in the plotting area (Cleveland 1994). In

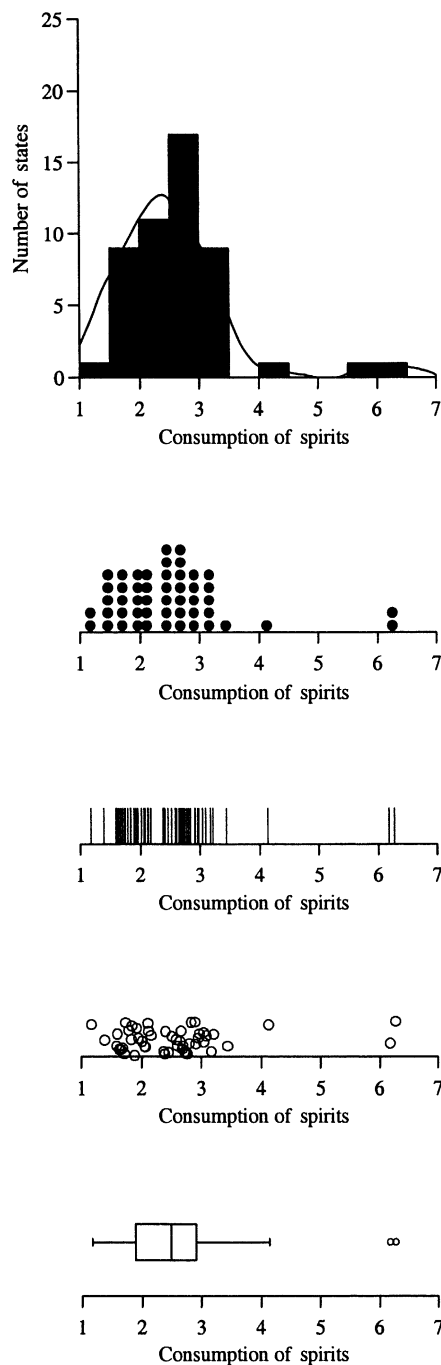


Figure 4
Consumption of spirits in gallons per capita, by US state, 1986

these cases, dots are preferable to bars because we do not need to include zero on the scale. The upper right graphic shows a dot plot of the same variable. In this

case, we would prefer the bar version because there are several values near zero and it is easier to anchor the bars visually on the partner categories. The locations of both the dots and the tops of the bars are easily decoded on the vertical scale.

The third graphic in Fig. 3 is a pie chart of the proportions of total category frequencies. The general antipathy statisticians have toward pie charts is misinformed. Pie charts are not always bad, although the issues on when to use them are complex. In general, bar charts are more suited to absolute judgments and pie charts are more suited to judging proportions (Simkin and Hastie 1987). Pie sectors near a fourth or half of a pie in size are judged especially accurately. When there are many sectors, however, pies are not usually appropriate. In this example, the bar chart would be more suitable.

The fourth graphic is a divided bar chart, which is nothing more than a pie chart in rectangular coordinates. Divided bars do worse than pies in studies of graphical perception (e.g., Simkin and Hastie 1987). Cleveland (1994) provides several examples to show why they are generally a bad choice.

The fifth graphic is a 3D pie chart. This is probably the worst of the popular ways to represent proportions. Three-dimensional graphics are not in themselves harmful. What creates the problem with the 3D pie chart is the distortion of angles that the perspective projection produces. In the 3D pie, the single partner category seems to cover almost 75 percent; the actual coverage is 68 percent.

Some users and computer programs apply line or area elements (see Fig. 7) to categorical variables. Popular examples are line graphics of means in analysis of variance, or area graphics of profiles across categories. These elements should generally be avoided with categorical data. Spacings between categories (even ordered categories) are not fixed, so slopes of line segments or profiles between categories are meaningless.

Figure 4 shows five ways of representing one continuous variable. The data are per-capita consumption of spirits for each of the 50 US states (Bureau of the Census 1986). The top graphic is a classic histogram overlaid with a kernel density estimate (Silverman 1986). Histogram bars look like ordinary bars, but they measure areas rather than intervals and they rest on a continuous scale rather than on a set of categories. Histograms are more useful than kernels when we are interested in displaying frequencies within intervals, especially when the intervals themselves are bounded by meaningful, round units. Kernels are more useful than histograms when we are looking for a smooth estimate of a continuous distribution.

The next graphic down is a dot histogram (Chambers et al. 1983, Wilkinson 1999a). This display represents each observation with a dot located at its scale value. If several values coincide on the scale, the

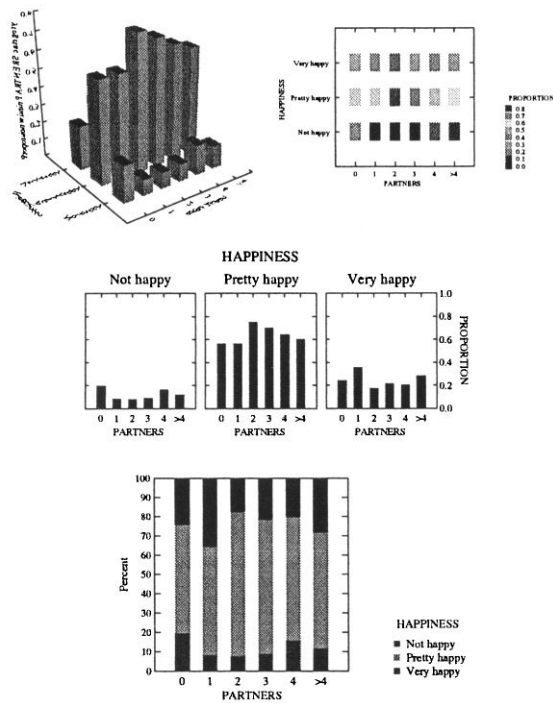


Figure 5

Number of reported sexual partners during previous year by reported level of happiness among 1,462 respondents to 1993 General Social Survey. Third variable is proportion of respondents in each sexual partner category

dots are stacked vertically. This display is most suited for small samples, when every value is to be displayed.

The graphic below that is a stripe plot (Chambers et al. 1983). A vertical stripe occurs at each data value. This display can handle more cases than the dot plot, although it is not suitable for large datasets.

The graphic following is a jittered density (Chambers et al. 1983). Random error is added to each value so that the plotting symbols do not overlap. This display is designed to handle roughly the same conditions as the stripe plot.

The last graphic in Fig. 4 is a box plot, or schematic plot (Tukey 1977). As Tufte (1983) notes, Tukey's plot is based on an earlier display that represents the quartiles of a distribution. In that earlier plot, the whiskers cover the range and the box covers the midrange. Tukey improved on this design by scaling the whiskers to allow for outliers. The display in Fig. 4 shows how important this feature can be. Two states (Nevada and New Hampshire) have unusually large values (see Fig. 6), possibly due to gambling and cheap liquor, respectively. The box plot is a clean representation of the summary values we usually want to see—the median, the twenty-fifth and seventy-fifth

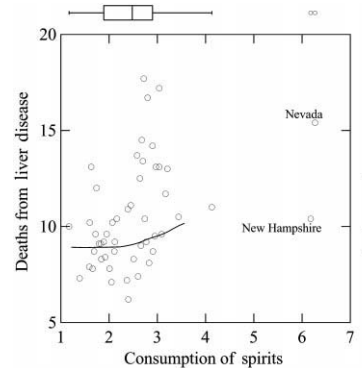


Figure 6

Consumption of spirits in gallons per capita by US state vs. number of deaths from chronic liver disease and cirrhosis per 100,000 by US state, 1986

centiles (hinges), the extreme values (usually bounding the range), and possible outliers (outside values).

3.1.2 Two variables. Figure 5 shows four ways of representing two categorical variables crossed. The data are from the 1993 General Social Survey used in Fig. 3. The additional variable is general happiness, as measured by the response to the question, "Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?"

The upper left plot shows a 3D bar chart of happiness against number of sexual partners. The vertical scale is the proportion of respondents within each partner category. This graphic has the disadvantage that short bars tend to be hidden by long ones. Generally it should be avoided.

The graphic to the right is a tiled bar chart. This graphic uses color or gray scale to indicate the relative frequency of each category combination. However, as Cleveland's research has shown, the use of color or gray scale to encode continuous values is problematic (Cleveland 1994).

The next graphic below shows a paneled categorical plot. Instead of using an extra dimension in 3D, we use an extra categorical dimension in 2D to panel the plot. This is probably the most effective method for handling the extra dimensionality required for representing the frequencies, plus the two categories. The bottom panel shows a multiple divided bar chart. This layout has the same problems as the single divided bar chart (see Fig. 3). Cleveland (1985) shows an example of this graphic and offers a more effective dot plot alternative.

Figure 6 shows the most popular way of representing two continuous variables crossed with each other—the scatterplot. The data are per capita consumption of spirits for each of the 50 US states, used in

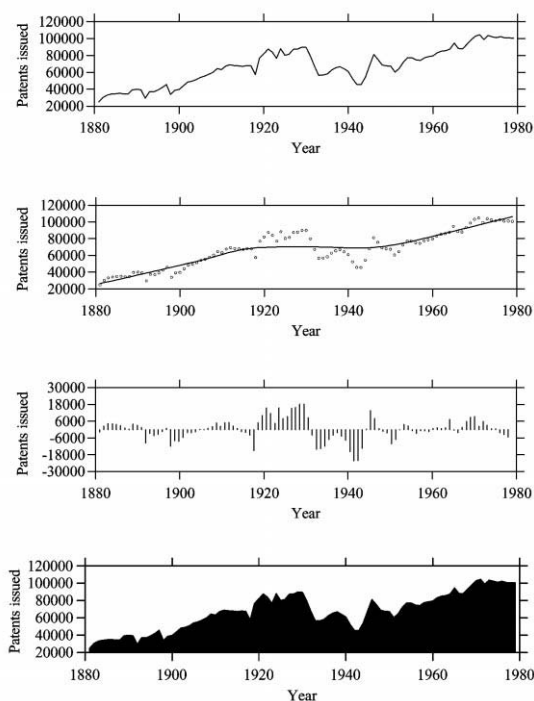


Figure 7

US patents issued from 1880 to 1980

Fig. 4 (Bureau of the Census 1986). The additional variable is number of deaths from chronic liver disease and cirrhosis per 100,000 people, by state. This scatterplot has been enhanced in a number of ways. The bordering box plots help us to assess the marginal distributions of the variables and highlight the Nevada and New Hampshire outliers. The smoother shows the conditional mode of deaths (estimated mode of deaths given consumption). This was computed through kernel regression (Scott 1992). This particular smoother limits itself to areas where there are relatively more data values, so it provides a conservative estimate of trend.

Time series data require special treatment. Figure 7 shows four examples. The data are number of US patents issued in the century from 1880 to 1980 (Wilkinson et al. 1996). The top panel shows a line element representing the series. Lines are most useful when a series is relatively smooth. The next panel shows a point element with a superimposed LOESS smoother. Points are best when we supplement them with a smoother; they tend to obscure time order if used alone, unless the series is especially smooth. The third panel shows spikes. These are vertical lines best used to reveal deviation from a constant level. This panel displays the residuals from the LOESS smooth. The lowest panel shows an area chart for the raw series. This highlights trend but prevents the use of

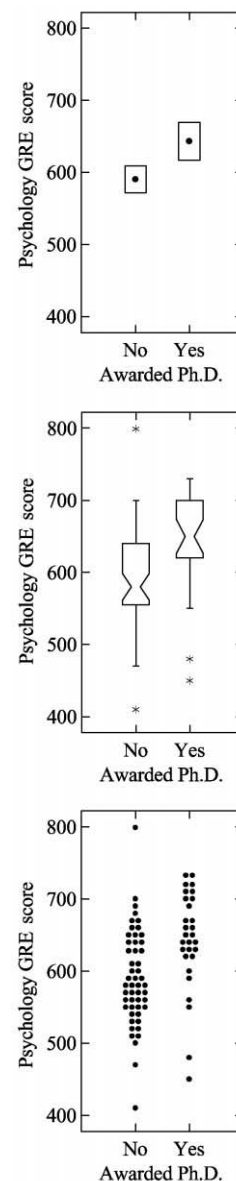


Figure 8

Advanced Psychology Graduate Record Examination (GRE) scores for 80 students taken from 10 years of a graduate psychology program. Students are grouped by whether or not they were awarded a Ph.D.

confidence intervals and other enhancements to the plot. Generally, it should be avoided. Cleveland (1994) discusses other graphical representations for time series data.

Figure 8 shows three ways of representing a categorical variable crossed with a continuous variable.

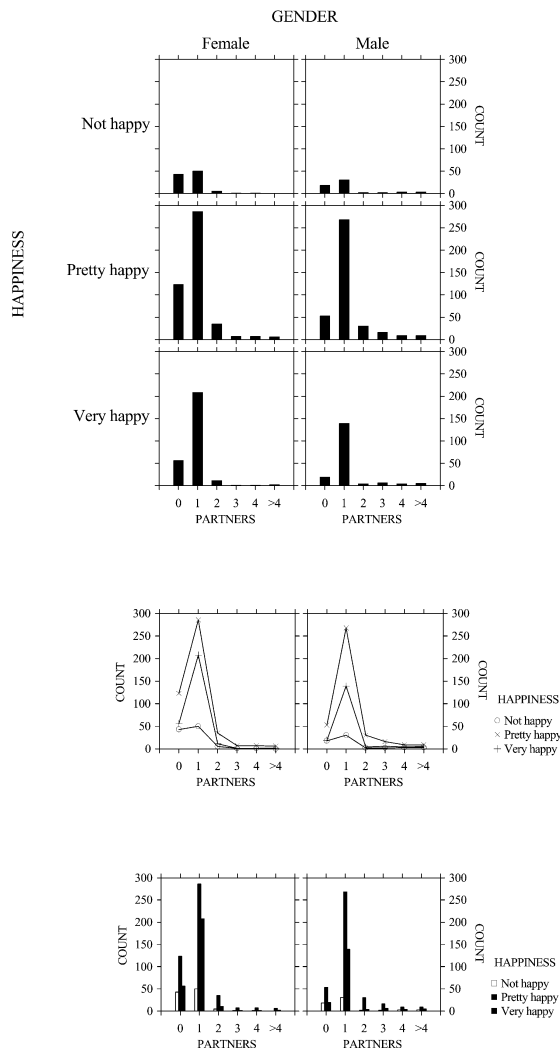


Figure 9

The number of reported sexual partners during last year by reported level of happiness among 1,462 respondents to 1993 General Social Survey. The third variable is the proportion of respondents in each sexual partner category

The data are based on 80 graduate students over a 10-year period in a US psychology department (Wilkinson et al. 1996). Graduate Record Examination Advanced Psychology Test scores are plotted against whether or not the students eventually received their PhD.

The first graphic employs a dot for the mean and a range bar to represent a 95 percent confidence interval on the mean. The bars do not overlap, which suggests that the advanced test can help identify those who achieve a PhD. (Interestingly, neither the Verbal nor

the Quantitative test scores predicted graduation in this sample.)

The second graphic is a notched box plot (McGill et al. 1978). This is a variation on the box plot that not only conveys more of the important data landmarks than the classic confidence interval plot but also provides an approximate confidence interval of its own. If the data are independent samples from identically distributed populations that are lumpy in the middle (approximately normal in their midrange), then comparing the notches yields an approximate 95 percent test of the null hypothesis that the true medians in the population are equal. In this example, the notches do not overlap, reinforcing what we concluded from the confidence intervals on the means. Because it relies on the median instead of the mean, the notched box plot procedure is more robust against outliers.

The third graphic is a dot plot (Wilkinson 1999a). Used frequently in the medical literature, this plot shows every data point. In a small sample (e.g., clinical case study), a dot plot can be useful for readers who wish to consider every data value. However, it is of little use in making informal graphical inferences concerning group differences. It is best to think of grouped dot plots as unadorned one-dimensional scatterplots within categories.

3.1.3 Three variables. Figure 9 shows a triple crossing of categorical variables. The data are from the 1993 General Social Survey used in Figs. 3 and 5. The additional variable is gender (observed by the interviewer, not reported by the respondent). We ask whether the relationship between reported happiness and number of sexual partners differs by gender.

The top plot in the figure shows a paneled graphic. The format is similar to a Trellis display (Becker et al. 1996), but the labeling of the paneling variables (happiness and gender) is placed outside the plotting area to improve readability.

The middle plot collapses happiness into a legend in order to reduce the number of panels. This is a popular method for saving space, particularly when representing factorial layouts in ANOVA and other designs. It has several problems. First, the line segments used to highlight trends have slopes that depend on the particular spacing and arrangement of partner categories. Unless we employ an external scaling procedure to determine the spacing of the partners categories, we have scant justification for using these lines. Second, the collapsing introduces a symbol choice problem. It is difficult to find symbols that are easily distinguishable for more than a few categories. The symbols collide, as well, at the upper end of the horizontal scales.

The lower plot introduces an even less appealing alternative. Clustered bar charts are used widely but they have several defects. First, it is difficult to discern

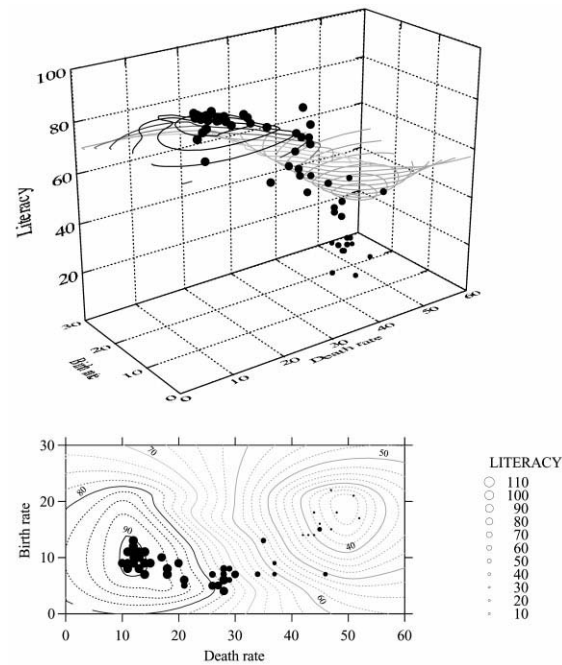


Figure 10
Birth rate and death rate vs. literacy for selected countries

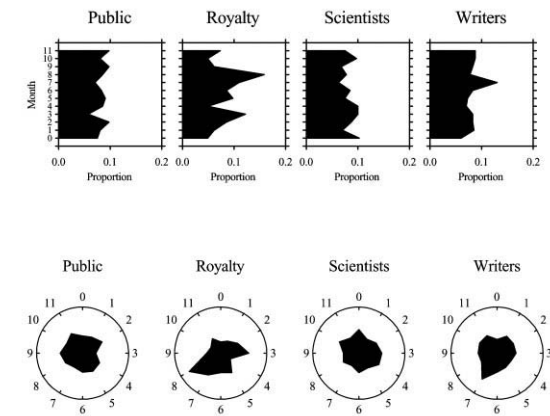


Figure 11
Difference between death month and birthday by occupational category

separate patterns for the categories. One needs to focus on one set of bars to do this, but there is visual interference from the other bars in each cluster. Second, these bars can become quite thin with more than a few categories. Decoding this graphic is problematic.

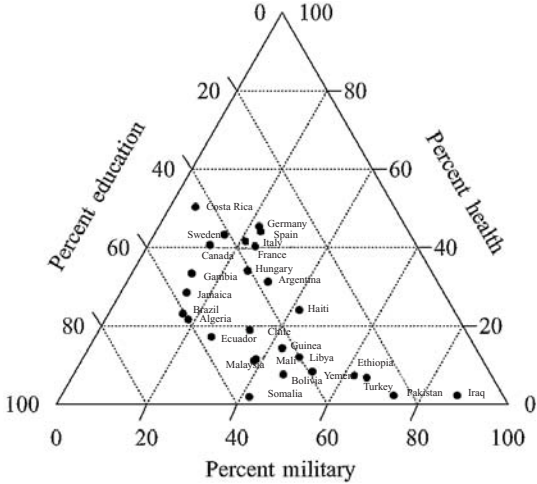


Figure 12
Percentage per-capita spending on education, military, and health

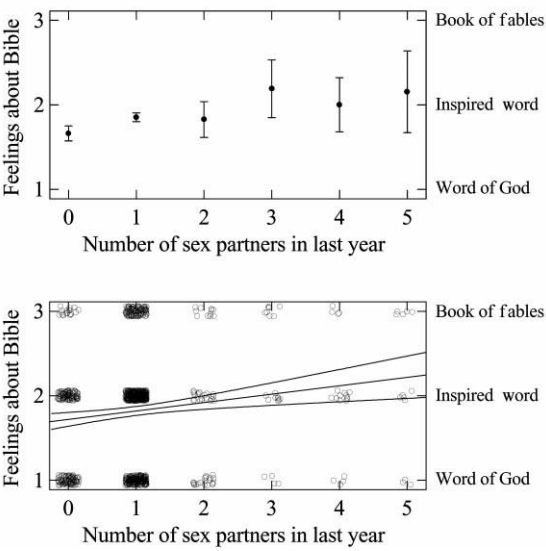


Figure 13
Sex partners vs. feelings about the Bible (GSS survey responses)

Figure 10 shows a triple crossing of continuous variables. The data are from the UN databank used in Fig. 1. The additional variable is literacy (percentage of the population that can read). We ask whether literacy is related to birth and death rates taken together.

The upper plot is a 3D frame that includes a contoured regression surface and a scatterplot. The

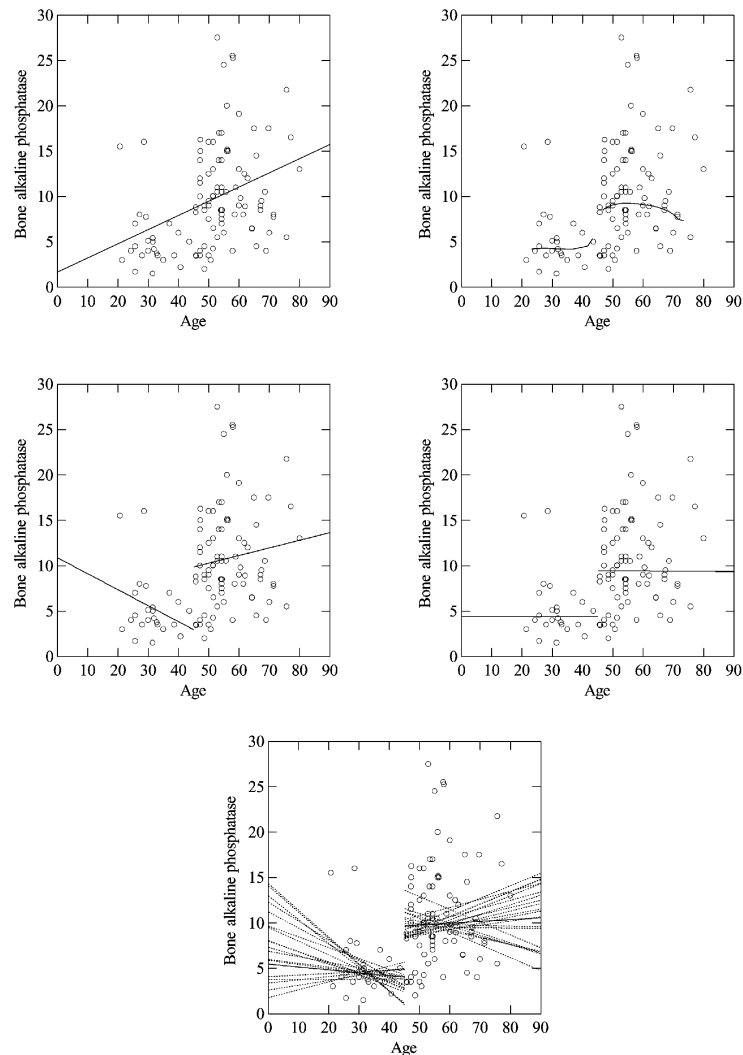


Figure 14
Bootstrapped data from Gonnelli et al. (1996)

contours (as opposed to a wire-frame or rendered surface) help viewers discern literacy levels in different elevations of the plot and keep the surface from hiding the data. Coloring the contours improves the coherence of the surface.

The lower plot is a 2D frame that includes the same contoured surface plus a scatterplot of symbols whose size is proportional to literacy. This plot makes it easier to decode specific xyz triplets.

Some critics tend to eschew 3D surface plots, but they have their uses. As Wilkinson (1999c) argues, surfaces elicit a holistic impression of a function. They are less useful for decoding individual values, however. Perhaps the best compromise is to present both displays simultaneously to provide different structural

views. Another alternative is to categorize one or more of the variables and display the results in a paneled plot such as those in Fig. 9.

3.2 Coordinates

We do not ordinarily think of charts being embedded in different coordinate systems, except for certain scientific graphics such as polar plots. Some chart types, such as radar charts, pie charts, and horizontal bar charts are nothing more than popular charts in unusual coordinate systems. Wilkinson (1999c) provides numerous examples.

Figure 11 shows two different graphics in different coordinate systems, based on data from Andrews and

Herzberg (1985), contributed by C. O'Brien. The data are months since their last birthday during which members of different social groups died. The upper plot shows a paneled horizontal area chart produced by rotating a vertical area chart 90 degrees (exchanging x and y coordinates in this case). This graphic reminds us that horizontal bar charts are a coordinate transformation of a vertical bar chart, not a new chart type.

If it makes sense to consider a death one month before a birthday as being more similar to a death one month after, than to one six months after, then we might want to plot these variables in polar coordinates. The lower plot shows a profile plot of the same results.

Figure 12 contains a triangular coordinate plot. Mosteller and Tukey (1968), Bishop et al. (1975), and Wainer (1997) show examples of this type of plot. The data are from the UN dataset used in Fig. 1. The variables represent expenditures in US dollars per person for health, education, and the military.

The outside tick marks signify how to read the scales. Grid lines collinear with ticks demarcate values on the three scales. Iraq and Pakistan, for example, allocated most to military spending and least to education and health. Costa Rica spent more than 40 percent on education, almost 50 percent on health, and less than 10 percent on the military. While triangular coordinates may be used most frequently in analysis of mixture models in industrial experiments, they can also be useful for social science models involving tradeoffs or compensatory mechanisms.

3.3 Uncertainty

There are two chief methods for representing error in a graphic. The first is sharp: error bounds are represented by clear edges, points, or lines. The second is fuzzy: error is represented by blurring an estimate.

Figure 13 shows two sharp ways of representing error in a graphic. The data are from the 1993 General Social Survey used in Fig. 3. The additional variable is feelings about the Bible, as measured by the response to the question, "Which of these statements comes closest to describing your feelings about the Bible?" The responses coded are 1 (Word of God), 2 (Inspired word), and 3 (Book of fables). We assume that the dependent variable is continuous (biblical absolutism vs. relativism?) even though the responses are integers.

The upper plot shows error bars representing 95 percent confidence intervals on the means by category. Error bars are also used to represent one standard deviation or one standard error. It is important to make clear in accompanying titles or notes, which type is used. If the spacing of values on the independent variable is meaningful, then we can fit a linear regression model to our data. We shall further assume the number of sex partners to represent values of a continuous variable (promiscuity?). Assuming the

assumptions for inference using a t distribution are appropriate, we can fit continuous confidence bounds on the regression line as in the lower panel of the figure. This plot uses jittering to reveal the concentration of data points where they are tied.

Figure 14 shows one way to represent error by fuzziness. The data are from Gonnelli et al. (1996). They represent concentration of bone alkaline phosphatase (BAP) in a sample of women of different ages. The authors fit a linear regression (shown in the upper left panel) to argue that BAP levels increase with age.

A modal regression in the upper right panel indicates that there is a discontinuity in this relationship at age 45 years or so, corresponding most likely to the onset of menopause. Accordingly, we fit separate linear regressions to the two subgroups split at age 45 years (third panel). These regressions appear to be sensitive to outliers, however, so we fit linear models using robust regression with t weighting (fourth panel). This fit indicates that a plausible model for predicting BAP from age involves level differences but no slope differences.

It is not easy to compute confidence intervals on the robust linear fits, so we resort to bootstrapping to provide an estimate of error. The bottom panel shows the result of 20 bootstrap robust fits displayed as faint dashed lines. The non-overlapping envelopes of the fits indicates that our level-change model is reasonable.

4. Conclusion

Figure 15 shows the principal result of a survey by Cleveland (1984) on the use of graphics in scientific articles. The horizontal axis represents the percentage of total page area devoted to graphs in 47 articles sampled from the 1980–81 volumes of selected journals. As Cleveland noted, the substantial differences among disciplines remain when one examines number of graphs per article instead of page area.

In Cleveland's survey, social science journals ranked lowest in the use of graphics. We might speculate that this is a legacy of the aversion toward graphics among

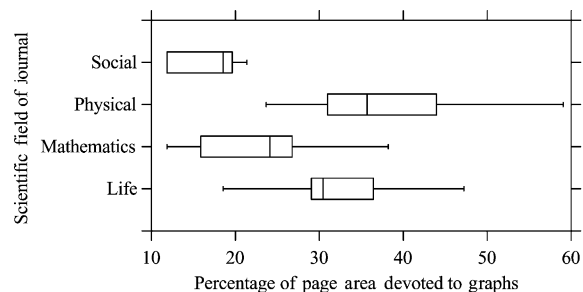


Figure 15

Use of graphics in journal articles (data from Cleveland 1984)

eighteenth century social scientists that Beniger and Robyn (1978) describe. Whatever the reason, there is little justification for this to continue. Software packages such as S-Plus, SAS, SPSS, and SYSTAT (used for the graphics in this article) now offer statistical graphics and publishing standards such as Adobe PDF format now make it possible to display graphics as easily as text.

See also: Bayesian Graphical Models and Networks; Graphical Models: Overview

Bibliography

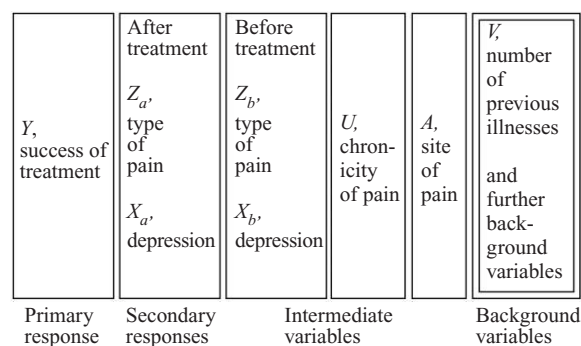
- Andrews D F, Herzberg A M 1985 *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. Springer-Verlag, New York
- Becker R A, Cleveland W S, Shyu M-J 1996 The design and control of Trellis display. *Journal of Computational and Graphical Statistics* **5**: 123–55
- Beniger J R, Robyn D L 1978 Quantitative graphics in statistics: A brief history. *The American Statistician* **32**: 1–11
- Bertin J 1967 *Semiologie Graphique*. Editions Gauthier-Villars, Paris [English translation Berg W J, Wainer H 1983 *Semiology of Graphics*. University of Wisconsin Press, Madison, WI]
- Bertin J 1981 *Graphics and Graphic Information-Processing*. [English translation Berg W J, Scott P. Walter de Gruyter, New York]
- Bishop Y M, Fienberg S E, Holland P W 1975 *Categorical Data Analysis*. MIT Press, Cambridge, MA
- Bureau of the Census 1986 *State and Metropolitan Area Data Book*. US Government Printing Office, Washington, DC
- Chambers J M, Cleveland W S, Kleiner B, Tukey P A 1983 *Graphical Methods for Data Analysis*. Wadsworth, Monterey, CA
- Cleveland W S 1984 Graphs in scientific publications. *The American Statistician* **38**: 19–26
- Cleveland W S 1993 A model for studying display methods of statistical graphics (with discussion). *Journal of Computational and Graphical Statistics* **2**: 323–43
- Cleveland W S 1994 *The Elements of Graphing Data*, rev. edn. Hobart Press, Summit, NJ
- Cleveland W S 1995 *Visualizing Data*. Hobart Press, Summit, NJ
- Cleveland W S, Devlin S 1988 Locally weighted regression analysis by local fitting. *Journal of the American Statistical Association* **83**: 596–640
- Cleveland W S, McGill R 1984 Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association* **79**: 531–54
- Collins B M 1993 Data visualization—has it all been seen before? In: Earnshaw R A, Watson D (eds.) *Animation and Scientific Visualization: Tools and Applications*. Academic Press, New York, pp. 3–28
- Davis J A, Smith T W, Marsden P V 1993 *The General Social Survey*. National Opinion Research Center, Chicago
- Fienberg S 1979 Graphical methods in statistics. *The American Statistician* **33**: 165–78
- Funkhouser H G 1936 Note on a tenth century graph. *Osiris* **1**: 260–2
- Funkhouser H G 1937 Historical development of the graphical representation of statistical data. *Osiris* **3**: 269–404
- Gonnelli S, Cepollaro C, Montagnani A, Monaci G, Campagna M S, Franci M B, Gennari C 1996 Bone alkaline phosphatase measured with a new immunoradiometric assay in patients with metabolic bone diseases. *European Journal of Clinical Investigation* **26**: 391–6
- Herdeg W 1981 *Graphis Diagrams: The Graphic Visualization of Abstract Data*. Graphis Press, Zürich, Switzerland
- Holmes N 1991 *Designer's Guide to Creating Charts and Diagrams*. Watson-Guptill Publications, New York
- Kosslyn S M 1994 *Elements of Graph Design*. W. H. Freeman, New York
- McGill R, Tukey J W, Larsen W A 1978 Variations of box plots. *The American Statistician* **32**: 12–6
- Mosteller F, Tukey J W 1968 Data analysis, including statistics. In: Lindzey G, Aronson E (eds.) *Handbook of Social Psychology*, 2nd edn. Addison-Wesley, Reading, MA, pp. 80–203
- Pinker S 1997 *How the Mind Works*. Norton, New York
- Robinson A H 1982 *Early Thematic Mapping in the History of Cartography*. University of Chicago Press, Chicago
- Scott D W 1992 *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, New York
- Silverman B W 1986 *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London
- Simkin D, Hastie R 1987 An information processing analysis of graph perception. *Journal of the American Statistical Association* **82**: 454–65
- Stigler S 1983 *The History of Statistics*. Harvard University Press, Cambridge, MA
- Tilling L 1975 Early experimental graphics. *The British Journal for the History of Science* **8**: 193–213
- Tufte E R 1983 *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT
- Tufte E R 1990 *Envisioning Data*. Graphics Press, Cheshire, CT
- Tufte E R 1997 *Visual Explanations*. Graphics Press, Cheshire, CT
- Tukey J W 1977 *Exploratory Data Analysis*. Addison-Wesley, Reading, MA
- Wainer H 1997 *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot*. Springer-Verlag, New York
- Wainer H, Spence I 1997 Who was Playfair? *Chance* **10**: 35–7
- Wainer H, Velleman P F 2001 Statistical graphs: Mapping the pathways of science. *The Annual Review of Psychology* **52**
- Wilkinson L 1999a Dot plots. *The American Statistician* **53**: 276–81
- Wilkinson L 1999b *SYSTAT*. SPSS, Chicago
- Wilkinson L 1999c *The Grammar of Graphics*. Springer-Verlag, New York
- Wilkinson L, Blank G, Gruber C 1996 *Desktop Data Analysis with SYSTAT*. Prentice-Hall, Upper Saddle River, NJ

L. Wilkinson

Graphical Models: Overview

1. Some General and Historical Remarks

Graphical models aim to describe concisely the possibly complex interrelationships between a set of variables. Moreover, from the description key, proper-

**Figure 1**

Initial partial ordering of variables in chronic pain study for a multivariate regression chain. Variables in the same box are treated on an equal footing. Variables in any one box are considered conditionally given all variables in boxes listed to their right. Associations among background variables in the double-lined box are not analyzed but taken as given

ties can be read directly. The central idea is that each variable is represented by a node in a graph. Any pair of nodes may be joined by an edge. For most types of graph a missing edge represents some form of independency between the pair of variables. Because the independency may be either marginal or conditional on some or all of the other variables, a variety of types of graph are needed.

A particularly important distinction is between directed and undirected edges. In the former an arrow indicates the direction of dependency from an explanatory variable to a response. If, however, the two variables are to be interpreted on an equal footing then an edge between them is undirected, or cyclic dependencies are permitted. For instance, systolic and diastolic blood pressure would typically be treated on an equal footing because they are two aspects of a single phenomenon, namely the blood pressure wave.

Graphical models started to be developed by Darroch et al. (1980) and Wermuth (1976) as special subclasses of loglinear models for contingency tables and of multivariate Gaussian distributions which are interpretable in terms of conditional independencies and can be represented by undirected graphs. See *Multivariate Analysis: Overview; Multivariate Analysis: Discrete Variables (Loglinear Models)*.

The first extension was to problems in which the variables can be arranged in sequence so that each variable is regarded as a response to all variables to the right of it in the sequence. This leads to representation by a directed acyclic graph. Then there are joint response models with graphs having blocks of undirected components, the blocks being again connected in a directed and acyclic way. Much recent and ongoing research is for more specialized situations such as

models for confounded dependencies, for event history data, for time series data, and for simultaneous dependencies in multivariate Gaussian distributions. The latter correspond, for instance, to independence graphs having directed cyclic components.

As an example of some of the above ideas we report here on a study concerning chronic pain. One main research question is: when investigating the development of this illness, is it necessary to include information on the patient's site of main pain? Figure 1 shows a first partial ordering of a set of variables within boxes. It reflects some of the prior knowledge on the development of chronic pain and its treatment. It also determines important aspects of the types of analysis to be carried out: relations among several background variables are taken jointly as given, univariate conditional distributions are studied when there is a single variable within a box, and joint conditional distributions are studied when there are several variables within a box. We return to a graph resulting from such statistical analyses in Sect. 6

We concentrate in this article on the use of models for the analysis and interpretation of empirical data, and mention some of their other uses in Sect. 7 on suggested further reading. The more recent statistical work quoted here by authors only is referenced in Edwards (2000), Studený and Bouckaert (1998), Richardson (1998), or Wermuth (1998).

2. Models for Contingency Tables and the Need for Extensions

For a q -dimensional contingency table, i.e., for counts on q categorical variables, the most complex undirected graph model has no constraints. It is called the saturated model and is represented by a graph in q vertices or nodes, with each node pair having exactly one edge, i.e., being connected by a full line. Each full line represents the conditional association of the connected variable pair given all remaining $q-2$ variables. Removing a line from the graph introduces a particular independence constraint for the corresponding variable pair: the pair are to be conditionally independent given all remaining variables. In a log-linear model formulation this is achieved by setting the two-factor and all higher-order interaction terms involving this variable pair to zero. If all full lines are removed, the graph of q disconnected nodes results. It represents the simplest model in the class, the one of mutual independence of all q variables.

This development built on previous work (quoted in monographs on graphical models) by M. S. Bartlett, M. W. Birch, L. A. Goodman, Y. Y. Bishop, S. E. Fienberg and P. Holland and much earlier work by A. A. Markov, who used the notion of conditional independence around 1900 to formulate simple multivariate models, now called Markov chains. A Markov

chain for q categorical variables can be represented by an undirected graph which is a single path of full lines, i.e., a sequence of nodes $i_1, i_2, \dots, i_q$ where just each consecutive pair is connected by a full line.

Compared to the larger class of loglinear models the graphical models are attractive for three main reasons: (a) the interpretation of a model can be much simpler if independencies are taken into account compared to only loglinear interactions; (b) parameter estimation can be strengthened by basing it on a sequence of possible small contingency tables which depends on the unique decomposition of a given graph into its prime graphs, i.e., into graphs which in a graphical sense cannot be split any further; and (c) a powerful separation criterion permits one to read all independence statements implied by a model directly off the graph.

This separation criterion for undirected graphs takes any three nonoverlapping subsets a, b, c of nodes in the graph of which the set c may be empty. Then the corresponding variables X_a are conditionally independent of X_b given X_c if every path from a to b has a node in c . This holds provided two conditions are satisfied: all pairwise independence statements specified with the missing edges of the undirected graph hold, and a technical positivity condition is fulfilled. One sufficient condition for checking the latter from an observed contingency table is that there are no zero frequencies.

In graphical models represented by undirected graphs all variables are treated on an equal footing, i.e., response variables and possible explanatory variables are not distinguished. This is appropriate for various different types of data, for instance for symptoms of a given disease, for items which are to measure slightly varying aspects of a particular attitude or of a behavior of persons, or for several aspects which all might contribute to a specific risk factor. However, for much multivariate research some response variables are of primary interest. There is in addition a set of background or context variables, and there are possibly sequences of intermediate variables which play the role of potential response variables to some and of potential explanatory variables to other variables under investigation. Typically, not all variables are only categorical; some are nominal, some are ordinally scaled, and still others are quantitative measurements. Subsets of variables may have to be considered as joint responses instead of univariate responses.

3. Questions Regarding Applications and Statistical Research

The original formulation of graphical models has already been extended in various directions to accommodate the need posed by different substantive research questions, to describe properties of different classes of independence models, and to prove corresponding results.

For research in different substantive fields such as genetics, economics, social, and life sciences, some important questions arising in applying graphical Markov models are:

(a) How do graphical Markov models relate to techniques applied traditionally in the given field? What additional possibilities do they offer to gain better insight into any given research question?

(b) How can a given graphical Markov model be best fitted to data? Is software available which permits fast application as well as comparison and integration into existing techniques and strategies of data analysis?

(c) Which of the associated new results in statistical theory are most useful for the substantive research at hand? Are case studies available in which graphical Markov models have been used fruitfully?

(d) Are criteria available to decide among alternative types of model to be considered for a given set of data?

The answers must depend partly on the specific substantive research questions. Also, the kinds of research questions which may be analyzed in the context of graphical Markov models widen as more theoretical results become available, and as new classes of model are developed. But many positive answers have become available, especially since, as described in Sect. 2 for general loglinear models, representations interpretable in terms of conditional independencies often result as a subclass of widely used tools for data analysis.

The following are some of the associated research questions in statistical theory.

(a) Which sets of independence statements can be specified within different classes of independence graphs? In particular, what are the defining independencies, i.e., what are (possibly alternative) sets of independence statements by which a given graph can be characterized? Which independence statements are implied in addition to those in a defining set?

(b) Under what conditions are two different graphs independence-equivalent; that is, when do they characterize the same independence structure, so that they imply an identical list of independence statements?

(c) Under what distributional assumptions do all models in a given class of independence graphs exist? When are they not only independence-equivalent but when are they also distribution-equivalent; that is, when do they define the same joint distribution?

(d) For which graphical Markov models do unique maximum-likelihood estimates exist? With which types of algorithm can maximum-likelihood estimates or parameter estimates in posterior distributions be obtained? When can large estimation problems be decomposed into smaller, simpler ones? When can complicated estimation problems be embedded in larger, simpler ones?

(e) Can the implied independence structure be predicted that results, for instance, after marginalizing over some of the variables and after conditioning on

other variables? When can even the strength and direction of implied conditional associations be derived explicitly?

(f) How can deterministic transformations of variables be utilized systematically in model building? Should definitions of variables, i.e., measurement questions, be integrated into the formulation of a graphical Markov model, or should they be introduced and studied separately?

While the last set of questions involve ongoing research, at least partial answers are available to all the others for most subclasses of graphical Markov models known until now. Here we give only examples of results, and point to the researchers who developed them.

4. Types of Independence Graph and Corresponding Markov Models

As mentioned before, there are a number of different types of graph which differ with respect to their defining sets of independencies. To describe them we denote by Y_v a random vector variable of d_v individual components. A graph for the node set V captures an independence structure for the joint distribution of all d_v components. For nonoverlapping subsets S and C of V , with S having at least two nodes, a graph for the node set S gives an independence structure for the joint conditional distribution of Y_S , given Y_C . When C is empty the graph is for the joint marginal distribution of Y_S . Conditional independence of Y_i and Y_j , given Y_C , is written more compactly as $i \perp\!\!\!\perp j \mid C$.

For each member of the following types of graph, there exists the corresponding independence structure for mixtures of discrete and continuous variables in some exponential family distribution, defined either locally for conditional Gaussian regressions, or globally for a general conditional Gaussian distribution or for one having no higher than two-factor interactions. For simplicity we illustrate here the different types of model only for a mean-centered vector variable Y_v having a joint Gaussian distribution and being partitioned into at most three components $V = (a, b, c)$. The overall covariance matrix as well as its inverse, the overall concentration matrix, are written accordingly, partitioned as

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} & \Sigma_{ac} \\ \cdot & \Sigma_{bb} & \Sigma_{bc} \\ \cdot & \cdot & \Sigma_{cc} \end{pmatrix}, \quad \Sigma^{-1} = \begin{pmatrix} \Sigma^{aa} & \Sigma^{ab} & \Sigma^{ac} \\ \cdot & \Sigma^{bb} & \Sigma^{bc} \\ \cdot & \cdot & \Sigma^{cc} \end{pmatrix} \quad (1)$$

where the dots denote symmetric entries.

The concentration matrix in the marginal distribution of Y_c is denoted by $\Sigma^{cc \cdot ab}$, and the one in the joint marginal distribution of Y_b and Y_c has components $\Sigma^{bb \cdot a}$, $\Sigma^{bc \cdot a}$, $\Sigma^{cc \cdot a}$. In the latter, the notation

reminds one that marginalizing is over a . Similarly, the conditional covariance matrix of Y_a is denoted by $\Sigma_{aa \cdot bc}$, and the conditional covariance matrix of Y_a and Y_b given Y_c has components $\Sigma_{aa \cdot c}$, $\Sigma_{ab \cdot c}$, $\Sigma_{bb \cdot c}$. Here the notation reminds one that conditioning is on c . Finally, the matrix of regression coefficients of Y_c when regressing Y_b on Y_c is $\Pi_{b|c}$, while in the regression of Y_a on both Y_b and Y_c the regression coefficient matrix is in partitioned as $(\Pi_{a|b \cdot c}, \Pi_{a|c \cdot b})$. For example, $\Pi_{a|b \cdot c}$ are the coefficients of Y_b . The notation reminds one that regression is also on c .

4.1 Directed Acyclic Graph

A *directed acyclic graph* of Y_v is a graph of arrows in d_v nodes without directed cycles, i.e., starting from any one node it is impossible to return to this node by following any path in the direction of the arrows. The (i, j) arrow is missing in it if

$$i \perp\!\!\!\perp j \mid \text{parents of } i. \quad (2)$$

Nodes from which an arrow points directly to node i are called the parents of i .

For a Gaussian distribution, the independencies show as zero coefficients in recursive linear equations having independent residuals. More precisely, we have equations $AY_v = \varepsilon$, with residuals ε having zero mean and covariance matrix, $\text{cov}(\varepsilon)$, being a diagonal matrix T . Further, A is upper triangular with ones along the diagonal, giving $\Sigma^{-1} = A^T T^{-1} A$. Therefore (A, T^{-1}) is a triangular decomposition of the concentration matrix and off-diagonal elements in row i and A are negative values of regression coefficients when regressing Y_i on all its potentially explanatory variables, i.e., variables ordered to have indices larger than i . Thus, the independencies show as zeros in A . If only a directed acyclic graph is given, typically more than one set of recursive equations is compatible with it, i.e., with several different orderings of the variables, an upper triangular matrix can result which reflects the same independence structure as the given graph.

However, often a unique full ordering is provided from substantive knowledge about how the data could have been generated. This was how the geneticist S. Wright introduced and used path diagrams for recursive linear relations. For other than linear relations, a generating process in terms of univariate conditional distributions still determines uniquely the edge matrix in a directed acyclic graph and a corresponding factorization of the joint density. Separation criteria for directed acyclic graphs apply therefore to any joint distribution generated in this way over a directed acyclic graph. A first path criterion was given by J. Pearl, who called it d-separation (for separation in directed graphs). Two other equivalent criteria are due to S. L. Lauritzen and coworkers, and D. R. Cox and N. Wermuth.

4.2 Types of Undirected Graph

A *conditional concentration graph* of Y_s , given Y_c , is an undirected graph of full lines in d_s nodes. The full (i, j) line is defined to be missing in it if

$$i \perp\!\!\!\perp j \mid C, S \text{ excluding } i, j. \quad (3)$$

For a Gaussian distribution, the independencies show as zeros in the conditional concentration matrix $\Sigma_{ss|c}^{-1}$. Such models have been introduced by A. P. Dempster under the name of covariance selection. The name concentration graph model for Gaussian variables is more common now. It records that for Gaussian distributions the above independence statement is equivalent to the vanishing of a concentration, which is a multiple of the partial correlation of Y_i and Y_j given all remaining variables in S and C .

The name concentration graph model does not imply that for general distributions concentrations vanish, but only that the distribution satisfies a set of independence statements of the above type. The separation criterion in Sect. 2 holds for such general concentration graph models and was given by J. Darroch, S. L. Lauritzen and T. P. Speed.

A *conditional covariance graph* of Y_s , given Y_c , is an undirected graph of dashed lines in d_s nodes. The dashed (i, j) line is missing in it if

$$i \perp\!\!\!\perp j \mid C. \quad (4)$$

For a Gaussian distribution, the independencies show as zeros in the conditional covariance matrix $\Sigma_{ss|c}$. The models are a special case of T. W. Anderson's linear in covariance models. They may not have a unique maximum likelihood estimate.

In general a covariance graph structure can capture all independencies in a joint distribution only if it has, just as a Gaussian distribution, no higher than two-factor interactions. For binary variables this holds, for instance, for the quadratic binary exponential distribution. For mixed variables it can be achieved for the conditional Gaussian distribution with all three and higher-order interactions being zero. Separation criteria were given by G. Kauermann and by Richardson (2001).

4.3 Types of Chain Graph

A *chain graph* of Y_v is a graph in d_v nodes which may be arranged in a sequence of ordered boxes $(1, \dots, k, k+1, \dots, K)$ such that there are undirected edges within boxes and arrows between boxes all pointing in one direction, e.g., from k to $1, \dots, k-1$; the arrows form no directed cycle, i.e., starting from any one box it is impossible to return to this box by following the direction of the arrows.

A *chain graph of block regressions* contains full lines within boxes and full arrows between boxes. The full

(i, j) line is missing within box k or the full (i, j) arrow pointing from a node in one of the boxes $k+1, \dots, K$ to a node in box k is missing if

$$i \perp\!\!\!\perp j \mid \text{all nodes in boxes } k, \dots, K \text{ excluding } i, j. \quad (5)$$

For three chain components in a Gaussian distribution, each independence statement corresponds to a zero entry (possibly symmetric) in one of the following concentration matrices: Σ^{aa} , Σ^{ab} , Σ^{ac} , $\Sigma^{bb|a}$, $\Sigma^{bc|a}$, $\Sigma^{cc|ab}$. The name block regression arises in spite of this concentration matrix representation since dividing all elements in row i of, say $(\Sigma^{bb|a}, \Sigma^{bc|a})$, by the element in position (i, i) , a transformation into corresponding block-regression coefficients results. Two different equivalent separation criteria were given by M. Frydenberg and by M. Studený.

Models of this type for the special kind of model for mixtures of continuous and discrete variables called CG-regressions are known to have unique maximum likelihood estimates, but it took more than 20 years before an algorithm was made available by D. Edwards and S. L. Lauritzen that converges in reasonable time.

A *chain graph of multivariate regressions* contains dashed lines within boxes and dashed arrows between boxes. The dashed (i, j) line is missing within box k if

$$i \perp\!\!\!\perp j \mid \text{all nodes in boxes } k+1, \dots, K \quad (6)$$

the full (i, j) arrow pointing from a node in one of the boxes $k+1, \dots, K$ to a node in box k is missing if

$$i \perp\!\!\!\perp j \mid \text{all nodes in boxes } k+1, \dots, K \text{ excluding } j. \quad (7)$$

For three chain components in a Gaussian distribution each independence statement corresponds to a zero entry (possibly symmetric) in one of the following matrices of covariances or of regression coefficients: $\Sigma^{aa|bc}$, $\Pi_{a|b|c}$, $\Pi_{a|c|b}$, $\Sigma^{bb|c}$, $\Pi_{b|c}$, Σ^{ca} . A separation criterion was given by N. Wermuth and D. R. Cox and by T. S. Richardson.

A *chain graph of concentration regressions* is a mixture of a concentration graph and a multivariate regression graph. It contains full lines within boxes and dashed arrows between boxes. The full (i, j) line is missing within box k if

$$i \perp\!\!\!\perp j \mid \text{all nodes in boxes, } k, \dots, K \text{ excluding } i, j \quad (8)$$

the dashed (i, j) arrow pointing to box k is missing if

$$i \perp\!\!\!\perp j \mid \text{all nodes in boxes } k+1, \dots, K \text{ excluding } j. \quad (9)$$

For three chain components in a Gaussian distribution, each independence statement corresponds to a zero entry (possibly symmetric) in one of the following matrices of concentrations or of regression coefficients: Σ^{aa} , $\Pi_{a|b|c}$, $\Pi_{a|c|b}$, $\Sigma^{bb|a}$, $\Pi_{b|c}$, $\Sigma^{cc|ab}$. For

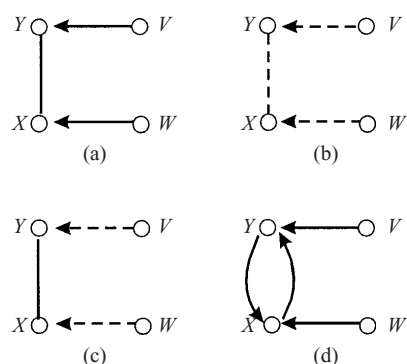


Figure 2

Simple examples of the four types of joint response graphs in Sects. 4.3 and 4.4. In each the joint responses are Y and X , and the explanatory variables are marginally independent: $V \perp\!\!\!\perp W$. The only additional independencies represented by each graph are in the (a) chain graph of block regressions: $Y \perp\!\!\!\perp W | (X, V)$ and $X \perp\!\!\!\perp V | (Y, W)$, (b) chain graph of multivariate regressions: $Y \perp\!\!\!\perp W | V$ and $X \perp\!\!\!\perp V | W$, (c) chain graph of concentration-regressions: $Y \perp\!\!\!\perp W | V$ and $X \perp\!\!\!\perp V | W$; and (d) cyclic graph: $V \perp\!\!\!\perp W | (Y, X)$.

Graphs (b) and (c) are examples of two independence equivalent graphs. In fact a distinction in independencies would emerge only with a third response Z in the left-hand box. Then a missing edge between, say X and Z , in (c) would refer to independency conditional on Y, V, W , and in (b) conditional only on V, W .

exponential families the concentration graph components refer to canonical parameters and the arrows correspond to (conditional) mean parameters. One intensively studied special case is for categorical data in which joint responses are modeled by loglinear interactions and dependencies via so-called marginal models. A separation criterion was given by S. A. Andersson, D. Madigan and M. Perlman.

Of these three types of chain graph only the second class, the graph of multivariate regressions, contains the general linear models in which every regression is of each component of the response separately on the same set of explanatory variables and the residual covariance matrix contains the remaining parameters. The class also contains seemingly unrelated regressions.

4.4 Cyclic Independence Graphs

Some fully directed graphs may contain directed cycles among a set of different variables and even for the same pair of variables. It has been shown by P. Spirtes and by J. Koster that the d-separation criterion can be applied to these graphs in unchanged form. Graphs of

traditional structural equation models may be slightly modified (Koster 1999) to have in our notation dashed lines, arrows, or combinations of both, as well as cyclic dependencies. Then a slightly extended version of the d-separation criterion can be applied to them as well. However, missing edges in these graphs need not point to any independence statement for the variable pair involved, and the meaning of edges in terms of distributional properties may have to be derived from scratch for a given graph.

Figure 2 shows examples of graphs for all four types of joint response models of Sects. 4.3 and 4.4. They have in common two connected joint responses, two marginally independent explanatory variables, and missing edges for two further pairs: (X, V) and (Y, W) .

5. Independence Graphs and Generating Processes

As mentioned above, multivariate models which have a univariate recursive generating process can be expressed in a stepwise fashion in terms of univariate conditional distributions. They are sometimes called association models generated over a directed acyclic graph. Because they can be compatible with causal interpretations they are attractive for much substantive research driven by causal hypotheses. See *Causal Inference and Statistical Fallacies*.

Multivariate regression chains are closest to a recursive generating process, since residual association may be regarded as a secondary feature. Also, after replacing each dashed line separately by two arrows starting from a common additional node, called a synthetic common source node, a directed acyclic graph results which implies the correct independence structure for the variables, in the given multivariate regression chain. It also implies covariance graph components which do not have interactions involving more than two variables.

In general, no similar simple generating processes are available for the other two types of chain graph containing chordless undirected cycles. The reasons were that (a) it is not possible to order the nodes in a chordless cycle of a concentration graph to generate it recursively in the given variables and (b) it is also not possible to enlarge the graphs by synthetic common sake nodes to generate undirected chordless cycles. At present the only ways known to generate such cycles are by conditioning on synthetic common responses or by postulating a dynamic process observed cross-sectionally in statistical equilibrium.

Experience suggests that multivariate regression chain structures fit well data from panel studies in several waves. Also, they permit prediction of individual responses without involving response variables considered on an equal footing, but only variables prior to all the joint responses. For many applications in medicine and social science this is an explicit goal of the investigation.

6. A Chain Graph as Partial Summary of Data Analysis

We now return to the variables in Fig. 1. The multivariate regression chain in Fig. 3 is derived by statistical analyses based on data provided by Judith Kappesser of the Department of Psychology at the University of Mainz. The data are for 201 chronic pain patients who have been given a three-week stationary treatment at a chronic pain clinic.

The response of primary interest in Fig. 1 is self-reported success of treatment. It is measured three months after discharge by a score which comprises several aspects of successful treatment. The background or context variables are age, gender, marital status, years of formal schooling, number of previous other illnesses, and duration of pain. There are a number of intermediate variables. Before and directly after treatment, questionnaire scores are available of depression and of self-reported type of pain, ranging from 'no pain' to 'pain as strong as imaginable.' The chronification index is a score incorporating different aspects of time, spreading of pain, use of pain relievers, and the patient's pain treatment history. The main site of pain has two categories: 'back pain' and pain on the 'head, face, or neck.'

The chain graph of Fig. 3 summarizes some important aspects of the results of analysis. It shows, in particular, which of the variables are needed for each response such that adding one more of the potentially explanatory variables does not improve prediction. Site of pain is an important intermediate variable and information on it should therefore be included in studies of chronic pain.

Some of the directions and types of dependency which cannot be read off the graph are as follows. Patients with many years of formal schooling (13 years or more) are more likely to be headache patients, while the others are more likely to be backache patients, while possibly because more of them have jobs involving physical work, and they are more likely to reach higher stages of intensity of pain before treatment. Backache patients reach higher stages of chronifi-

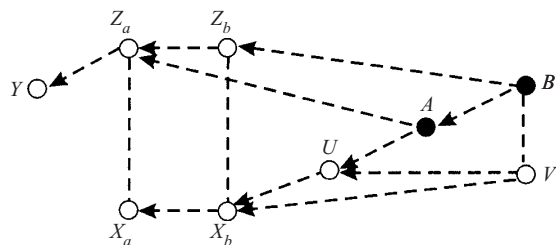


Figure 3

Multivariate regression chain of significant relations, well compatible with the data. Discrete variables are drawn as full circles, continuous ones as open circles

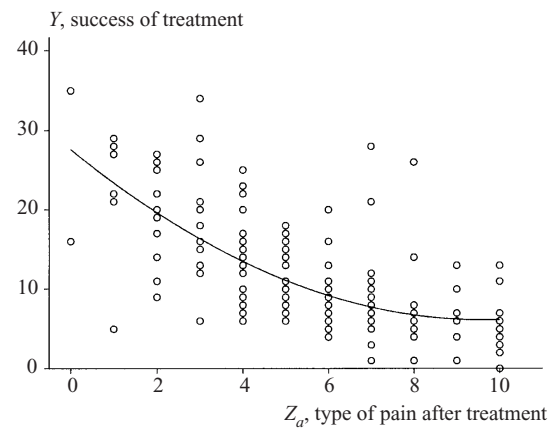


Figure 4

Form of dependence of primary response Y on Z_a

cation and report higher intensity of pain after treatment. Figure 4 shows that treatment is more successful the lower the intensity of pain, but for scores of intensity of pain of about six or higher this dependence vanishes. One path of development is that patients with shorter periods of formal schooling may get chronic backache, and patients with chronic backache get help too late and respond less well to the type of treatment offered.

Screening for nonlinear relations and interactive effects did not point to important other relations. The model is said to fit the data well because for each response taken separately no indication was found that adding a further variable would improve prediction. Had there been no nonlinear relations and no categorical variables as responses the overall model fit could have been tested within the context of available software for models of linear structural relations.

7. Suggested Further Reading

We have described some of the statistical aspects of graphical Markov models. Statistical monographs documenting the development are by Wermuth (1978), Whittaker (1990), Edwards (2000), Lauritzen (1996), Cox and Wermuth (1996). For an exposition of important different types of independence structure present in small data sets containing only linear relations, see Cox and Wermuth (1993). Models for single and joint response chain graphs can be viewed as a generalization of Sewall Wright's path analysis different from the generalization to models of linear structural relations (Bollen 1989). The two model classes have some subclasses in common, for instance linear systems represented by directed acyclic graphs and linear multivariate regression chains.

Monographs treating the use of graphical Markov models as a basis for probabilistic calculations in expert systems and for artificial intelligence applications are by Pearl (1988) and by Cowell et al. (1999). For their use in a specifically decision-making context, see Oliver and Smith (1990). For discussions in relation to causal reasoning, see the monographs by Spirtes et al. (1993) and by Pearl (2000).

See also: Bayesian Graphical Models and Networks; Graphical Methods: Presentation

Bibliography

- Bollen K A 1989 *Structural Equations with Latent Variables*. Wiley, New York
- Cowell R G, Dawid A P, Lauritzen S L, Spiegelhalter D J 1999 *Probabilistic Networks and Expert Systems*. Springer, New York
- Cox D R, Wermuth N 1993 Linear dependencies represented by chain graphs (with discussion). *Statistical Science* **8**: 204–18
- Cox D R, Wermuth N 1996 *Multivariate Dependencies*. Chapman and Hall, London
- Darroch J N, Lauritzen S L, Speed T P 1980 Markov fields and log-linear models for continuous tables. *Annals of Statistics* **8**: 552–39
- Edwards D 2000 *Introduction to Graphical Modelling*, 2nd edn. (1st edn. 1995) Springer, New York
- Koster J T A 1999 On the validity of the Markov interpretation of path diagrams of Gaussian structural equations systems with correlated errors. *Scandinavian Journal of Statistics* **26**: 413–31
- Lauritzen S L 1996 *Graphical Models*. Oxford University Press, Oxford, UK
- Oliver R M, Smith J Q 1990 *Influence Diagrams, Belief Nets and Decision Analysis*. Wiley, Chichester
- Pearl J 1988 *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA
- Pearl J 2000 *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, UK
- Richardson T S 1998 Chain graphs and symmetric associations. In: Jordan M I (ed.) *Learning in Graphical Models*. Kluwer, Dordrecht, The Netherlands, pp. 231–60
- Richardson T S 2001 A local Markov property for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*.
- Spirtes P, Glymour C, Scheines R 1993 *Causation, Prediction and Search*. Springer, New York
- Studeny M, Bouckaert R R 1998 On chain graph models for description of conditional independence structures. *Annals of Statistics* **26**: 1434–95
- Wermuth N 1976 Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics* **32**: 95–108
- Wermuth N 1978 *Zusammenhangsanalysen medizinischer Daten*. Springer, Heidelberg, Germany
- Wermuth N 1997–99 Graphical Markov models. In: Kotz S, Read C, Banks D (eds.) *Encyclopedia of Statistical Sciences*. Wiley, New York, Vol. 2, pp. 284–300
- Whittaker J 1990 *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester, UK

N. Wermuth and D. R. Cox

Green Parties

Green parties have become established political players in most industrialized democracies and are emerging in several industrializing ones. The electoral strides of green parties, especially in Western Europe, have resulted in a situation unimaginable a decade ago: green parties in many national parliaments, green ministers sharing power in several European national governments, greens in the European Parliament forming one of the largest party groups, as well as sustained green representation at subnational levels of government. Yet greens' electoral fortunes are decidedly uneven: electoral success in northern countries contrasts with near invisibility in the South. Moreover, electoral success has brought new strategic dilemmas for the greens as they struggle to maintain their alternative 'green' credentials while joining parliaments and supporting, if not forming, national governments.

1. Emergence

The emergence of green parties needs to be understood in the context of a broader shift from industrial to 'postindustrial' politics. For younger, better-educated citizens in particular, the prosperity and rise in living standards enjoyed in the postwar era formed the backdrop for a change from materialist to 'post-materialist' values, or from 'old politics' to 'new politics' (see Inglehart 1977, Poguntke 1993). As the basic material needs (food, shelter, etc.) of a far higher share of the population were being satisfied, political attention shifted from materialist concerns to 'quality of life' issues, such as enhanced political participation, gender and racial equality, and, perhaps above all, environmental protection.

Growing environmental concern among the public was heightened by the continued degradation of the environment and increased media attention of environmental issues. Toxic chemical spills in major rivers, air pollution alerts in cities, and oil slicks in pristine coastal waters further increased public anxiety, especially in Europe and the USA. Meanwhile, the accelerated construction of nuclear plants in the wake of the energy crisis, as well as the increased deployment of nuclear weapons, awakened or reinforced fears about the safety of nuclear power, the problems of disposing nuclear waste, and fears of nuclear annihilation.

To many citizens in advanced democracies, none of these concerns was adequately addressed by existing parliamentary structures. A proliferation of extra-parliamentary, nonpartisan citizens' movements emerged in the 1960s and 1970s to protest issues of peace, ecology, and women's rights. These new social movements served as the antecedents for a new type of

'green' party, ideologically focused on new politics issues and an alternative paradigm of politics. Vowing to be different from mainstream parties in beliefs and practice, these new parties embraced a 'green' ideology which profoundly challenged conventional politics.

2. Green Party Ideology

The core green values forming this ideology (see *Environmentalism*) encompass a wholesale critique of advanced industrial society and advocate instead a 'sustainable society' (Dobson 1995) based on decentralization and reduced consumption of resources and material goods. For greens, the postindustrial 'good life,' revolving around economic growth and globalization, sophisticated technology, expanding services, and material goods needs to be replaced by a simpler lifestyle offering a cleaner planet, less resource depletion, social equality, a nature-centered spirituality, and local action to solve global problems.

Green parties are characterized not only by their values and beliefs, but by the way they express and pursue these beliefs. Above all, greens emphasize 'grassroots democracy': a decentralized, direct democracy within a radically participatory society. Policy decisions should be made at the lowest possible level and matched with varying local requirements. Diversity is not only tolerated, but celebrated. The organizational structure of most green parties thus tends to be decentralized and nonhierarchical, offering a wide range participatory channels to members, gender equality, and significant autonomy to local branches (Müller-Rommel 1998).

Green party values and practices do not represent a coherent or consistent ideology. Demands for grassroots direct democracy do not always mesh with the need for a sustainable society and the imposition of regulations such a society may require. Structures imposed to deter party hierarchy (e.g., midterm rotation of posts, refusal to appoint party leaders) have proved unworkable. Nor are greens a coherent or monolithic set of thinkers or actors. In particular, several ideological and strategic divisions characterize green parties.

3. Typology of Green Parties

The primary ideological distinction separating green parties is between 'purist' and 'rainbow/leftist' parties (Müller-Rommel 1993). Some refer to this divide as 'green-green' vs. 'red-green' parties. Green-green parties, such as *De Groenen* in The Netherlands, adopt radical environmental policies, reject leftist ideology as too proindustrial, and seek to create a clear distance between their views and views of mainstream parties, left or right. Red-green, or 'rainbow' parties, include the German Greens and the Dutch Green-Left. These

parties link ecological issues with leftist concerns of social equality and a critique of capitalism. The predominance of the latter parties has led some writers (Kitschelt 1989) to view green parties merely as an extension of a broader New Left ideology, in which environmental themes are adopted as one theme of many. This view neglects the existence of purist parties or purist elements within green parties. More importantly, it neglects the central role of green ideology and practice which defines green parties and separates them from their leftist counterparts. Green parties are best understood not merely as New Left parties, but as manifestations of an 'ecological cleavage' (Franklin and Rüdiger 1995). Despite their differences in emphasis, all green parties have broadly similar core ecological concerns and demands which serve as a pivotal part of their agenda and political identity.

Another broad distinction between green parties is their strategic orientation: the extent to which they embrace parliamentary, mainstream tactics, or favor instead more radical unconventional means to achieve their aims. All green parties pride themselves on alternative strategies, but the intensity of 'radicality' differs across parties. Generally, 'purer' parties eschew traditional forms of political activity such as party alliances, coalitions, or deals with mainstream parties. The small Dutch *De Groenen* party fits this bill. Other parties do not. For example, the French Greens (*Les Verts*) while originally adopting a 'purist' or 'green-green' ideology, were willing to engage in traditional forms of political activity such as participation in presidential elections. The strategic divide applies not only between but also within parties, as illustrated by the tension between German *Realos* (pragmatists) and *Fundis* (fundamentalists). These cleavages are damaging but not debilitating. Despite the obvious difficulties in pulling together different strands and types of green parties, the Green Group in the European Parliament has managed to exercise significant policy influence on issues such as auto emissions and biotechnology controls (see Bomberg 1998).

4. Electoral Fortunes

The varied electoral fortunes of green parties can be explained as a product of the wider political and economic climates, as well as more specific opportunity structures in which these parties operate. Examining the macroeconomic environment in which parties develop is a good starting point for explaining green party success. It is widely recognized that green parties fare better in 'postmaterialist' countries with higher levels of economic development and lower unemployment. Contrast green party emergence and development in the industrialized north with green parties' struggle to emerge and survive elsewhere. But these features are less able to explain differences among developed industrialized states (e.g., why do

green parties fare better in Ireland than in the USA?). More specific opportunity structures better account for such variation.

Electoral laws are often cited (especially by green parties themselves) as a major factor in determining electoral success. Certainly proportional representation (PR) has facilitated the development of new parties in several countries (e.g., Belgium and The Netherlands) while the first past the post electoral system as operates in the USA or British national elections has made it difficult for any small party to succeed unless they are regionally concentrated. But unsuccessful green parties also exist in countries with PR systems (Denmark, Greece) and other factors may be more important in explaining green success. A federal structure offers parties the opportunity first to articulate their demand at subnational level where they can gain experience, publicity, and often funds before launching a nationwide campaign (Müller-Rommel 1998). Green parties have been more successful in federal states (Germany, Belgium Austria, Switzerland, Australia) than in unitary states, and the weak US Greens may be the exception that proves the rule.

To understand change in green party fortunes, other explanations are needed. Evolving party systems help explain the swings in green party support, which often has floundered when major established parties (British Labour Party, US Democrats, Australian Labor Party) or other smaller protest or radical parties (such as the PPR in The Netherlands) either share or poach the ideas of greens and adopt green policies in deed or at least in word. Similarly, green parties face competition from other green actors—environmental interest groups or protest movements—who absorb green demands without themselves seeking government office. To illustrate the point, environmental awareness in the UK and USA is not noticeably lower than in countries with strong green parties, but this support is reflected in swelling membership of environmental interest groups rather than green parties (Rootes 1995).

These various explanations may be supplemented by an examination of political culture (compare France's embrace of radical culture and parties with the UK's tradition of political mediation and compromise) and episodic factors (environmental disasters, rise of new controversies such as genetically modified foods). Finally, a focus on agency and structure is required. Whatever the structural constraints in which they must perform, green parties shape their own fortunes by their (in)ability to develop and articulate a coherent party platform, run a professional electoral campaign, and avoid internal combustion through infighting and personality squabbles.

Taken together these contending explanations provide a plausible if nuanced explanation of green party development and change. Certainly green parties have developed and changed dramatically since their emergence in the 1970s. By the early twenty-first century,

green parties were represented in 15 national parliaments (mainly in Western Europe but also in the East), as well as in supranational organizations and hundreds of subnational and local assemblies. Moreover, by the late 1990s greens held ministerial positions in several major European countries. Most successful by far have been the German Greens (*die Grünen*) who first entered federal parliament in 1983, formed a governing national coalition with Social Democrats in 1998 and took on federal ministerial posts, including that of German foreign minister (Joschka Fischer). In late 1999 German Green Michaela Schreyer was the first green to be appointed as a member of the European Commission, the European Union's powerful executive.

5. Conclusion

The greens' success has clear policy implications, especially on issues of nuclear power, ecological tax reform, and citizenship rights (see Collier and Golub 1997). But success also has implications for green parties themselves. Greens have always faced a unique 'strategic conundrum' (Bomberg 1998) arising from their unique beliefs and movement roots. Put simply, how can they reconcile their radical alternative politics with participation in mainstream or 'grey' parliamentary and government structures? Throughout the 1990s most green parties shed their radical cloth in an attempt to capture votes, even at the expense of green party unity and purity. Most were rewarded with electoral success well beyond what had been imaginable in the 1980s. The price to pay has been tortured internal debates about strategy, and new questions about green party identity and purpose. Today the key questions facing green parties revolve around not whether to embrace power, but what to do with it. More specifically, green parties face three new challenges in the new millennium: first, how to carve out a policy niche as established parties and governments become wiser to green demands, and as green concerns themselves appear more mainstream. Second, how to take green ideas beyond the confines of rich industrialized states into Eastern Europe and the developing world where green parties remain marginal and environmental problems acute. Third how to ensure that the broader role of green parties—as consciousness raisers, agitators, conscience of parliament and politics—is not sacrificed on the altar of electoral success.

Green parties have come a long way since their emergence and development in the 1970s and 1980s. They have become established players able to shape party competition, government formation, and government policy. But this very 'establishment' carries risks for a party whose core values and identities depend mightily on their ability to challenge the conventional order, to agitate and to annoy. For most green parties,

the greatest fear is not electoral decline so much as the prospect of becoming a party with parliamentary platform, ministerial voice, but nothing new to say.

See also: Environmental Change and State Response; Environmental Planning; Environmental Policy; Environmental Policy: Protection and Regulation; Environmentalism, Politics of; Greening of Technology and Ecotechnology; Participation: Political; Parties: Litigants and Claimants; Party Systems; Political Culture; Political Parties; Political Parties, History of; Third Parties: United States

Bibliography

- Bomberg E 1998 *Green Parties and Politics in the European Union*. Routledge, London
- Collier U, Golub J 1997 Environmental policy and politics. In: Rhodes M, Heywood P, Wright V (eds.) *Developments in West European Politics*. Macmillan, Basingstoke, UK, pp. 226–41
- Dobson A 1995 *Green Political Thought*, 2nd edn. Routledge, London
- Franklin M, Rüdig W 1995 On the durability of green politics. Evidence from the 1989 European election study. *Comparative Political Studies* 28: 409–39
- Inglehart R 1977 *The Silent Revolution. Changing Values and Political Styles Among Western Publics*. Princeton University Press, Princeton, NJ
- Kitschelt H 1989 *The Logic of Party Formation: Ecological Politics in Belgium and West Germany*. Cornell University Press, Ithaca, NY
- Müller-Rommel F 1993 *Grüne Parteien in Westeuropa. Entwicklungsphasen und Erfolgsbedingungen*. Westdeutscher Verlag, Opladen, Germany
- Müller-Rommel F 1998 Explaining the success of green parties: A cross-national analysis. *Environmental Politics* 7: 145–54
- Poguntke T 1993 *Alternative Politics. The German Green Party*. Edinburgh University Press, Edinburgh, UK
- Rootes C 1995 Britain: Greens in a cold climate. In: Richardson D, Rootes C (eds.) *The Green Challenge. The Development of Green Parties in Europe*. Routledge, London, pp. 66–90

E. Bomberg

Green Revolution

The term 'green revolution' was applied to land reform in Eastern Europe in the 1930s and recently to the attempts to persuade governments and business to adopt policies and practices that are not inimical to the environment. But the term is most commonly used in the social sciences to describe the changes that took place in farming in Asia after the introduction of semi-dwarf wheat and rice varieties in the mid-1960s. If planted on irrigated land with the use of chemical

fertilizers and pesticides these crop varieties gave higher yields and greater profits than traditional ones. These innovations led to a substantial increase in the yield of wheat and rice, which was sustained until the 1990s when the rate of increase of yield and output declined; there may be difficulty in keeping cereal output ahead of population growth in the twenty-first century.

1. Early History

The period immediately following World War II in South Asia was characterized by rapid population growth, food shortages, and poverty. A great variety of wheat and rice varieties were in use, each adapted to their local micro-environment; little chemical fertilizer was used and no pesticides, and yields were low. In 1966 a semi-dwarf variety of rice known as IR8, bred at the International Rice Research Institute in the Philippines, was distributed amongst farmers in parts of India, whilst a semi-dwarf wheat originally bred in Mexico was also introduced. Where these High Yielding Varieties (HYVs) were adopted, yields and output increased dramatically. This led W S Gaud to state in a lecture in the USA in 1968: 'this and other developments in the field of agriculture contain the makings of a new revolution. It is not a violent Red Revolution like that of the Soviets ... I call it the Green Revolution' (quoted in Dalrymple 1979). The importance of the Ford and Rockefeller foundations and of the US government in sponsoring the breeding programs in Mexico and the Philippines meant that much of the criticism of the new technology and its alleged consequences was ideological. The term 'revolution' was perhaps unfortunate. The breeding of higher-yielding or disease-resistant varieties had been going on in Europe, North America, and parts of Asia since the rediscovery of Mendel's work at the beginning of the twentieth century, whilst the use of chemical fertilizers and applications of pesticides had been practiced since the 1920s in Western Europe, although adopted at far greater rate after 1945. More relevant were the developments in Japan.

Much Japanese rice had long been irrigated, but the area irrigated was increased in the 1890s when the most successful existing rice varieties were selected and their seed distributed to farmers. Chemical fertilizers were also applied. By the 1930s, rice yields in Japan and its colonies Korea and Formosa were substantially above those elsewhere in Asia. Even in South Asia there had been advances in technology before 1966. Indian grain output increased by 62 percent before 1950 and 1964, and in Pakistan by 59 percent (Farmer 1981). Although rice yields in China in the 1950s were low compared with the rest of East Asia, semi-dwarf rices were bred there independently of the research in the Philippines and distributed to Chinese farmers before IR8 was released.

2. *The Nature of High Yielding Varieties (HYVs)*

The principal characteristic of the HYVs were their high response to chemical fertilizers. They were also short with a stiff straw, and so, unlike the tall traditional varieties, did not lodge with higher yields. They were insensitive to photoperiodicity and matured in about 110 days rather than 180 days; it was thus possible to grow two or even three crops in a year. The yield potential of these varieties was greater in the temperate regions of Asia and in the dry season in the monsoon region than in the humid tropics, because of the longer hours of sunshine and hence the greater potential photosynthesis available to the plant. The highest yields are achieved in areas of well-managed irrigation and with adequate amounts of chemical fertilizers. At first the new varieties were highly susceptible to disease, and pesticides were necessary to obtain the high yield potential, but later varieties have been bred with a greater resistance to pests. Breeding of new varieties of rice and wheat has continued since the 1960s, increasingly in research stations in the countries where they are to be grown. High yielding varieties of maize have also proved successful and to a lesser extent new varieties of sorghum. Since the 1970s Chinese farmers have used hybrid rice seed. The new varieties introduced since the first HYVs are now often referred to as MVs (modern varieties).

3. *The Diffusion of the Modern Varieties*

The inputs of the Green Revolution are divisible: seed, fertilizer, and pesticides can be bought in small amounts and so in theory could be obtained by farmers with only very small holdings. In practice however, the first adopters were generally occupiers of larger farms with better credit facilities and a greater awareness of the new technology. By the 1980s however, there was near universal adoption in irrigated areas with no differentiation of inputs by farm size or tenure. The principal incentive to adopt was the greater profit per hectare, in spite of the increased unit costs. At first the use of fertilizers and seeds were subsidized, and the improvement and extension of irrigation was undertaken by government. In the 1980s however, subsidies

began to be withdrawn, and the further extension of irrigation had become very expensive. The introduction of modern varieties began at different times in different countries and different regions. Rice was introduced into India, for example, in 1965 but not in South Korea until 1972. The rate of adoption also differed; by 1976 four-fifths of all the rice in China was sown with modern varieties, but only 7 percent in Burma, 11 percent in Thailand, 14 percent in Bangladesh. National differences persisted in the 1990s (see Table 1).

Within countries there have also been major regional differences in the rate of adoption; thus in India in 1990, 95 percent of the rice in Punjab and Tamilnad was sown with modern varieties, but only one-third in Bihar and Kerala (Singh 1997). By the 1990s some of the major international differences in the rate of adoption had been reduced, but a greater proportion of Asia and Latin America's wheat was sown with modern varieties than Africa or the Near East, whilst more of China's was in new varieties than the rest of Asia (see Table 1). But of course there is a difference in the proportion of each continent sown in wheat, rice, maize, and other crops. Little wheat or rice is grown in tropical Africa, so the proportion of these crops in modern varieties is of little significance. This distribution explains why, in the mid-1980s, modern varieties were planted on less than one-third of the total cereal area in the developing countries: only 1 percent in Africa, 22 percent in Latin America and 36 percent in Asia (Tangley 1987).

4. *The Consequences of the Green Revolution*

In the 1950s and 1960s it was feared that future population growth would outrun cereal output in Asia and increase malnutrition and hunger. This has not occurred. The rate of increase in output of both rice and wheat in Asia exceeded population growth rates between 1966 and 1995. Population increased 76 percent but rice output doubled; and the real price of rice has fallen steadily since the 1950s, halving between 1970 and 1990. Cereal consumption per capita has increased by 25 percent, imports have been greatly reduced, and grain reserves built-up. The increase has

Table 1

The diffusion of modern varieties—percentage of crop area sown in modern varieties

	1970	Rice 1976–7	1991	1970	Wheat 1976–7	1990	Maize 1990
Sub Saharan Africa	4	3	n.a.	5	23	52	43
West Asian and North Africa	0	4	n.a.	5	17	42	53
Asia (excluding China and West Asia)	12	30	67	42	72	89	45
China	77	80	100	n.a.	25	100	90
Latin America	4	13	58	11	41	82	46

Source: D Byerlee 1996
D G Dalrymple 1979

Table 2
Cereal output in Asia

	Rice: million tonnes					
	1948–52	1961–5	1971–3	1979–81	1987–9	1995–7
Japan	11.9	16.4	15.1	13.3	12.9	12.9
South Korea	2.9	4.8	5.6	6.8	8.0	6.8
China	58.5	86.0	108.1	145.7	175.8	194.3
Indonesia	9.4	12.4	19.0	29.6	41.7	50.5
Philippines	2.8	4.0	5.2	7.9	9.0	11.3
India	33.4	52.7	63.4	74.6	99.7	121.4
Bangladesh	12.4 ¹	15.0 ¹	16.1 ¹	20.1	24.2	26.6
Thailand	6.9	11.3	13.3	17.0	20.8	21.7
	Wheat: million tonnes					
	1948–52	1961–5	1971–3	1979–81	1987–9	1995–7
China	18.1	22.2	34.1	59.1	87.4	111.8
India	6.0	11.2	25.0	34.6	48.2	65.9
Pakistan	3.7	4.2	7.0	10.8	13.0	16.8

¹ includes Pakistan

Food and Agricultural Organization *Production Yearbooks*, various years

Table 3
Crop yields in Asia, 1948–1997

	Rice: tonnes per hectare					
	1948–52	1961–65	1971–75	1979–1981	1987–89	1995–97
Japan	4.0	5.0	5.7	5.6	6.1	6.4
South Korea	2.8	4.1	4.7	5.5	6.4	6.5
China	2.1	2.8	3.1	4.2	5.4	6.2
Indonesia	1.6	1.7	2.3	3.3	4.1	2.8 ¹
Philippines	1.2	1.3	1.5	2.2	2.7	2.9
India	1.1	1.5	1.7	1.9	2.4	2.7
Bangladesh	n.a	1.7	1.7	1.9	2.4	2.6
Thailand	1.3	1.8	1.9	1.9	2.0	2.2
	Wheat: tonnes per hectare					
	1948–52	1961–65	1971–75	1979–1981	1987–89	1995–97
China	0.7	0.9	1.4	2.0	3.0	3.7
India	0.7	0.8	1.3	1.5	2.1	2.7
Pakistan	0.9	0.8	1.2	1.6	1.7	2.0

¹ 1.44 tonnes 1991–94

Food and Agriculture Organization, *Production Yearbooks*, various years

varied between countries (Table 2), and has been greater in wheat than rice output, but there have been dramatic increases in cereal harvests throughout the region. In the 1950s and 1960s increases in cereal output were largely due to an extension of the cultivated area; since the 1960s, overwhelmingly due to the increases in crop yields following the adoption of the new high-yielding varieties (Table 3).

The shorter growing season for HYVs has also allowed more double cropping in some parts of Asia. Yields have stagnated in the 1990s, however, particularly in Japan and Korea, which have the highest rice yields. The yield potential of the original HYVs has not been increased since the 1960s; the many new varieties bred since then have had greater disease resistance or better quality grain, but not greater yield potential. Additionally, in the high-producing irrigated rice areas, waterlogging, salinization, and disease have led in some places to a decline in yields. The fall

in rice prices and the continued increase in import prices has squeezed cereal producers, and many have left farming or turned to more rewarding crops. In the last few years there have been renewed fears for the future of cereal output in Asia.

5. Criticisms of the Green Revolution

Initially the Green Revolution was greeted with extravagant praise and high expectations. These were excessive, for it was unlikely that many farmers could reach the yields obtained on experimental farms, or that all farmers could profitably use the new techniques. A barrage of criticism followed to the present, although with changing targets. Controversy remains over the social and economic effects of the Green Revolution, nor is this surprising. There are three technical reasons why interpretations may differ. First,

national and state statistics in India and other parts of Asia are not reliable and so different interpretations of changes are possible, particularly with reference to land tenure, farm size and labor. Second, it would be surprising if there were identical responses to new technology over such a vast area and population with greatly varying physical and socioeconomic environments. Third, research produces different results at different stages of the adoption cycle of agricultural innovations. It is unlikely that the consequences of technological change in the 1960s would be the same in the 1990s.

But there are other criticisms that are a result of the unequal distribution of farm sizes and land ownership. Early critics believed that the adoption of HYVs benefited the larger farmers most, often led small farmers to sell their land, increased landlessness, and forced down real wages. It was argued that while those with larger farms—and in most of Asia more than 20 acres is a large farm and most farmers have less than five acres—were able to rapidly adopt the new technology, and make considerable profits; smaller farmers were unable to do so and often sold their holdings to larger farmers, whilst many tenants were evicted to allow landowners to farm larger units. The number of landless households increased and so forced down real wages. In some areas the larger farmers adopted machinery to replace labor, and real incomes of the landless fell. Consequently income inequality increased.

These criticisms were made most fiercely in the early 1970s. At the time it was pointed out that many of these trends predated the 'green revolution,' or had alternative explanations. Machinery was being adopted before 1965, and rural population growth alone was sufficient to increase landlessness. Since then there have been more sample studies over longer periods (David and Otsuka 1994, Hazel and Ramasamy 1991). They have shown that small farmers have been able to adopt the new techniques, seed, and other inputs; that real wages have not fallen; and that the adoption of machinery has often been the result of farm laborers seeking higher paid jobs elsewhere. Other critics pointed out that the HYVs were only profitable in irrigated areas, and that improved varieties of only wheat and rice had been bred, yet in many parts of India farmers grew other cereal crops so that the benefits of the 'green revolution' occurred only in limited regions. This was true at that time, but subsequently modern varieties were bred that gave higher yields than traditional varieties in rainfed areas and with low amounts of fertilizer. Furthermore, high-yielding varieties of other crops have been bred: high-yielding maize were bred in Kenya and Rhodesia before 1965, and now occupy half the African maize acreage, whilst in India high-yielding varieties of sorghum and millet have been bred. The early HYV rices had an inferior taste, and although subsequent varieties were of higher quality, richer consumers still

prefer the taste of traditional varieties. Other critics were alarmed that the inputs for the new farms were imported, whilst some believed that that monoculture was replacing the diversity of cropping in traditional Asian farming, particularly the replacing of protein rich pulses with cereals.

By the 1990s new fears replaced these traditional concerns. Yields were—and still are—stagnating and even declining as a result of waterlogging and poor drainage in irrigated areas, whilst chemical pollution was becoming a problem. Falling cereal prices and rising costs are now threatening the future of cereal output, whilst Asia's population may increase by 53 percent by 2030. It may be that a new technological advance will be necessary to maintain Asia's freedom from famine that was the result of the 'green revolution' of the 1960s.

See also: Agricultural Change Theory; Agricultural Sciences and Technology; Agriculture, Economics of; Environment and Development; Environmental Planning; Food Security; Greening of Technology and Ecotechnology; Technological Innovation

Bibliography

- Burmeister L 1987 The South Korean green revolution: Induced or directed innovation. *Economic Development and Cultural Change* 35: 767–90
- Byerlee D 1996 Modern varieties, productivity and sustainability: Recent experience and emerging challenges. *World Development* 24: 697–718
- David C C, Otsuka K 1993 *Modern Rice Technology and Income Distribution in Asia*. Rienner, Boulder, CO
- Dalrymple D G 1979 The adoption of high-yielding grain varieties in developing nations. *Agricultural History* 53: 704–26
- Evenson R E, Herdt R W, Hossain M 1996 *Recent Developments in the Asian Rice Economy: Challenges for Rice Research*. CAB International, Wallingford, UK
- Farmer B H 1977 *Green Revolution? Technology and Change in Rice-growing Areas of TamilNad*. Westview, Boulder, CO
- Farmer B H 1981 The green revolution in South Asia. *Geography* 66: 202–7
- Griffin K 1974 *The Political Economy of Agricultural Change: An Essay on the Green Revolution*. Macmillan, London
- Hazell P B R, Ramasamy C 1991 *The Green Revolution Reconsidered: The Impact of High-yielding Rice Varieties in South India*. Johns Hopkins University Press, London
- Karim M B 1986 *The Green Revolution: An International Bibliography*. Greenwood Press, New York
- Lipton M, Longhurst R 1989 *New Seeds and Poor People*. Unwin Hyman, London
- Pingali P L, Hossain M, Gerpacio R V 1997 *Asian Rice Bowls. The Returning Crisis?* CAB International, Wallingford, UK
- Shiva V 1989 *The Violence of the Green Revolution*. Natrasy Publishers, Dehra Dun, India
- Singh J 1997 *Agricultural Development in South Asia. A Comparative Study in the Green Revolution Experiences*. National Book Organisation, New Delhi, India

Tangley L 1987 Beyond the green revolution. *Bioscience* 37: 176–80

D. B. Grigg

Greening of Technology and Ecotechnology

‘Greening of technology’ and ‘ecotechnology’ have similar meanings. Both terms define a perspective regarding the impact of technology as a mediator and/or cause of the environmental problems in modern societies. Both imply a move toward a different kind of technology that is environmentally friendly.

1. ‘Green’ and ‘Ecological’

‘Greening’ and the prefix ‘eco’ refer to technology’s association with environmental processes and to its role, and thus humankind’s, as a cause of environmental change, from pollution to global climate change. Greening of technology and ecotechnology call for more environmentally benign technology, including systems that operate with a minimum impact on species and habitat. So both terms define a vision for change and refer to a set of criteria to be taken into account when evaluating the qualities of given technologies in terms of their impact on the natural environment, ranging from wastes and emissions to the consumption of non-renewable resources. They call for intentional technological designs to achieve the goals in question. Also in other fields these terms are used to pinpoint intended changes and policy needs, as exemplified by ‘greening of industry,’ ‘ecotourism,’ ‘greening of consumption,’ ‘eco-efficiency,’ and ‘green accounting,’ each pointing to specific areas of socioeconomic activity.

Although the term ‘green’ is used as a synonym for nature and natural in a number of situations, the close link to environmental problems is rather new and follows from the systematic and public recognition of pollution impacts in the 1960s and subsequently. In this sense, the term differs from that used in, for example, the ‘green revolution’ and its emphasis on increasing crop production from improved varieties and petro-chemical inputs (see *Green Revolution*). Indeed, some of the impacts following from the green revolution and identified by Rachel Carson (1962) in *Silent Spring* are precisely those illustrative of the double-edged character of technology and the impetus for green technology. The origins of the new uses of ‘green’ and ‘eco-’ in regard to technology have not been adequately addressed. The linkages between technology and environmental damage, however, were highlighted in the Club of Rome report *Limits to Growth*: ‘For the first time, it has become vital to

inquire into the cost of unrestricted material growth and to consider alternatives to its continuation’ (1972, p. 191) and brought to the international policy discussions in 1972 at the United Nations’ first global environment conference in Stockholm.

2. ‘Technology’

Technology is at the same time the engine of economic growth and human well-being and the immediate cause of environmental degradation. Technologies as machine embedded or enacted processes are the basis for transforming materials using energy. Industrial mass production and transport technology extract and concentrate materials, increasing energy demands and amounts of waste, both from the industrial process and from the ‘throwaway culture’ it supports. This waste includes materials that cannot be reused or recycled but remain concentrated and combined in ways that cause health hazards and reduce the stock of raw materials accessible for future generations. Such resource consumption is accompanied by impacts on the structure and function of the biosphere, be it increasing atmospheric carbon dioxide or the extinction of species.

The definition of ‘technology’ is important for the scope of its impact on nature. Technology does not only cover machines and other artifacts, but also includes processes and procedures, like the methods used in rural agrarian production and in the structure of settlements (e.g., contemporary continued urbanization leading to mega-city societies). The question extends beyond the domains covered by the notion of ‘technology,’ however, to the sociology of its use and the organization of economic processes. These systemic uses of technology are registered in such terms as technological system, or socio-technical ensemble (see *History of Technology*). Through these broader uses of the term, technology addresses such human activities as the movement, transformation, and consumption of materials and energy. Many environmental impacts from technology are not only linked to the production of technology *per se* or its basic function, but also to how production and consumption are connected (see *Industrial Metabolism*).

3. Is ‘Green’ an Attribute or a Process?

The labels ‘ecotechnology’ and ‘green technology’ are commonly applied to environmentally friendly production technologies and/or products, such as wind turbines and solar cells for renewable energy, natural textiles, and organic farming reducing, for example, the use of inputs from petrochemical industries. This use has been contested on the grounds that, in absolute terms, no technology is truly green and, in some cases, green and technological risks (or perception of risks) are confused. This is the case for genetically modified crops, for example. They are usually labeled as ‘non-

green,' in public debate. But from a different perspective, it can be argued that such hybrid crops could produce more food on less land. It is, however, not questioned that specific technologies can be less polluting and greener than others, in relative terms.

What is identified as environmentally friendly, green, or ecological is dependent on how the environmental problems and risks are perceived and prioritized. Every technology has an impact on the environment, and the choice between technologies must be assessed in terms of these impacts and the risks involved when recognizing the limits to existing knowledge. This leads the decision in the direction of a political process that has to deal with uncertainties as well as conflicting interests in the outcome.

Consequently in more sophisticated uses of the notions of 'green' or 'ecological,' they are used to characterize the direction of change based on selected measures referring to environmental performance of technologies. The notions are thereby used to establish distinctions between the outcome of different change processes and their focus.

4. Promotions and Problems

While traditional environmental policies developed and installed in the 1970s focused on the discharge of pollutants from production, the growing awareness for a need to engage industry in a constructive way resulted in new policy strategies in the later part of the 1980s. Here the strategy was to engage companies in environmental protection measures by introducing the concept of cleaner technologies and industrial ecology (see *Industrial Ecology*). The cleaner technology (could also read 'greener') approach value materials and energy reductions based on improved design and production as the highest priority, followed by substitution of nonrenewable materials, recycling, and ending with filtering and waste handling as the absolute lowest priority.

These new policies were supported by the introduction of environmental management as a new subdiscipline and the implementation of new management standards like the international ISO 14000 series and EMAS in the EU together with economic fees on pollution and waste. By focusing their interest on waste and the inefficient use of materials and energy, many companies improved their environmental performance. Again the ambiguous character of technology is illustrated: cleaner is not a simple quality of technology, but a product of a specific direction in industrial transformation.

Following these insights, environmental regulation has also changed from a focus on industrial point sources as the object of regulation (e.g., smokestack emissions) to broader aspects of societal production and consumption, such as the impacts from product use and disposal (e.g., recycling). Product lifecycle and

eco-design approaches have begun to be implemented by companies and supported by government environmental policies, especially in the developed world. The lifecycle approach addresses environmental impacts from the perspective of the complete cycle of a product's life: the production of raw materials, the transport of materials and products, the production process, the use of the product, and the final recycling or disposal of the product as waste. Consequently it introduces a rather extended obligation for companies in managing not only their own pollution and resource consumption but in being responsible for the complete chain of supplies and for the minimization of environmental impacts and resource consumption through consumption and disposal. Procedures for eco-design have as their main focus the translation of environmental impacts within the design phase and the reorganization of innovative processes to include both environmental concerns but also user perspectives and new types of expertise (Mitch and Jørgensen 1989).

The problems from which green and ecotechnology have evolved are now also registered in the term 'sustainable development,' signifying the dual needs of providing improved human well being with less environmental impacts (see *Sustainability Transition: Human-Environment Relationship*). This idea was officially recognized in the World Commission on Environment and Development's report *Our Common Future* (The Brundtland Report), which stated that 'When the century began neither human numbers nor technology had the power to radically alter planetary systems. As the century closes, not only do vastly increased human numbers and their activities have that power, but major, unintended changes are occurring in the atmosphere, in soils, in waters, among plants and animals, and in the relationships among all of these' (1987, p. 343).

The introduction of cleaner technologies has in principle shown results in reducing waste and the discharge of pollutants in quite substantial ways. But voluntary actions in industry do not guarantee the introduction of these green technologies (Kemp 1997). And these results do not satisfy those needed for a future whose material and energy demands appear to be enormous. Rather radical reductions in materials and energy consumption and also the reduction of the 'ecological footprint' of individual consumers are required. Phrased as the 'factor 4' and 'factor 10' strategies for reductions in materials and energy use, a new agenda for technological change has been outlined, including also the need for increased global wealth. This agenda will influence product designs, work organization, transport, and culture in radical ways in order to reach the result: a radical reduction in the amount of materials and energy circulated and used per produced unit of good or service (Weizsäcker et al. 1997). Such changes cannot be reached in business and industry alone, but will have to include

behavioral changes of consumers as well (see *Consumption, Environmentally Significant*). This strategy is not seen as different from or in competition with strategies focusing on industrial performance and eco-design, but as supplementary to them (Fisher and Schot 1993).

5. Involved Actors

Such greening and ecological concerns engage a large part of industry. This engagement is partly a response fueled by industrial interests in promoting and marketing their green improvements, and partly a response to pressures from governments and consumers to reduce environmental impacts. During the 1980s a number of new national and international organizations and networks were established in response to these challenges. These included cooperating networks as the World Business Council for Sustainable Development, the European Roundtable for Cleaner Production, the American Coalition for Environmentally Responsible Economics, and many others established on both regional and global levels. Consumers and environmental grassroots movements also took part in the construction of new agendas for green technologies and greener products, some even becoming involved in innovations, for example, Greenpeace's development of a low-energy refrigerator and cooperation with an industrial company about its manufacture, and other organizations' contribution to innovation in the wind turbine. The 'green' impact is also seen in the commercial use of environmental qualities in marketing products, ironically termed 'green speak' and 'green wash' when not based on substantiated environmental performance improvements.

6. Academic and Disciplinary Context

The study of 'greening' and 'ecotechnology' is not the subject of one specialist discipline but involves a large number of academic disciplines, ranging from engineering and design to social perception and innovation management. Specific subfields in the technical sciences advance green or ecological technologies, including energy-efficient cars, urban ecology, ecotextiles, organic farming, and energy efficient household equipment. Special attention has been given to methodologies assessing the impact of a product during its complete lifecycle (in short: LCA for life cycle assessment) and in searching for tools to be used as a basis for environmental prioritization. Design procedures and ways of translating and integrating environmental concerns into design practices have escalated as research interest lately showing both in the design journals and in new design manuals (*Ecological Engineering* is an example). New inter-

disciplinary research centers have also been established.

Technology studies, drawing on the experiences from the STS (Science Technology Society) field, and the studies of innovation in both management and evolutionary economics also sometimes include studies of the environmental dimensions of technology. The technology studies field, in particular, has developed a special interest in studying green technologies. But in the mainstream academic social science disciplines, as in the mainstream technology design fields, environmental concerns are often not explicitly mentioned or even left aside. Researchers have therefore built their own networks and journals and organize special sessions within the meetings of other formal organizations. One example, however, of a cross- or interdisciplinary organization, is the 'Greening of Industry' Network, established in 1990.

Among the mainstream social sciences, political science and economics have been active in regard to environmental regulation and politics lately. The interest in 'green industrial policies,' 'greening of business activities,' and 'green tax reforms' has been growing. So has the support for 'environmental economics' focusing on how to make external impacts visible in the economic system (i.e., the costs and problems caused by the economic system, but not accounted for in the price and market mechanism). Although 'technology' is often referred to in this context, in most studies both the environmental problems and the technologies stay in the 'black box,' and are not analyzed in detail.

7. Future Directions of Research

The rather dispersed character of research in 'greening of technology' illustrates two fundamental problems: the need for an interdisciplinary approach to tackle the problems, and the need for an integration of environmental concerns into a number of technical and social science disciplines. The need for interdisciplinary approaches is obvious in design and in management, where the traditional focus has been on economic priorities and organization rather than environment matters. These priorities have less to do with the character of the field of research than with the tradition of disciplines. The need for integration among disciplines is obvious, given the range of linked issues, from specific technologies, to business management, to government policy.

The complexity of the problems and interested communities is confounded by the juxtaposition of policymakers' and managers' desire for simple solutions. Such solutions are problematic, of course, especially when they are not measured by assessments of knowledge limitations and uncertainty (see *Precautionary Principle*). More generally, in studies of innovation and technological design there is a need to

develop methodologies to cope with integrating environmental concerns into design and management of processes and to deal with the conditions for use and consumption.

The development of greener technologies is not simply a question of improvements and integrations in the technical sciences and in design practices, though we should not underestimate the importance of these areas. A change in the dominant priorities for technological development is dependent on priorities set in the business sector, and changes therefore must address both management practices and societal governance. Likewise, since business is not unresponsive to public pressures and consumer actions these concerns must be addressed as well. In short, the greening of technology is a multi-sectoral and multidisciplinary subject which is likely to increase in importance as future generations cope with increasing demands to meet human needs, needs that are being redefined to include environmental quality (IHDP 2000).

See also: Agricultural Sciences and Technology; Consumption, Environmentally Significant; Ecotourism; Environmental and Resource Management; Green Revolution; History of Technology; Industrial Ecology; Industrial Metabolism; Land Reclamation; Land Use and Cover Change; Precautionary Principle; Sustainable Development

Bibliography

- Carson R 1962 *Silent Spring*. Houghton Mifflin, Boston
Club of Rome 1972 *Limits to Growth*. Universe Books, New York
Fischer K, Schot J 1993 *Environmental Strategies for Industry*. Island Press, Washington, DC
IHDP 2000 *Industrial Transformation: Science Plan*. International Human Dimensions Programme on Global Environmental Change, Report No. 12, Amsterdam
Kemp R 1997 *Environmental Policy and Technical Change*. Edward Elgar, Cheltenham, UK
Mitch W J, Jørgensen S E (eds.) 1989 *Ecological Engineering: An Introduction to Ecotechnology*. Wiley, Chichester, UK
Roome N (ed.) 1998 *Sustainability Strategies for Industry: The Future of Corporate Practice*. Island Press, Washington, DC
Weizsäcker E von, Lovins A, Lovins L H 1997 *Factor 4. Doubling Wealth—Halving Resource Use*. Earthscan, London
World Commission on Environment and Development 1987 *Our Common Future* [The Brundtland Report]. Open University Press, Buckingham, UK

U. Jørgensen

Grounded Theory: Methodology and Theory Construction

Grounded theory is an inductive methodology that provides systematic guidelines for gathering, synthesizing, analyzing, and conceptualizing qualitative data

for the purpose of theory construction. In addition, one of its creators, Barney G. Glaser (1978, 1992), has long proclaimed that grounded theory methods may be used with quantitative data. The term, 'grounded theory,' derives from its central premise: theory must be developed from systematic analysis of empirical data. This methodology consists of flexible strategies to guide qualitative data collection, and, particularly, data analysis. The strength of the grounded theory method lies in articulating: (a) logical steps for handling data collection and analysis, (b) a means of correcting errors and omissions and of refining analytic ideas, (c) tools for studying basic social and social psychological processes in natural settings, and (d) strategies for creating middle-range theories.

Prior to publication of Barney G. Glaser and Anselm L. Strauss's *The Discovery of Grounded Theory* (1967), training in qualitative research had largely occurred through an oral tradition of mentoring and immersion in a field setting. Thus, the grounded theory method represented a significant advance because it codified steps in analyzing qualitative data. The distinctive features of the grounded theory method include: (a) simultaneous data collection and analysis, (b) reliance on comparative methods, (c) early development of categories, (d) intermediate analytic writing between coding data and writing the first draft, (e) sampling for developing ideas, (f) delay of the literature review, and (g) a thrust toward developing theory.

1. The Intellectual Heritage of Grounded Theory

Grounded theory derives from the intellectual traditions of each of its founders. Strauss brought Chicago School pragmatism, symbolic interactionism, and field research to grounded theory (see also *Symbolic Interaction: Methodology; Field Observational Research in Anthropology and Sociology*). Hence, grounded theorists study processes in natural settings and invoke pragmatic criteria of usefulness to evaluate the completed study. Grounded theory methods themselves echo pragmatist and symbolic interactionist assumptions of social life as emergent and open-ended, and answer Herbert Blumer's (1969) call to study social action in natural settings. Glaser's training in survey research at Columbia University lent grounded theory its systematic approach, positivist proclivities, and procedural language. Glaser codified steps of analysis of qualitative data in an analogous way as quantitative research had been codified. His efforts culminated in an explicit statement of how to handle data analysis (Glaser 1978, Glaser and Strauss 1967). Glaser's approach assumed a knowable world waiting to be discovered, unbiased observers who are uninfluenced by preconceived logico-deductive theories of this world or by prior research about it, and a view of grounded theory

categories as arising from the data. As a result of its divergent origins, grounded theory contains elements of both positivism and constructivism.

Glaser and Strauss developed grounded theory methods at a time when quantification had gained hegemony throughout the social sciences. Theory and research had become separate pursuits. The quest for quantified research findings resulted in the waning of qualitative studies. Except for an occasional classic study, most quantitative methodologists ignored qualitative research and relegated it to disciplinary sidelines, or treated it only as a precursor to rigorous quantitative research. Qualitative research was deemed impressionistic and anecdotal, unfitting the scientific quest for quantified facts. In contrast, Glaser and Strauss (1967) argued that qualitative research could stand as science in its own right, demonstrate rigor, and generate theory.

Glaser and Strauss's arguments found receptive audiences. Their methods appealed to social scientists who wished to conduct qualitative research but lacked tools for doing it. In the late 1960s and throughout the 1970s, grounded theory methodological rationales contributed significantly to re-establishing the legitimacy of qualitative research. This method provided researchers with ready justifications for conducting qualitative studies and strong rationales that their research inquiry was systematic. Not surprisingly, diverse qualitative researchers still claim to use grounded theory to establish their credibility and the legitimacy of their research enterprise. However, their claims of using grounded theory may rest on having conducted some form of qualitative research or on only following the initial steps of the grounded theory method.

2. *Developments and Changes in the Method*

Although *The Discovery of Grounded Theory* called for conducting qualitative research and developing theory, it was less clear on details about using the method. Glaser's *Theoretical Sensitivity* (1978) set forth the most explicit statement of how to conduct grounded theory research, yet its abstract and dense writing made it resonate most with those who have had prior training in grounded theory. Since then, Glaser's position has remained consistent with his early formulations. For Glaser (1992), grounded theory builds on discovered data, avoids preconceiving interpretation of them through extant theories or categories, relies on comparative methods, and aims toward theory development. Thus, the grounded theorist compares data with data, data with concept, concept with concept, and theoretical category with theoretical category. For example, the grounded theorist compares one interview excerpt with another, an interview excerpt with a concept that may explain it, a concept with another concept and so forth.

Strauss with his co-author, Juliet Corbin (1990, 1998), took grounded theory in somewhat different directions than the early formulations. They introduce several new techniques, stress description, and include verification as part of grounded theory. They emphasize adherence to technical procedures which Glaser views as forcing data into preconceived categories rather than letting categories emerge through comparing data with data. Both Strauss and Corbin and Glaser's approaches remain objectivist because they each assume the reality of an external world, discovery of data within that world, a neutral observer of it, conceptual categories as emanating from the data, and representation of data and subjects as nonproblematic.

A constructivist grounded theory (Charmaz 1995, 2000) adopts methods of grounded theory without adhering to earlier objectivist, positivist assumptions (see *Constructivism/Constructionism: Methodology*). It places priority on the studied phenomenon rather than methods of studying it; hence, it adopts grounded theory strategies as useful tools, not as rigid prescriptions. Because these grounded theorists build reflexivity into the research process, they scrutinize their research experience, ways of knowing, and products of knowing (see *Reflexivity: Method and Evidence*). A constructivist grounded theory attends to data collection closely. It does not assume that data simply exist in an external world, awaiting discovery. Nor does it assume that an observer can enter the research scene as a *tabula rasa*. Rather, what observers see and hear depends upon their prior interpretative frames, biographies, and interests as well as the research context, their relationships with research participants, and modes of generating and recording data. The questions that observers ask of the empirical world shape data that they select from it. Similarly, constructivists view their conceptual categories as constructed through their interpretations of the data rather than emanating from them. Thus, they assume that their resulting theoretical analyses provide interpretive renderings of a reality rather than an objective reporting of the reality.

3. *Data in Grounded Theory Research*

Although grounded theorists have given priority to data analysis over data gathering, the quality of their gathered data affects the final analysis. Full, rich data provide a more thorough view of the studied topic. The kind of data obtained for a study should suit the research topic and research participants. Grounded theory methods have become closely associated with interview studies, however, they are suitable to use with varied forms of data such as ethnographic field notes, written personal accounts, and documents (Clarke 1998).

A hallmark of grounded theory studies is that data collection and analysis proceeds simultaneously. Each

provides a check on the other. Grounded theorists code and categorize data during the initial stages of data collection. This strategy helps them to shape further data collection to gather material that fits the topic and field setting. Hence, they are less likely to pursue their preconceptions about their topic and more likely to follow leads obtained through their early analysis to find new data.

Grounded theory methods are emergent; both the type and nature of data and analysis emerges throughout the research process. Before grounded theory methods were developed, researchers separated data analysis and collection phases of research. Often they discovered later that they had gathered extensive, but thin, data that contained gaps. Grounded theory strategies include returning to the field to gather more complete data that fills gaps and answers questions. Because grounded theory methods are aimed to further the theoretical power of the analysis, these methods tailor data gathering to the researcher's emerging analysis. Thus, adopting these methods makes data collection more efficient and, therefore, streamlines the research process.

4. Grounded Theory Strategies for Theory Construction

The purpose of grounded theory strategies is to develop middle-range theories. Each stage of grounded theory analysis moves the work toward theoretical formulations. The techniques involved in three major strategies of the grounded theory method, coding, memo-making, and theoretical sampling, serve to both build theory and to distinguish grounded theory from other kinds of qualitative analysis.

4.1 Coding Qualitative Data

Coding is the first phase of analysis. Rather than apply extant concepts to their data, grounded theorists create codes as they study their data and, in abbreviated terms, define what they see in it. Coding helps the researcher begin to conceptualize what basic processes occur in the research setting or situation. While coding, grounded theorists ask the essential question: 'What is happening here?' Objectivist grounded theorists believe that the most significant processes in the setting will be apparent and that research subjects will inform the researcher of them. Constructivist grounded theorists believe that the most significant processes may be liminal and taken for granted by research subjects.

Grounded theory coding includes the following characteristics: (a) a focus on action and process, (b) a practice of line by line initial coding, (c) a simultaneous involvement in coding and further data collection, (d) an emphasis on analytic development rather than description. In initial or open coding the researcher

proceeds to code line-by-line to begin breaking bits of data into specific types of action. After assessing which initial codes appear most frequently, grounded theorists then apply these codes to large amounts of data during a second stage of focused coding. Strauss and Corbin (1990, 1998) also recommend 'axial coding' which means coding for the dimensions of a category.

4.2 Memo-writing

Memo-writing constitutes the pivotal intermediate stage between coding data and drafting the theoretical analysis. Through memo-writing, grounded theorists fill out their codes and identify gaps in them. They define the code, delineate and analyze its properties, specify conditions under which it exists and changes, demonstrate its relationship to other codes, and weigh its significance for processes discovered in the field. An objectivist grounded theorist proceeds by emphasizing external characteristics of the code and the processes with which it is linked. Constructivist grounded theorists look for their own assumptions and implicit meanings, as well as those of their research subjects. Throughout data collection and analysis, constructivists remain more interpretive and less tied to overt behavior than their objectivist counterparts.

4.3 Theoretical Sampling

Theoretical sampling means sampling to develop or to refine emerging theoretical categories, not to describe populations chosen before the research begins. Thus, theoretical sampling occurs after the grounded theorist has defined and analyzed core theoretical categories through focused coding and memo-writing, but needs more data to develop, refine, and check the properties, boundaries, causes, and consequences of these theoretical categories. Thus, theoretical sampling builds precision, density, and complexity into the emerging theoretical statements and keeps them grounded in data.

When engaging in theoretical sampling, the researcher makes brief inquiries into fruitful areas for core categories of the emerging theory. Techniques of data gathering in theoretical sampling vary and are aimed to answer questions about the emerging theoretical categories. Theoretical sampling raises the conceptual level of the analysis and does so through systematic comparative analysis. The scope of the theory increases as the grounded theorist samples and compares different groups for their relevance to the theoretical category.

Gaps and questions may necessitate either returning to the same field setting or research problem or obtaining data in a different setting that possesses the characteristics the grounded theorist needs to test. Conducting comparative work across groups or field

settings allows the grounded theorist to build variation and complexity into the analysis. The emerging theory determines whether seemingly disparate groups make relevant sources to sample for theoretical comparison. For example, a study of establishing trust may sample such groups as mountaineering partners, con-men and women and their marks, and married couples. Theoretical sampling fosters a developing a full range of the properties of the category. Grounded theorists stop theoretical sampling when they have saturated their core categories and find no new data to shed light on them.

5. Current Emphases and Future Directions

The grounded theory method has significantly influenced the development of qualitative research in the social sciences and professions, particularly nursing and education. Strauss and his colleagues trained several generations of graduate students in sociology and nursing, whose students and colleagues have subsequently adopted the grounded theory method. In sociology, this method has, perhaps, had most influence in medical sociology and social studies of science (see Baszanger 1998, Clarke 1998, Star 1989). However, recent debates between Glaser and Strauss and Corbin, as well as larger epistemological debates about scientific inquiry, have intensified interest in examining and using the method. Although questions have been raised concerning objectivist premises within grounded theory, a constructivist revision of the method provides answers to them and points to new directions. In short, the grounded theory method can be expected to gain renewed support from those social scientists and professionals who value conceptual analysis of rigorous empirical qualitative research.

See also: Content Analysis; External Validity; Qualitative Methods, History of

Bibliography

- Baszanger I 1998 *Inventing Pain Medicine: From the Laboratory to the Clinic*. Rutgers University Press, New Brunswick, NJ
- Blumer H 1969 *Symbolic Interactionism*. Prentice-Hall, Englewood Cliffs, NJ
- Charmaz K 1995 Grounded theory. In: Smith J A, Harre R, Van Langenhove L (eds.) *Rethinking Methods in Psychology*. Sage, London
- Charmaz K 2000 Grounded theory: constructivist and objectivist methods. In: Denzin N K, Lincoln Y S (eds.) *Handbook of Qualitative Research*, 2nd edn. Sage, Thousand Oaks, CA
- Clarke A E 1998 *Disciplining Reproduction: Modernity, American Life Sciences, and the Problems of Sex*. University of California Press, Berkeley, CA
- Dey I 1999 *Grounding Grounded Theory*. Academic Press, New York
- Glaser B G 1978 *Theoretical Sensitivity*. Sociology Press, Mill Valley, CA

- Glaser B G 1992 *Emergence vs. Forcing: Basics of Grounded Theory Analysis*. Sociology Press, Mill Valley, CA
- Glaser B G, Strauss A L 1967 *The Discovery of Grounded Theory*. Aldine, Chicago
- Star S L 1989 *Regions of the Mind: Brain Research and the Quest for Scientific Certainty*. Stanford University Press, Stanford, CA
- Strauss A L 1987 *Qualitative Analysis for Social Scientists*. Cambridge University Press, New York
- Strauss A L, Corbin J 1990 *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Sage, Newbury Park, CA
- Strauss A L, Corbin J (eds.) 1997 *Grounded Theory in Practice*. Sage, Thousand Oaks, CA
- Strauss A L, Corbin J 1998 *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory Procedures and Techniques*, 2nd edn. Sage, Thousand Oaks, CA

K. Charmaz

Group Decision Making, Cognitive Psychology of

A group is a set of three or more persons who engage in joint activity directed at a common goal. For many groups, the goal is to make a decision. For other groups, the goal is achievement of a political, social, educational, or recreational objective. Regardless of their overall purpose, every group must make decisions so that member's actions are coordinated, and directed toward the common objective. This article presents major developments in the understanding of the cognitive processes involved in group decision making. It begins with a description of the forces leading to the current importance of group decision making, and proceeds to explain how the cognitive processes in decision making groups can be understood with a social information processing model. Next, some of the conceptual and methodological problems surrounding group decision making theory and research are covered. The theme throughout this discussion concerns differences among group members in terms of their information, beliefs, and decision preferences. The significance of two recent advances in addressing differences among group members, the Judge Advisor System and Information Sharing paradigms, is explained.

1. Groups as Information Processors

Group decision making has become increasingly important in the world of work for many reasons. One is that a greater proportion of jobs now involve cognitive work, so just as individual workers shifted from physical to cognitive work, so did groups.

Another force promoting the use of groups is the effort to increase representation and participation in organizational and societal decision making. Decisions that might have once been dictated by an authority are now made by a group representing various levels and segments of the population. Much of the current group research occurs in the context of organizational theory and practice (Sanna and Parks 1997)

The growth in the practice of using groups for decision making has been accompanied by a similar growth in the study of cognition in groups. The benefits of this research extend to all groups because all groups, regardless of their purpose, are dependent on cognitive processes for functioning. Simply put, people must coordinate their individual processes of attention, reasoning, judgment, memory, etc. during interaction in order to achieve their common goal. Group decisions are essential to coordination, and occur at many points during interaction. A group must choose its goal and the nature of the process that will be used to achieve it, select information deemed relevant, and integrate information, beliefs, etc. to determine the group's course of action.

The influence of 'real-world' practice is evident by the efforts at realism in modern group decision making research. Whereas early group research used games, trivial and 'eureka' decision problems, research tasks now capture such essential features of real decisions as the inability to know, with certainty, the quality of the decision at the time it is made. Motivational issues are taken more seriously by, for example, using incentives for making good decisions to reflect better the consequences facing persons in decision making groups. Other advances in group decision making research are the use of ongoing groups, and the study of interactions using technology (email, videoconferencing, etc.).

Part of the current scientific activity pertaining to cognitive processes in groups can be traced to the recent growth of cognitive psychology and the birth of social cognition (Fiske 1992), a subdiscipline of psychology devoted to the study of cognition about social objects (see *Cognitive Psychology: History*). Advances in these areas, combined with the increased prevalence of groups engaged in decision-making tasks, led to the modern view of conceptualizing groups as information processors. As described by Hinsz et al. (1997), key leaders in this development, the information processing approach to groups emphasizes the social processing of information, thereby integrating the group decision making and social cognition research approaches. In the social information processing view, groups exchange ideas, information, resources, strategies, etc. in the service of a processing objective. Research is directed at understanding attention, encoding, processing, storage, retrieval, feedback, and response processes in groups. Although recent group research tends to be guided by an individual-level information processing model, group and individual

level processing are not assumed to be identical, and a pervasive research agenda is to understand the qualitative and quantitative differences between individual and group processing.

A primary difference between individual and group decision making arises from the greater complexity of group level processing of information, that is, group decision making requires coordination of individual members' cognitive processes and communication of the outputs of these processes. For example, consider how the memory problem in groups requires development of common strategies for transactive memory (Wegner 1987), a shared representation of the organization of information over members. It is through transactive memory that groups are able to identify and use the unique information that members previously knew or are able to recall from group interactions. Further, groups have more complex processes because they have a number of unique problems from which individual performers are free. At the start, group members face social uncertainty about the other group members. This refers to an inability to predict the behavior of other persons, making it difficult to know the extent to which they are willing or able to work toward the group goal. Group members may have different ideas about what the group goal is or should be. If so, it is necessary to negotiate them before purposeful interaction can occur. At any point in time, groups can have social goals that detract attention, time, and efforts from the task of making a decision. Even if all members are clearly focused on a single objective directly related to their decision problem, they are likely to have differences in information, beliefs, and preferences that need to be integrated in order to reach that objective.

2. Levels of Behavior in Groups

Despite the differences between individual and social information processing, it is sometimes possible to apply 'individual' cognitive constructs at the group level. Consider the case of human judgment. Judgment is the location of an object in an abstract representation such as a scale from one to nine, or a set of categories. An example of judgment in group decision making is a committee's recommendation for a salary offer. On occasion, group members' independent judgments may agree closely enough for the purpose at hand. But often, group members have cognitive conflict, that is, differences in beliefs. In some cases, groups may begin with varying judgments but end the task with a consensus group judgment of the amount. Yet, in other circumstances, members will adhere to discrepant judgments, and the judgment for the group must be a compromise such as the central tendency of the members' judgments. Then an accurate picture of the group's judgment needs to include some indicators of the variation in the members' judgments. The

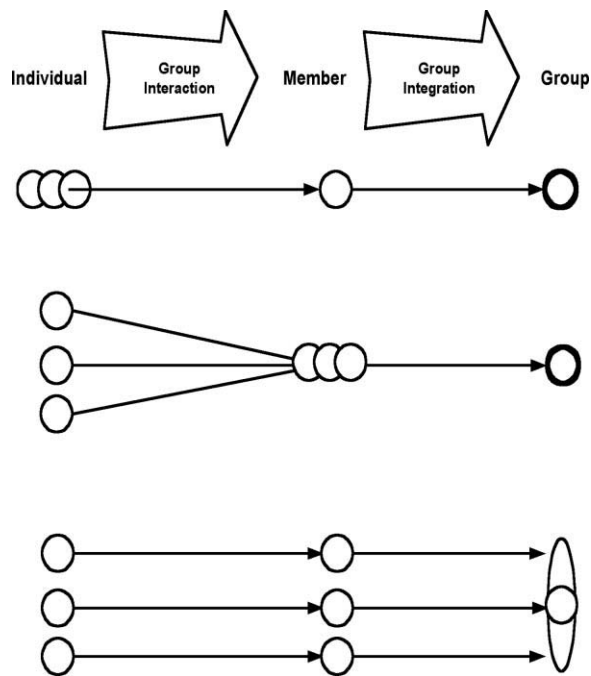


Figure 1
Three types of relationships between individual, member, and group judgments

problem is similar for cognitive tasks with discrete outcomes, such as deciding to uphold or overturn a law. When all are in agreement, there can be a group determination; if not, there can be only majority and minority member opinions. In general, we can describe members' cognitions at the group level if observations of the members of a group are indistinguishable, that is, every group member provides a response identical to that of every other group member. Under such conditions, we can speak about group goals, group confidence, group memory, and so forth. Otherwise, it is necessary to represent the diversity of group members' cognitions, with or without information on the cognitive value for the average or majority.

Figure 1 illustrates the distinction between the application of the cognitive construct of judgment to individuals, group members, and the group. Prior to interaction, there are three independent individual judgments representing individual information processing. These are similar in the first case, but different in the bottom two cases. These individual judgments or the beliefs surrounding them may be discussed during group interaction. Group interaction allows for social influence processes to transform the initial individual judgments into revised member judgments. Thus, the term 'member' is used to refer to individual members' judgments during or following group interaction to distinguish these judgments from the

initial individual judgment, whether or not they have changed. In the first case, members learn that they are in agreement on a single judgment that becomes the consensus group judgment. Member judgments have become similar in the second case, but not the third. As in the first case, the integration of member judgments into a single group judgment is straightforward in the second case. But in the third case shown at the bottom of the figure, there is no consensus group judgment, only an average judgment for the group. There may be no alternative but to use this average judgment to represent the group judgment. But an accurate picture of the group's judgment must include the persistent differences among member judgments.

3. Problems in Conceptualizing Groups

The study of cognition in decision making groups extends to many aspects of the groups behavior. In practice, groups are comprised of individuals who can have considerable cognitive conflict in their processing of information relevant to their interactions and objectives, that is, the various persons in a group may have different schemas or frames for the problem at hand, attend to different problem features, or know or recall different information (see *Mental Models, Psychology of*). Due to such discrepant views within the group, one of the first problems encountered by group researchers is appropriately representing a group, that is, research needs to begin with a model of the group that describes features of its structure and function that are important to the research question. This is often difficult because the members themselves can differ with respect to their own mental representations of the group. They can lack agreement in their beliefs about who is in the group, what its goals are, and how the group should proceed to achieve them.

There are four ways in which group researchers can deal with the problem of group representation. The first is to use the researcher's model, with the implicit assumption that it represents not only the reality of the group, but also the members' mental models of the group. For example, a group may be characterized as egalitarian without any assessment of members' beliefs about status and authority within the group. Second, the researcher can elicit the perspectives of a representative of the group. Although it is often convenient to collect information from only one group member, that person may not accurately represent the entire group. A third approach to the problem of group representation is to identify and use members' cognitive representations of the group. Typically these are represented mentally by a few discrete structures or categories. A final solution is to impart a model to group members prior to group formation and interaction. This latter option is especially popular in the laboratory study of *ad hoc* groups. With this approach,

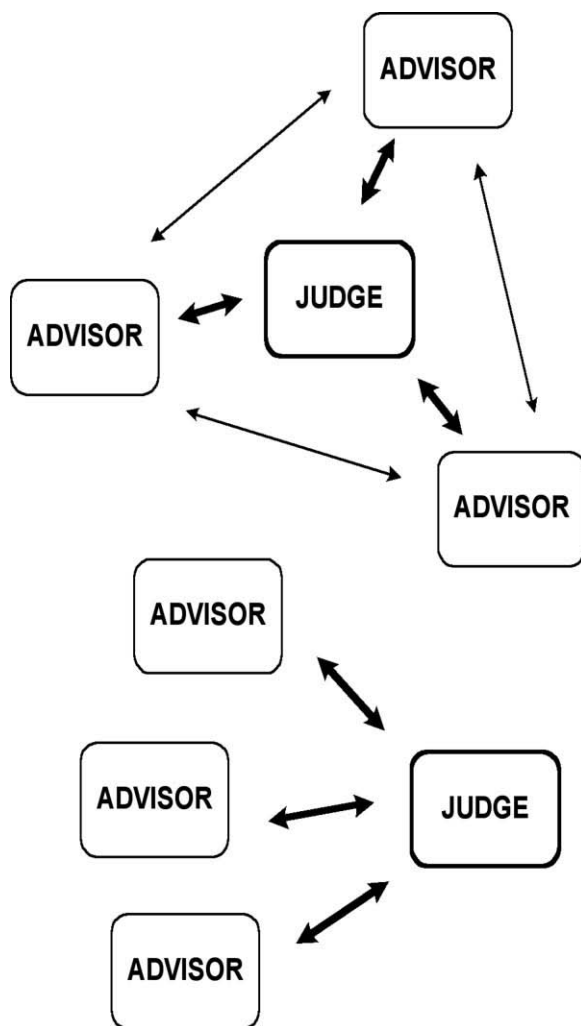


Figure 2
Illustrations of Judge Advisor Systems showing role differentiation in groups

group members may be informed that they were selected for the group because of their knowledge in a given domain, and instructed to discuss a problem until they reach a consensus on its solution. The advantage of this last approach is that data collection begins with a common understanding within the group—and between the group and the researcher—about the group's composition, task, and procedures.

One aspect of groups in natural settings that is not captured by most laboratory research is role differentiation (see *Decision Making (Naturalistic)*, *Psychology of*). What may appear to an outsider or the researcher as five persons sharing information and power to achieve their goal may be better represented as something else. Many groups are structured with a

leader who is authorized to act on behalf of the group, or with a single member who emerges to take on the role of leading the group to its goal. In many circumstances, information processing becomes social when an individual seeks advice from others. The traditional group model does not adequately describe either the group with a leader or the individual seeking advice from others. A recent paradigm that accounts for role differentiation reconceptualizes the group as a Judge Advisor System. The Judge Advisor System, illustrated in Fig. 2, is a model of social decision making that was designed to distinguish between the member responsible and accountable for making a decision, known as the Judge, and the other members, called Advisors, who provide input to the Judge (Sniezek and Buckley 1995, Sniezek and van Swol 2001). The top diagram shows a group where each member can interact with any other, but the Judge is the leader who is responsible for the goal. In the bottom diagram, the Judge is also responsible for the goal, but seeks and interacts with a number of independent Advisors to elicit resources needed for goal attainment. Whenever a decision maker interacts with another person in the course of making a decision, the process becomes social, and cannot be portrayed accurately by individual decision making theory and research.

A central cognitive activity according to this view is information search. The Judge must decide to solicit advice, whom to consult, and when to stop. Research on advising in decision making has become an international scientific activity (e.g., Harvey et al. 2000). Although similar to information search from non-social sources, this social information search differs in that the social sources are capable of strategic and reactive behavior in the process.

4. Cognitive Conflict in Groups

At this point the reader can reasonably wonder why cognitive tasks should be done by groups and not a single individual. With so many coordination costs groups must be much more inefficient than individuals. Nevertheless, there are many reasons for using groups for cognitive tasks even when the costs of coordinating members' difference are high. Some of these reasons are unrelated to the task itself; for example, democracy requires broad representation of views regardless of costs. In some situations, the task is simply not divisible; multiple persons are required. Yet, the primary motivation for using groups to make decisions is that the problem is important, and there is uncertainty about the quality of the various decision alternatives.

Ironically, the very same cognitive conflict in group inputs that creates challenges often provides the rationale for the use of groups. With a diversity of inputs, there is the potential for the group to follow the

one that best serves members' collective interest. Put simply, there are more ideas and strategies from which to choose. The best input in a group will, by definition, be superior to that of the average member, and superior to those of most members. For the most part, groups performing cognitive tasks have been found to produce outcomes that are better than those of the average member, though not as good as those of the best member. The extent to which the group outcome is better than that of the average individual input increases with the variation in group members' inputs. An important point is that this result has been obtained even when analyses control for the initial bias in group members' inputs. This assures us that the benefit of input variation cannot be attributed to groups simply discounting their poorest member.

In some cases, group outcomes can be even better than those of any group member (Sniezek and Henry 1989). Clearly the potential for such dramatic improvements in results can greatly outweigh the inefficiencies and other costs of group information processing. However, this potential benefit comes with the risk that groups can sometimes perform below the level of the poorest individual member. Therefore, a more realistic picture is that groups have the potential to be superior information processors compared to individuals working independently, but they simultaneously pose the risk of costly, inefficient, and seriously inferior performance.

Heterogeneity of inputs in a group is necessary but not sufficient for process gain. A second requirement is that those inputs be shared with other group members. One stream of research on information sharing in groups, originated by Stasser and co-workers, identifies some barriers to the sharing of information that is unique to individual group members (Stasser and Titus 1985). For starters, there is a sampling advantage for common information that means it will be more likely to be discussed in the group than unique information. Even if both common and unique information is discussed, groups tend to weight common information more than unique information. Unique information that is contributed to discussion is influential only on the judgment of the group member who contributed it. While the reduction of belief heterogeneity is necessary for group consensus, it is not always easy. This is especially true when disagreements go beyond differences in preferences to differences in preference structures. In the latter case, members have cognitive conflict, that is, differences concerning the interpretation or importance of information.

The picture that emerges from research on cognitive processes in groups shows a series of challenges unique to social information processing. Most of these problems arise from the fact that group members are not replications of an individual but varying individuals. We have seen that the persons comprising a group can vary in their beliefs about the nature of the group and its goals as well as their processing objective, focus of

attention, encoding format, and memory processes. The essence of the group challenge is to reconcile member differences so that coordinated action is possible.

See also: Group Decision Making, Social Psychology of; Group Processes in Organizations; Group Processes, Social Psychology of; Group Productivity, Social Psychology of; Groups, Sociology of; Organizational Decision Making

Bibliography

- Fiske A P 1992 The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review* **99**: 689–723
- Harvey N, Harries C, Fischer I 2000 Using advice and assessing its quality. *Organizational Behavior and Human Decision Processes* **81**(2): 252–73
- Hinsz V B, Tindale R S, Vollrath D A 1997 The emerging conceptualization of groups as information processors. *Psychological Bulletin* **121**(1): 43–64
- Sanna L J, Parks D C 1997 Group research trends in social and organizational psychology: Whatever happened to intragroup research? *Psychological Science* **8**(4): 261–67
- Sniezek J A, Buckley T 1995 Cueing and cognitive conflict in judge-advisor decision-making. *Organizational Behavior and Human Decision Processes* **62**(2): 159–74
- Sniezek J A, Henry R 1989 Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes* **43**(1): 1–28
- Sniezek J A, van Swol L 2001 Trust, confidence, and expertise in Judge Advisor Systems. *Organizational Behavior and Human Decision Processes*
- Stasser G, Titus W 1985 Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology* **48**: 1467–78
- Wegner D 1987 Transactive memory: A contemporary analysis of the group mind. In: Mullen B, Goethals G R (eds.) *Theories of Group Behavior*. Springer-Verlag, New York, pp. 185–208

J. A. Sniezek

Group Decision Making, Social Psychology of

People do many things in groups, of which one of the most significant is making decisions (see Baron et al. 1992, Brown 2000, Hogg and Tindale 2001, Levine and Moreland 1998, Moreland and Levine 1994). Indeed decision making is one of the most important functions of groups (see also *Group Processes, Social Psychology of*). Our lives are largely guided by decisions we make with others (e.g., groups of friends), or are made by groups of others (e.g., parliament), and many of us also spend a significant portion of our

working lives engaged in group decision making (e.g., committee meetings). Group decision making is also influenced in varying ways by the wider intergroup context in which the group exists—that is, by awareness of the existence of other groups (i.e. outgroups) with which one's own group may be in competition (see also *Intergroup Relations, Social Psychology of*). Finally, although group decision making usually involves people getting together to exchange information, it can also involve virtual groups making decisions via computer.

Social psychologists have long been interested in processes involved in group decision making, and whether groups make better or different decisions than do individuals—it is a field with significant applied relevance. There are many different ways to approach the study of group decision making (e.g., social communication approaches, social influence perspectives), there are many different facets of group decision making (e.g., persuasion, negotiation), and there are many different group decision making phenomena (e.g., juries, groupthink). Group decision making almost always involves psychological movement of group members from a relative diversity of positions to a more consensual position or to agreement on a group position. Group decision making reduces opinion differences within groups. Group decision-making can therefore be considered a social influence phenomenon (even a conformity phenomenon) that often also involves discussion, argument, persuasion, and leadership (see also *Group Processes, Social Psychology of*).

1. Social Combination Models

Social combination models relate the distribution of initial opinions in a decision making group to the final group decision (Stasser et al. 1989). These models, unlike social communication models, are not really concerned with interpersonal communication processes that transform individual positions into a group decision, but rather with the decision making rules that determine how group members combine their preferences (initial positions) into a group decision—the focus is on group preferences and their change over time. There is now an array of social combination models (Baron et al. 1992). Some are complex computer-simulation models, whereas others are more immediately related to real groups.

1.1 Social Decision Schemes

Social decision schemes are explicit or implicit decision-making rules that groups can adopt to decide how to combine information into a group decision (Davis 1973). Knowledge of the initial distribution of individual opinions in the group, and what rules the group is operating under allow reliable prediction of the final group decision. These rules include: (a)

unanimity—discussion pressurizes deviants to conform; (b) majority wins—discussion confirms the majority position, which is then adopted as the group position; (c) truth wins—discussion reveals the position that is demonstrably correct; (d) two-thirds majority—unless there is a two-thirds majority the group is unable to reach a decision; (e) first-shift—the group adopts a decision consistent with the direction of the first change in opinion by any member of the group. For intellectual tasks (where there is a demonstrably correct solution, as in, e.g., a mathematical puzzle) groups tend to adopt the truth wins rule, and for judgmental tasks (with no demonstrably correct solution, e.g., for aesthetic preference) the majority wins rule.

Rules vary in (a) strictness—i.e., the degree of agreement required by the rule (unanimity is very strict and majority less strict), and (b) distribution of power among members—i.e., authoritarian rules concentrate power in one member, whereas egalitarian rules spread power evenly. Generally, the stricter the rule the less the power concentration—unanimity is very strict but low in power concentration, whereas two-thirds majority is less strict but has greater power concentration. The type of rule adopted can, mainly due to its strictness, influence not only the group's decision but also members' preferences, their satisfaction with the group decision, the perception and nature of group discussion, and members' feelings for one another. For example, stricter decision rules can make the process of reaching final agreement exhaustive, slow, and difficult, but can enhance liking among members, and satisfaction with the quality of the decision.

1.2 Social Transition Schemes

Kerr's (Stasser et al. 1989) social transition scheme model focuses on the actual pattern of member positions that a group, operating under a particular decision rule, moves through *en route* to its final decision. To track opinion transitions, member opinions are monitored during the process of discussion, either by periodically asking the discussants, or by getting them to note any and every change in their opinion. These procedures can be intrusive, and therefore one issue concerns the extent to which they affect the natural ongoing process of discussion.

2. Jury Decision Making

Social combination models have often been tested with simulated juries, as the laboratory analogue has similar characteristics to an actual jury (e.g., *ad hoc* collection of people who initially do not know one another). Of course, juries are important decision making groups in their own right (Hastie et al. 1983), not only because they are held up as a symbol of all that is democratic, fair and just in a society, but also

because of the consequences of their decisions for defendants, victims, and the wider community (see also *Justice: Social Psychological Perspectives*). For instance, the 1992 Los Angeles riots were sparked by a 'not guilty' jury verdict, and the murder trial of O. J. Simpson in the mid-1990s became a national obsession in the USA. Juries are groups, and thus prey to all the consequences of group processes involved in group decision making.

Research on juries indicates that the initial majority position usually prevails in determining the final verdict. Close cases, which are associated with an even split of initial juror positions, tend to produce a hung jury. Hung juries, which are considered a relatively favorable outcome from the perspective of the defendant (the case is often dismissed), are less likely as jury size decreases from 12 to six persons, and if the decision rule changes from unanimity to some form of majority. Because jurors are instructed to presume the defendant innocent while reasonable doubt remains, juries have a leniency bias. This bias tends to surface (via acquittal) in close cases where there is an even split of opinions.

Juries are decision making groups that make a decision about a person. Thus, characteristics of the defendant and of the victim can affect the jury. Physically attractive defendants are more likely to be acquitted or to receive a lighter sentence, though biases can be reduced by furnishing sufficient factual evidence, presenting the jury with written rather than *in vivo* testimony, or explicitly directing the jury to consider the evidence alone. In the USA (and probably elsewhere) race can also affect the jury. Blacks are more likely to receive prison sentences, and people who murder a black are less likely than those who murder a white to receive the death penalty. Another issue is the influence of laws and penalties on the jury. Harsh laws with stiff penalties (e.g., the death penalty) tend to discourage juries from convicting. The jury foreman is important in guiding the jury to its verdict, because he/she occupies the role of leader. Research suggests that the foreman is likely to be someone of higher socioeconomic status, who has had previous experience as a juror, or simply initially occupies the seat at the head of the table.

3. Risky Shift and Group Polarization

Folk wisdom has it that people in groups make inherently more conservative decisions than they do on their own—individuals are likely to take risks, while group decision making is a tedious averaging process that errs towards caution. This is consistent with traditional social psychological perspectives on conformity and group influence processes—through mutual influence, people's attitudes tend to converge on the group mean (see also *Social Influence, Psychology of*, Turner 1991). However, under some circumstances groups can make more risky decisions than

individuals. Specifically, groups that already favor risk to some extent (i.e., the mean of the members' prediscussion opinions favors the risky pole of a decision making dimension) tend, through group discussion, to arrive at a group decision that is even more risky than the mean. This has been called risky shift. Later research showed that if the group initially favors caution then there is a cautious shift, and indeed that group decisions can be more extreme than the mean of the members' prediscussion opinions on a range of decision making dimensions that do not involve risk or caution, provided that the group initially tends towards one direction. Hence, the effect was renamed group polarization. It is defined as a tendency for groups, as a consequence of discussion, to make decisions that are more extreme than the mean of individual members' initial positions, in the direction already favored by that mean.

Since 1960, research has produced many different explanations of group polarization. They can be simplified to three major perspectives: persuasive arguments, social comparison/cultural values, and self-categorization.

(a) *Persuasive arguments theory*. Focuses on the persuasive impact of novel information. A group that leans in one direction will generate, during discussion, more novel arguments favoring that direction than the opposing direction, and, because novel arguments are persuasive, will cause members to become more committed to their original position and thus the group to shift more in that direction.

(b) *Social comparison/cultural values*. Focuses on people's motivation to avoid social censure and to seek social approval. Group discussion reveals the group's views, which are assumed by members to reflect what is socially desirable or culturally valued, and so group members publicly shift in the direction of the group in order to gain approval and avoid disapproval.

(c) *Self-categorization theory* (see also *Social Identity, Psychology of*). Argues that people in decision making groups psychologically identify with the group and use the group's position, which is cognitively represented as a prototype, to define themselves. The prototype is formed not only to capture similarities within the group but also to distance the group's position from outgroup or non-ingroup positions. Thus, extreme groups form polarized prototypes. The process of categorizing oneself as a group member (self-categorization) causes people to conform to this prototype and, thus, produces group polarization.

4. Groupthink

Groups sometimes employ defective decision-making procedures that produce poor decisions that can have disastrous consequences. Janis (Janis and Mann 1977) coined the term groupthink to define a style of thinking in cohesive groups in which the desire to reach

unanimous agreement overrides the motivation to adopt proper rational group decision-making procedures. The main cause of groupthink is excessive group cohesiveness, but there are other factors that relate to structural faults in the group (e.g., lack of impartial leadership) and to the immediate decision making context (e.g., high stress). Together, these factors generate a set of symptoms that are associated with defective decision making procedures. For example, there is inadequate and biased discussion and consideration of objectives and alternative solutions, and a failure to seek the advice of experts outside the group. A number of American foreign policy decisions with unfavorable outcomes (e.g., the 1961 Bay of Pigs fiasco, the unsuccessful defense of Pearl Harbor in 1941) have been attributed to groupthink in Presidential decision-making committees. Groupthink is probably a lot more common than one would like to think. It is a serious pitfall of group decision making, and groups need to be particularly vigilant in order to avoid it. A special decision making procedure, called the nominal group technique, is available to help groups avoid groupthink.

5. Group Remembering and Transactive Memory

Group decision making involves collective recall of information (see Clark and Stephenson 1995). Group remembering is not simply a retrieval activity in the pursuit of veridical recall, rather it is a constructive process where comparisons among group members shape the resultant group memory. Complex ideas can sometimes be transformed through group discussion into much simpler, and thus distorted, shared knowledge—a process called social representation, that is similar to rumor transmission. Group remembering may not be more accurate than individual remembering, but may be subject to implicit or even intentional bias.

Another feature of group remembering is transactive memory. Groups often need to bring a large amount of information to bear upon a decision, and thus they develop a differentiated memory structure where individuals or subgroups specialize in different memory domains. Where this occurs people not only have to remember their own part of the puzzle, but they also need to know who specializes in what memories—they develop a transactive memory structure. Research shows that group functioning and decision making is compromised where people have not acquired a transactive memory, or where such a system has not developed in the group.

6. Intergroup Decision Making

Decision making groups often contain factional rifts. In fact, many decision making groups can be considered interactive decision-making contexts where representatives from different groups are brought

together to reach a mutually satisfactory agreement (Brown 2000). In this sense, the group is actually a context for intergroup decision-making aimed at reducing intergroup conflict (see also *Intergroup Relations, Social Psychology of; Conflict and Conflict Resolution, Social Psychology of*). Examples of intergroup decision making include negotiations between nations, hostage crises, union/management confrontations, and parliament. The intergroup dimension introduces an additional dynamic to group decision making, because decision makers are charged with the responsibility to argue on behalf of their group, not themselves (Pruitt and Carnevale 1993).

When people are bargaining on behalf of groups to which they belong they tend to bargain more fiercely and uncompromisingly than if they were bargaining for themselves. The effect is enhanced when negotiators are aware that they are being observed by their constituents, either directly or through the media. This bullish strategy of intransigence is less likely to secure a satisfactory compromise than a more interpersonal orientation in which both parties make reciprocal concessions. Direct negotiation between group representatives is therefore likely to reach an impasse where neither group feels it can compromise without losing face.

To break the deadlock a third party can be recruited to mediate between the groups (Pruitt and Carnevale 1993). To be effective, mediators should have power, must be seen by both groups to be impartial, and the groups should already be fairly close in their positions. Biased mediators are ineffective because they are not trusted, and weak mediators are ineffective because there is little pressure for intransigent groups to be reasonable. Although mediators have no power to impose a settlement they can help in several important ways: (a) they can reduce the emotional heat associated with deadlock; (b) they can help reduce misperceptions, encourage understanding, and establish trust; (c) they can propose novel compromises that allow both groups to appear to win—i.e., change a zero-sum conflict (in which one group's gains are the other group's losses—the more one gains the more the other loses) into a nonzero-sum conflict (i.e., both groups can gain); (d) they can help both parties make a graceful retreat, without losing face, from untenable positions; (e) they can inhibit unreasonable claims and behaviors by threatening to publicly expose the group as being unreasonable; and (f) they can reduce intra-group conflict and thus help a group clarify its consensual position.

Where mediation fails, the last resort is arbitration, where a third party is invited to impose a mutually binding settlement. Arbitration can backfire because both groups adopt outrageous final positions in pursuit of a favorable compromise. One way to combat this is through 'final-offer arbitration,' where the third party chooses one of the final offers. This tends to encourage more reasonable final positions.

7. Motivation and Status

Group decision making is influenced by how hard members are prepared to work towards a group decision. Research on performance motivation suggests that people are often less motivated to work hard in groups than they are on their own—a phenomenon called social loafing (see also *Group Productivity, Social Psychology of* and Baron et al. 1992). A final aspect of group decision making is that some people have greater influence over the decision making process because they have higher status or occupy a leadership position. Research on status within groups shows that influence over group decisions and group action stems from possession of specific status characteristics, abilities and skills directly related to the group's function, and of diffuse status characteristics, membership of social categories that have high status in society at large (see Ridgeway 2001). Finally, leadership research suggests that effective leaders are people who can instill and inspire others with their own vision for the group, so called, transformational leaders who can transform a group so that its decisions reflect the leader's preferences (see also *Leadership, Psychology of*).

8. Future Directions

Due to its applied relevance the study of group decision making is always popular. However there are some key directions for future research. One such direction is the study of shared cognition—how groups create shared memories and representations and influence people to think in a collective manner. Another direction is the study of how intergroup relations impact decision making in groups. And finally, the study of computer-mediated decision making by virtual groups is a new area that is already attracting great attention.

See also: Decision Making (Naturalistic), Psychology of; Decision Making, Psychology of; Decision-making Systems: Personal and Collective; Group Decision Making, Cognitive Psychology of; Group Processes in Organizations; Group Processes in the Classroom; Group Processes, Social Psychology of; Group Productivity, Social Psychology of; Leadership, Psychology of; Organizational Decision Making

Bibliography

- Baron R S, Kerr N L, Miller N 1992 *Group Process, Group Decision, Group Action*. Brooks/Cole, Pacific Grove, CA
Brown R 2000 *Group Processes: Dynamics Within and Between Groups*, 2nd edn. Blackwell, Oxford, UK
Clark N K, Stephenson G M 1995 Social remembering: individual and collaborative memory for social information. *European Review of Social Psychology* 6: 127–60

- Davis J H 1973 Group decision and social interaction: A theory of social decision schemes. *Psychological Review* 80: 97–125
Hastie R, Penrod S D, Pennington N 1983 *Inside the Jury*. Harvard University Press, Cambridge, MA
Hogg M A, Tindale R S (eds.) 2001 *Blackwell Handbook of Social Psychology: Group Processes*. Blackwell, Oxford, UK
Janis I L, Mann L 1977 *Decision-Making: A Psychological Analysis of Conflict, Choice, and Commitment*. Free Press, New York
Levine J M, Moreland R M 1998 Small groups. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *The Handbook of Social Psychology*. McGraw-Hill, Boston, Vol. 2
Moreland R L, Levine J M 1994 *Understanding Small Groups*. Allyn and Bacon, Boston, MA
Pruitt D G, Carnevale P 1993 *Negotiating and Social Conflict*. Brooks/Cole, Monterey, CA
Ridgeway C L 2001 Social status and group structure. In: Hogg M A, Tindale R S (eds.) *Blackwell Handbook of Social Psychology: Group Processes*. Blackwell, Oxford, UK pp. 352–75
Stasser G, Kerr N L, Davis J H 1989 Influence processes and consensus models in decision-making groups. In: Paulus P B (ed.) *Psychology of Group Influence*, 2nd edn. L. Erlbaum, Hillsdale, NJ
Turner J C 1991 *Social Influence*. Brooks/Cole, Pacific Grove, CA

M. A. Hogg

Group Processes in Organizations

The success or failure of social and commercial organizations depends not only on specific individuals but also on the quality of the interaction and communication within groups (intragroup relationships) and between groups (intergroup relationships). Groups inside organizations influence individual performance, motivation, and work satisfaction as well as productivity, decision quality, learning, innovation, absenteeism, and fluctuation. Group processes are a central subject of social psychology as well as organizational psychology. During the 100-year history of group research, a substantial amount of literature has been amassed (for current reviews, see West 1996, Hinsz et al. 1997). In the following, we will concentrate on a selection of group research issues that we believe are of particular relevance to the understanding of group processes in organizations.

1. Definition and Structural Dimensions of Groups

Groups are defined as a number of people who are oriented toward the same or similar goals and who interact and communicate in order to achieve these goals. All groups share at least three structural principles: norms, cohesion, and role differentiation.

Norms. Norms can be defined as the shared expectations of the group members about how they should act or behave in certain situations. Norms make the behavior of group members more easily predictable and thus help to clarify social relationships.

Cohesion. When group membership is highly attractive—because goal achievement and need satisfaction are perceived to be achieved better in the group—group members are motivated to sustain the contact with the group. The sum of all the forces that increase group attractiveness is called ‘cohesion.’

Role and status. Role differentiation in groups is the social-psychological basis for work differentiation and organization. Role is defined as the expectation of the group members about the appropriate behavior in a certain position. Role-conformant behavior is regularly positively sanctioned, and deviations are sanctioned negatively. Status means that people differ in the hierarchy of prestige, power, influence, and privileges. High-status group members are allowed to deviate from group norms more strongly without being sanctioned than low status members (idiosyncrasy credit).

2. Conformity and Dissent

Conformity and dissent within groups are central aspects of group functioning. Behavior according to the group norm is necessary for relatively conflict-free interaction. Normally, group pressure brings the groups’ behavior and attitudes into line with the norms. This is important for group functioning and in order to achieve the relevant goals. However, normative behavior can also be dysfunctional, as has been shown by Asch’s studies (1951) in which group members expressed clearly wrong and irrational perceptions just because of the groups’ pressure towards conformity.

2.1 Groupthink: The Conformity of Groups

Groupthink describes the phenomenon by which existing information that is relevant to a certain decision is not processed by the group members in a rational and systematic way, but instead information processing is biased in the direction of a preferred decision alternative within the group. This reduces the quality of information processing within the group and increases the risk of faulty decisions. Analyzing historical political disasters following group decisions, such as, for example, Pearl Harbour, the American invasion of Cuba, or the Vietnam War, Janis (1982) has specified the symptoms of *groupthink*. These symptoms can be observed in the economical field too, such as, for example, in project teams, or on boards or committees. The risk of groupthink is amplified when groups are homogeneous (absence of minorities and

contradicting views), when the group structure is markedly hierarchical, and when there is a strong commitment to particular decision alternatives before group discussion begins.

The symptoms of groupthink include the following (cf. Janis 1982). (a) *Illusion of invulnerability*; that is, high conformity concerning the views of the group members and concerning their thinking styles makes them optimistic and overconfident. (b) *Willingness to take high risks* resulting in warnings, potential hazards, and other forms of negative feedback being ignored or not taken seriously. This behavioral tendency in groups serves the purpose of avoiding questioning one’s own basic beliefs and value systems when facing a challenging decision. (c) *Stereotyping of members with a different opinion* by labeling them as incompetent. (d) *Pressure towards conformity* exerted by the formal leader or the majority of the group in order to stabilize the group’s homogeneity. Conformity pressure helps members to neglect counter arguments more easily and supports self-censorship processes which makes group members (falsely) assume that silence means approval.

There is considerable empirical evidence regarding prevention and intervention strategies for counteracting groupthink. Some examples are (for a summary of these interventions, see Frey et al. 1996):

- (a) the group leader acts as an impartial coordinator,
- (b) systematically strengthening self-confident minority members,
- (c) bringing external experts into the group,
- (d) implementing artificial conflict in the group decision process (e.g., by devil’s advocacy or dialectical inquiry techniques),
- (e) holding each group member individually responsible for the decision process and the outcome,
- (f) creating different subgroups separately engaging in decision making.

2.2 Minority Influence

The problem of conformity and homogeneity in groups is that it may come at the expense of divergent thinking, which can be helpful when scrutinizing premature conclusions, and innovative thinking. Due to conformity pressure, the default information processing style in groups is consent oriented. Minority influence, however, often generates productive dissent.

According to Moscovici (1976), minority members are able to oppose the pressure towards conformity in groups under certain conditions—and are thereby able to influence the attitudes and behavior of the majority. To achieve this, minority members should present their point of view in a consistent behavioral style, though not rigidly (flexibility of negotiation and argumentation), while signaling a willingness to compromise.

Why can minorities exert an influence on majority members? According to Moscovici's Conversion Theory (cf. Moscovici 1980), minority members induce an internal cognitive conflict within members of the majority; majority members then ask themselves *What is in fact the right course?* This leads to an intensification of information processing, and is accompanied by divergent and more creative thinking styles, resulting in a less biased information search and greater openness for the minority position. In contrast to this, majorities exert a pressure to conform, producing social conflict between people or subgroups of people within the group rather than a cognitive conflict. Group members ask themselves *Which person is right now?* This leads to a more superficial processing of information, and is accompanied by conventional and conformist styles of thinking.

What is the relevance of the minority research for group processes in organizations? Very often minority groups or minority persons have the feeling that they *cannot* change a big organization. Taking minority research into account, the opposite is in fact true. Minority influence in organizations is a precondition for scrutinizing common patterns of thought and practice, and thus is an important ingredient for organizational learning, change, and innovation. The interesting thing is that a minority can be influential independently of the hierarchical structure.

3. Bringing Together Individual Contributions

How can individual contributions be brought together in an optimum way so that production losses (e.g., loss of motivation and coordination loss) can be minimized, and procedural gains or synergies (e.g., motivating each other, learning from each other, and sharing innovations) are maximized?

3.1 Activating or Impeding Individual Performance

Whether social inhibition or social facilitation occurs via the mere presence of other persons is a question of the complexity and newness of the task to be performed. If other persons are present, difficult or new tasks are carried out less satisfactorily, since these tasks will lead to uncertainty and fear of negative feedback ('social inhibition'). Simple or well-trained tasks, on the other hand, are usually carried out better in the presence of other people, as the active subject can expect success and a positive judgment from his/her colleagues ('social facilitation').

In the case of social inhibition, no matter whether the task performed is difficult or not, loss of motivation is involved. For example, if the individual contributions are not perceptible in the groups' overall output, or their importance for group productivity is not clearly seen, social loafing occurs. When group

members expect other people in their group to be 'free-riders,' they may be inclined to conclude that their own individual contribution might be too high compared with the contributions of other group members. Consequently, individual performance will be decreased since no one wants to be left to carry the can; this is called the 'sucker effect.' Additionally, if group members fear negative comments from others, or if they are not sure about the norms and standards of evaluation as a whole, they will also withhold optimal performance (for fear of a negative evaluation). These motivational impediments on optimum individual performance can be met by transparency of performance feedback, by conveying concrete individual as well as collective performance norms, by fostering intrinsic motivation (e.g., interesting and challenging tasks), by strengthening group cohesion in combination with a high achievement norm, and by pointing out how important each and every individual contribution is for the group's productivity (for an overview see Sheppard 1993).

On the other hand, when group cohesion is high and the group's task is highly important, one can observe that groups will increase individual performance via motivation gain. If the group members' capabilities do not differ very much, and if the success of the group as a whole is very important for all the members, very strong and competent persons often try to compensate for the weaknesses of their colleagues by trying even harder to do their best. The weaker group members then likewise try to increase their output, as they do not want to be the 'losers' bringing the whole team down. These gains in motivation are especially likely if group cohesion is high—that is in case of a very attractive group (cf. Brown 1992), in case of a high mutual sympathy (cf. Latané 1986), and if the group's task is highly important (cf. Williams and Karau 1991).

3.2 Task Structure and Social Combination of Individual Resources

According to a very well-known model of group performance in social psychology (i.e., Steiner 1972), the *de facto* group performance is a function of the potential group performance, minus procedural losses, plus procedural gains. In order to find out whether groups are working to their full potential, one needs to determine the potential level of group performance. For this, one needs to know the individual capabilities of the group members and the type of task at hand.

In general, the structure of work tasks may be *additive*, in which case the potential group performance is the sum of all the individual performances, for example in the case of brainstorming. If the task is structured in a *compensatory* way, then the potential group performance is the result of the mean of all individual performances, for example in the case of

guessing tasks or majority decisions). If work tasks are *disjunctive*, the group potential is the result of the best individual performance (for example in the case of problem-solving tasks or decision problems). Thus, determining a group's potential level of performance requires consideration of the group task. In order to further determine whether and to what extent a group is working below (or sometimes even above) its potential, the actual group performance is obtained and compared with the potential level of performance. The differences tell us to what extent group-induced procedural losses and procedural gains have occurred.

3.3 Synergy in Groups

Practitioners very often speak about 'group synergy.' This means that a group is more than the sum of all its parts, or more precisely: that a group generally does better than its aptest single member. However, synergy effects have only rarely been demonstrated in social-psychological group research (for an overview, see Hill 1982, Brown 1993). Stroebe et al. (1992) have argued that the common belief in group synergy may rest on an overestimation of one's own output in group contexts. For example, where brainstorming is concerned, single individuals usually believe they are better and more productive when they work in groups—probably because they feel 'enriched' by the ideas of the others and because many new perspectives emerge. However, Stroebe et al. (1992) have shown that individual group members are inclined to claim other group members' ideas as their own body of thought (and, hence, believe in 'having produced a lot'). Consequently, one feels being 'more productive' as a participant of a brainstorming group compared with working alone. Each group member extrapolates this impression onto the other persons, and in the end believes that the group is much more efficient than the sum of its components (i.e., members) working alone. This overestimation feeds the idea of 'group synergy.'

3.4 Top Performance of Work Teams

There is growing evidence that top performances in groups occur despite dominant forces such as conformity, groupthink, loss of motivation, and coordination loss (Levine and Moreland 1990, West 1996). A high decision quality and top performance of teams is likely under the following conditions.

(a) Homogeneity of values and rules exists, meaning that team members agree on rules for dealing with each other; they are able to communicate honestly and to respect each other, and show loyalty.

(b) A heterogeneity of talents, experiences, education, and backgrounds exists within the team.

(c) The team has a vivid commitment to excellent output and an ethos of achievement. All members take responsibility for team success.

(d) Team members fit together technically and personally.

(e) Team members are able to profit from their individual strengths.

(f) Orientation towards the best practice idea (benchmarking). This technique means comparing oneself with best practice in one's own branch or with the best in the world, but also with the best within the organization.

The team is able to use the tool of team reflection, meaning regular joint reflection on social and task issues. What is good and should be sustained, what is bad and should be improved (West 1996).

4. Group Dynamics

Group interaction may be viewed from three different theoretical perspectives. First, one can look at socio-emotional relations where cohesion, role-forming, open communication, the building of trust, and linear models concerning the development of groups are relevant (forming, storming, norming, performing; Tuckman and Jensen 1977). Second, one can look at the development of goal-directed and output-oriented groups, mainly to be found in work groups in organizations where goal finding, role differentiation, leadership, quality of decisions, coping with problems, and action-oriented models of group dynamics (cf. Pinto and Prescott 1988) are relevant. Third, one can look at the development of interpersonal and task-related structures within and between groups, in organizations where high-quality work, customer orientation, and innovation are sought to be improved. Scientific progress has been made in all these fields (cf. Brodbeck 1996, Hinsz et al. 1997). In the following we will discuss two subjects that are particularly related to the development of goal and task-oriented group dynamics: group learning and approaches to team development in practice.

4.1 Group Learning

Group learning in general has not been explored very well to date (Argote and Epple 1990). However, transactive knowledge systems have been researched quite thoroughly; in other words the way in which a group chooses to encode, store, and retrieve information systematically. Transactive knowledge systems make information accessible for team members, even if this information has been (physically) stored by other members of the group (Wegner 1987). According to Liang et al. (1995), teams that have been trained to do complex tasks together (i.e., who have been trained as a team) do better afterwards than groups where the members have been trained individually to solve the task. The authors hold that this lead in the output is due to the fact that teams who have been trained as a group have established transactive knowledge systems. These structures improve the group's performance, for

example by means of group processes such as feedback about weaknesses, helping each other out, sharing information that others need right now. A model of group learning as a multilevel phenomenon; that is, on the individual level, the group level, and respective interactions between these two levels; has been presented by Brodbeck and Greitemeyer (2000a, 2000b).

4.2 Team Development

Typical techniques of team development usually deal with the analysis and layout of group processes in order to increase work performance, personality development, team climate, and cooperation between departments. Four team development approaches can be distinguished. (a) The interpersonal approach tries to increase competence as well as cooperation by promoting self-reflection within groups. (b) The goal-oriented approach tries to increase motivation, performance, and subjective well-being by optimizing goal finding and goal negotiation, for example by way of participation, mutual feedback, and intelligent reward systems. (c) The role approach tries to reduce misunderstandings and conflicts (due to poor coordination or bad communication): underlying conflicts are to be made visible and a better role understanding achieved using techniques of role building and role reconstruction. (d) The process-oriented approach aims to obtain an increase in performance as well as achievement by restructuring collective work processes (such as problem solving and decision making) by way of well-established methods (e.g., Delphi Technique), to some extent assisted by computer technology (for example, group decision support systems (GDSS)).

Practitioners often combine these approaches. For example, it is common to combine the negotiation of new goals with role negotiation so that a corresponding new understanding of expectations and norms can be established. (For concrete techniques and measures of team development that can be utilized in this process, see West 1997.)

5. Social Identity: In-group–Out-group Processes

Until now we have discussed group processes and their consequences within groups. Of course, in organizations there are also group processes between groups. Here the most important theory which is relevant for these processes is Tajfel's social identity theory (SIT). Interestingly, it is a psychological theory about the social cognition of individuals (not a sociological one that directly addresses entire collectives such as organizations). The predictions made by SIT are among the best proven in social psychology.

Individuals identify with the groups they belong to and derogate others who do not belong to their group

(the in-group). Consequently, there is a high degree of identification with the in-group and the respective organization, but at the cost of respect for and cooperation with individuals perceived to belong to other groups (out-groups). Tajfel's theory is very important for understanding the in-/out-group processes underlying conflict and conflict resolution.

Social identity theory emphasizes that all individuals are motivated to shape a positive social identity. The underlying processes are: (a) categorization; that is, a basic cognitive mechanism for structuring the manifold stimuli from the environment in an orderly fashion—in particular, within their social environment, humans categorize on the basis of who belongs to the in-group and who does not (in-group/out-group distinction). (b) Social comparison; that is, a basic sociocognitive mechanism for gaining information about where one stands within the social context—in particular, humans compare their own in-group with other out-groups, preferably in such a way that the dimension on which the comparison is performed produces positive distinctiveness (i.e., a feeling of positive uniqueness). Having the choice of a dimension is important in that when the in-group compares badly (e.g., with a highly performing group), it can be argued that it is more important to have character than to have strong skills and success, thereby protecting the positive distinctiveness of the in-group. Thus, in summary, SIT describes the dynamics between groups on the basis of the individual's motivation to achieve social identity and positive distinctiveness—at the cost of derogating others because they are perceived to belong to an out-group.

Tajfel (1978) shows that even a minimal group context (minimal group paradigm) which may come about through a more or less arbitrary distinction between two groups (e.g., admirers of Vasily Kandinsky vs. admirers of Paul Klee), suffices to produce the in-group favoritism and out-group discrimination processes. People from the out-group are discriminated against more—for example concerning the evaluation of positive attributes, concerning judgments about liking, and concerning the distribution of material rewards (compared to persons from the 'in-group').

These processes are highly relevant for groups within organizations because very often subgroups within organizations think in terms of in-groups and out-groups instead of seeing the overall organization as an in-group. An organization is always in a conflict, on the one hand to increase the in-group feeling because a strong feeling of togetherness is positive, on the other hand the in-group feeling must not be too strong because this could mean shutting down borders so that there are not that many contacts between groups and departments. Therefore the question is what can be done in order to increase both. According to Tajfel's SIT, it is (a) good to have a common goal which can only be achieved by joint action and

cooperation of all groups, (b) important that groups have background information about what the tasks of the other groups are, (c) good to have a rotation of positions between heterogeneous groups, and (d) good to have a lot of informal activities and personal relationships between members of different groups. The important point is that a group must be open and have open borders.

6. Group Processes in Mergers, Acquisitions, and Change Management

When organizations are merged or taken over, or when organizational changes take place, group processes as described in SIT play a central role. In mergers and acquisitions, it is important that the norms and values of the distinct groups can be combined and are not too different from the start. Very often, these groups behave towards each other like an in-group towards an out-group. Very often with mergers or acquisitions, two actual competitors which merge as an in-group or out-group come together and have to continue as one group. Very often, an earlier empty competitor has to be integrated. Many intergroup conflicts can be reduced by taking into account certain psychological findings about the acceptance of change management processes, which are listed below.

(a) There is a cultural fit between the two merging companies and therefore an identification with the new emerging culture is possible (principle of identification and social identity).

(b) The change process is perceived to be definite, inevitable, and irreversible (principle of inevitability).

(c) The vision is clearly perceived and shared, and one understands the reasons why change is necessary (principle of explainability and sense).

(d) Participants have the impression that they can foresee and predict the processes to come (principle of predictability).

(e) There is a consensus about rules and responsibilities (principle of clarity).

(f) Participants feel they are an active part of the decision process (principle of participation).

(g) They feel they are being treated fairly; for example, there is perceived fairness about resources and positions (procedural fairness).

(h) Top managers are perceived as being trustworthy and are positive role models (principle of truthfulness).

Very often, groups and people who are responsible for change processes in organizations, who have to make the relevant decisions and who are responsible for the implementation of those decisions, have a poor understanding of how human beings and groups function in psychological terms. This is one reason why so many mergers, acquisitions, and change processes fail. Very often, the culture and the mentality

of groups do not fit with each other; when the above-mentioned aspects are neglected in the change process, it will fail.

It is the inability to judge employees' and groups' anxieties, hopes, and expectations correctly, and to communicate why and how and what changes are necessary (Hogan and Overmyer-Day 1994). Taking the above-mentioned conditions into account will help to make change processes more successful in the future.

7. Future Directions

Group processes in organizations are a typical example of good interdisciplinary cooperation between social psychologists and organizational psychologists. This research shows that nothing is as practical as a good theory. Of course, there are no integrating theories of groups but as we have shown there are different small group theories which can be applied for different heterogeneous phenomena. In the future, it is important to have an integration of these small group theories and in the empirical field it is also necessary to do more longitudinal studies within organizations.

See also: Group Decision Making, Social Psychology of; Group Processes, Social Psychology of; Group Productivity, Social Psychology of; Intergroup Relations, Social Psychology of; Problem Solving (Joint), Psychology of; Psychological Climate in the Work Setting; Psychology: Overview; Situated Learning: Out of School and in the Classroom; Social Psychology; Social Psychology, Theories of

Bibliography

- Argote L, Eppler D 1990 Learning curves in manufacturing. *Science* **247**: 920–3
- Asch S E 1951 Effects of group pressure on the modification and distortion of judgements. In: Guetzkow H (ed.) *Groups, Leadership, and Men*. Cernegie, Pittsburgh, PA, pp. 177–90
- Brodbeck F C 1996 Criteria for the study of work group functioning. In: West M (ed.) *Handbook of Workgroup Psychology*. Wiley and Sons, Chichester, UK, pp. 285–315
- Brodbeck F C, Greitemeyer T 2000a A dynamic model of group performance: Considering the group members' capacity to learn. *Group Processes and Intergroup Relations* **3**: 159–82
- Brodbeck F C, Greitemeyer T 2000b Effects of individual versus mixed individual and group experience in rule induction on group member learning and group performance. *Journal of Experimental Social Psychology* **36**: 621–48
- Brown R 1993 *Group Processes: Dynamics Within and Between Groups*, 4th edn. Blackwell, Oxford, UK
- Frey D, Schulz-Hardt S, Stahlberg D 1996 Information seeking among individuals and groups and possible consequences for decision making in business and politics. In: Witte E, Davis J H (eds.) *Understanding Group Behavior—Small Group Processes and Interpersonal Relations*. Lawrence Erlbaum, Mahwah, NJ, pp. 211–25

- Hill G W 1982 Group versus individual performance: Are two heads better than one? *Psychological Bulletin* **91**: 517–39
- Hinsz V B, Tindale R S, Vollrath D A 1997 The emerging conceptualization of groups as information processors. *Psychological Bulletin* **121**: 43–64
- Hogan E A, Overmyer-Day L 1994 The psychology of mergers and acquisitions. In: Cooper C L, Robertson I T (eds.) *International Review of Industrial and Organizational Psychology*. John Wiley and Sons, Chichester, UK, Vol. 9, pp. 247–81
- Janis I L 1982 *Victims of Groupthink, a Psychological Study of Foreign-Policy Decisions and Fiascoes*, 2nd edn. Houghton Mifflin Company, Boston, Atlanta, Dallas
- Latané B 1986 Responsibility and effort in organizations. In: Goodman P S and Associates (eds.) *Designing Effective Work Groups*. Jossey-Bass, San Francisco, pp. 277–304
- Levine J M, Moreland R L 1990 Progress in small group research. *Annual Review of Psychology* **41**: 585–634
- Liang D W, Moreland R, Argote L 1995 Group versus individual training and group performance: The mediating role of transactive memory. *Personality and Social Psychology Bulletin* **21**: 384–93
- Moscovici S 1976 *Social Influence and Social Change*. Academic Press, London
- Moscovici S 1980 Towards a theory of conversion behavior. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*. Academic Press, New York, Bd. 13, S. pp. 208–39
- Pinto J K, Prescott J E 1988 Variations in critical success factors over the stages in the project life cycle. *Journal of Management* **14**: 5–18
- Sheppard J A 1993 Productivity loss in performance groups: A motivation analysis. *Psychological Bulletin* **113**: 67–81
- Steiner I D 1972 *Group Process and Productivity*. Academic Press, New York
- Stroebe W, Diehl M, Abakoumkin G 1992 The illusion of group effectivity. *Personality and Social Psychology Bulletin* **18**: 643–50
- Tajfel H 1978 *Differentiation Between Social Groups: Studies in the Social Psychology of Intergroup Relations*. Academic Press, London
- Tuckman B, Jensen M 1977 Stages of small-group development. *Group and Organizational Studies* **2**: 419–27
- Wegner D M 1987 Transactive memory: A contemporary analysis of the group mind. In: Mullen B, Goethals G R (eds.) *Theories of Group Behavior*. Springer, New York, pp. 185–208
- West M 1994 *Effective Teamwork*. BPS Books, Leicester, UK
- West M 1996 *Handbook of Work Group Psychology*. Wiley and Sons, Chichester, UK
- West M 1997 *Developing Creativity in Organizations*. BPS Books, Leicester, UK
- Williams K, Karau S J 1991 Social loafing and social compensations: The effects of expectations of co-worker performance. *Journal of Personality and Social Psychology* **61**: 570–81

D. Frey and F. Brodbeck

Group Processes in the Classroom

Interest in group learning and group processes has a very long history and remains a key focus of current efforts to improve education. Group processes in the classroom occur when peers engage in a common task and can include a variety of both social and cognitive

processes. The focus of this entry is on cooperative or collaborative groups of peers and the various theoretical approaches that attempt to explain group processes in these kinds of group. These approaches propose varying mechanisms for the effectiveness of group learning. Comparisons of empirical findings across studies conducted from different theoretical perspectives are very difficult because of the wide variances in tasks used, duration of studies, roles of teachers and students, outcome measures, and other variables such as group composition. One consequence of the variance in theoretical approaches is the fact that these approaches require different decisions with regard to key variables such as group size or group composition. Group processes can best be understood within particular theoretical approaches. This entry will not address the influences of specific variables but will instead focus on the delineation of various theoretical perspectives and the linkage of research and practice.

1. Theoretical Approaches to Group Processes

Different theoretical approaches to peer learning propose different underlying mechanisms or processes that come into play when a group works together (O'Donnell and O'Kelly 1994). Slavin (1996) described two main classes of orientation toward peer learning: a social-behavioral orientation that includes motivational and social cohesion approaches, and a cognitive orientation that includes elaboration and developmental perspectives. There are two distinct developmental theories. Table 1 delineates key differences among these perspectives as they relate to instructional decisions and problems that may arise from such decisions.

1.1 Social-behavioral Perspectives

The concept of interdependence among group members is fundamental to the motivational and social cohesion perspectives on cooperative learning, although they differ in how interdependence is created and sustained. When group members are interdependent, the success of one member can occur only if other members are also successful. In competitive situations, one person's success requires that another person fail. In the social motivational approach, interdependence is created by the use of group rewards. In Student Teams Achievement Divisions (Slavin 1995), students work in small groups to master material that the teacher has presented. Each student takes an individual quiz on the material, and points are assigned to the group based on each individual's improvement score. Thus, students are accountable for improving over past performance and learners of varying abilities can contribute equally to the total group score. Group rewards are assigned based on the team's number of points. A key component of this

Table 1

Key considerations in cooperative learning from a variety of theoretical perspectives

Considerations	Theoretical perspectives				
	Social-behavioral			Cognitive	
	Motivation	Social cohesion	Elaboration	Developmental	
				Piagetian	Vygotskian
Goals/incentives	Rewards essential	Rewards nonessential	Rewards nonessential	Rewards nonessential	Rewards nonessential
Group size	Large (4–6)	Large (4–6)	Small (2–4)	Small	Dyads
Group composition	Heterogeneous	Heterogeneous	Heterogeneous	Homogeneous	Heterogeneous
Tasks	Rehearsal	Rehearsal/integrative	Rehearsal/integrative	Exploratory	Skills
Teacher role	Director	Facilitator	Facilitator	Facilitator	Model/guide
Potential problems	Use of reward	Poor social skills	Poor help-giving	Inactive	Poor help-giving
Averting problems	Group size	Social loafing	Unequal	No cognitive	Providing adequate
	Improvement scores	Cognitive loafing	participation	conflict	time/dialog
	Simpler tasks	Team building	Direct instruction in	Structuring	Direct instruction in
	Social cohesion	Conflict resolution strategies	help-giving	controversy	help-giving
		Discuss group process	Model help-giving		Modeling help-giving
			Scripting interaction		

Source: O'Donnell and O'Kelly 1994, p. 327, © Kluwer Academic/Plenum Publishers. Reproduced with permission

approach is the reliance on individual accountability. Evidence of the effectiveness of this approach comes from a large number of studies (Slavin 1996). The kinds of learning task used in this approach, however, tend to focus on basic skills or factual knowledge, and it is unclear whether this approach is effective in promoting higher-order skills.

1.2 Social Cohesion Perspectives

In the social cohesion approach, interdependence derives from a motive of care and concern. Group members help one another because they are concerned about their peers. Teaching social skills that will support this type of interdependence is an important emphasis in social cohesion approaches. The Johnsons's Learning Together technique (Johnson and Johnson 1994) illustrates the social cohesion approach. The empirical evidence supporting this orientation toward group learning as a means of enhancing academic achievement is much weaker than that supporting the motivational perspective. Slavin (1996) notes that cooperative learning methods that emphasize team building and group process, but do not provide group rewards, are not very effective but that when they are combined with the use of rewards or accountability, student achievement from group learning exceeds individualistic learning. Slavin concludes that individual accountability and group rewards are critical to the success of cooperative learning methods.

Not everyone subscribes to the idea that rewards are necessary to create cohesion or interdependence. Cohen (1994) believes that cooperative learning that uses group rewards may be useful for lower-level skills but may not be either necessary or helpful for promoting higher-order skills. Thus, decisions about whether to adopt a motivational or social cohesion approach to peer learning may depend on the desired outcomes from group interaction.

Cohen (1994) emphasizes a different source of social cohesion, suggesting that the interest value of the task in which students engage will create the kind of interdependence necessary for students to succeed. In such situations, difficulties associated with status differences among students are ameliorated. When students have low status in the group, they participate less and in ways that do not promote achievement. Explicit rewards in this context may only exacerbate the problems created by differential status in a group. Cohen and her colleagues have successfully promoted social cohesion in groups by reducing status differences among group members so that they are oriented to accept that all students *can* contribute to task accomplishment. Complex and interesting tasks are necessary for this level of inclusion to occur. Cohen and her colleagues have been very successful in promoting student achievement. Results from Cohen and her colleagues suggest that the nature of the task

may have important implications for the kind of accountability that is needed to promote effective group learning.

1.3 Cognitive/Elaborative Perspectives

The third perspective identified by Slavin is a cognitive/elaborative approach and is based on information processing theory. The role of the group interaction from this perspective is to increase active processing by providing opportunities for restructuring and elaborating knowledge. Group learning techniques influenced by this perspective tend to provide some structure to the interaction of students by emphasizing the cognitive processes in which students must engage. Examples of work influenced by this perspective include the use of scripted cooperation by O'Donnell, Dansereau, and colleagues (O'Donnell and Dansereau 1992), reciprocal peer questioning (King 1999), reciprocal teaching (Palinscar and Brown 1984), and the work of Noreen Webb (Webb 1992). The empirical evidence supporting the use of these techniques is strong.

The quality of discourse is critical from this perspective and will have an influence on the outcomes from group interaction. Webb (1992, Webb and Palinscar 1996) has shown that the quality of explanations that are provided in a group are linked to achievement. Giving explanations is more strongly associated with achievement than receiving explanations. King (1999) has also shown that instructional efforts to improve the quality of discourse by teaching students to ask questions that facilitate knowledge construction is very effective.

1.4 Developmental Perspectives

A fourth approach to understanding cooperative groups relies on two different theories of development. From a Piagetian perspective (De Lisi and Golbeck 1999), peers can learn from one another because they can provide opportunities for cognitive conflict and resolution. Groups that experience mutuality of influence and power are most likely to provide a context within which cognitive development can occur because they are more likely to include differences of viewpoint. Children who have not attained the principle of conservation often develop and maintain conservation concepts when they have worked with a child who has already developed this concept (De Lisi and Golbeck 1999). Much of the work on conceptual change in science education has been influenced by this perspective.

From a Vygotskian perspective (Hogan and Tudge 1999), children acquire knowledge and skills that are first modeled in the community and are subsequently internalized by an individual. When cooperative learning is anchored in Vygotskian theory, it is

concerned with the larger community of practice in which individual learning is accomplished. Key processes include the scaffolding of the learner's performance by a more expert individual. The difference between what an individual can accomplish alone and with the assistance of an adult guide or more capable peer is defined as the zone of proximal development.

1.5 Sociocultural and Combined Theoretical Perspectives

A fifth perspective that undergirds such group learning techniques such as Communities of Learners (Brown and Campione 1990) may be identified. It includes the developmental perspective of Vygotsky, with its emphasis on the sociocultural basis of learning, and a social cohesion approach in which such learning is motivated by care and concern. Other forms of cooperative learning borrow from multiple perspectives in making choices about the selection of tasks and the organization of learning (Van Meter and Stevens 2000).

2. Research and Practice

Despite the burgeoning research findings that might assist teachers in making choices about grouping for instructional purposes, teachers do not seem to rely on the empirical research for guidance when making such choices (Antil et al. 1998). However, few studies have examined teachers' beliefs about group processes in the classroom.

2.1 Teachers' Use of Cooperative Learning

The research findings from any of the perspectives previously described seem to have little systematic influence on teachers' choices about groups. Although Slavin (1996) is convinced that empirical evidence supports the use of group rewards, he notes that there is quite a reluctance on the part of classroom teachers to use them.

Antil and his colleagues (1998) conducted a study of elementary school teachers' conceptualization and utilization of cooperative learning. Eighty-five teachers responded to an initial survey, and 93 percent of them reported that they used cooperative learning for both academic and social reasons. In interviews with 21 of these teachers, Antil found that the majority structured tasks to create positive interdependence, and taught students skills for working in groups. Few, however, employed recognized forms of cooperative learning and, in particular, did not link individual accountability to group goals. No data were presented, however, that linked classroom practices in relation to cooperative learning and student achievement. It is clear from this study that teachers do not implement the particular practices of cooperative learning that

research has shown to be effective. However, when teachers receive extensive professional development, the effects on student achievement can be powerful (Schultz 2000).

2.2 Usable Models of Group Processes

The availability of multiple perspectives on peer learning and the proliferation of specific techniques may contribute to the disconnection between the available research findings and their use by classroom teachers. The kind of classification system represented in Table 1 is rarely presented to either pre-service or in-service teachers. Usable models of peer learning that can guide practical decisions are needed. Such models need clearly to link expected outcomes to classroom processes and teachers' actions or decisions.

Slavin (1996) proposed a model of effective peer learning with motivation at its core. Motivation is necessary in order for students to engage in the active processing (e.g., explanations) that is linked to achievement. The keys to motivating students are group rewards and individual accountability. Teachers can act to create and sustain children's motivation that will allow them to link will and skill in pursuit of achievement.

Van Meter and Stevens (2000) emphasize the quality of discourse in a group as the key variable in effective group learning. Other instructional choices should be such that the quality of discourse is enhanced. Thus, task instructions and group composition are only interesting insofar as they influence the quality of the discourse among peers.

These two models provide a simple set of goals for teachers using group learning. Coherence among classroom decisions about fundamental aspects of arranging instruction (e.g., group size, group composition, type of task) is more likely to occur if there is a clear goal in place (e.g., motivate students to engage in active learning). Teachers in the Antil et al. (1998) study reported that they did not subscribe to particular cooperative techniques because they wanted to maintain their decision-making autonomy. The two models that are briefly described here with their emphases on motivation and discourse provide a set of superordinate goals that can accommodate teachers' needs for autonomy, while preserving a principled approach to decision making about group processes.

3. Future Directions

Given that teachers seem reluctant to use group rewards (Slavin 1996), it is important to specify the conditions under which alternative methods of creating interdependence among group members can be effective. What are the kinds of tasks for which extrinsic reward are not necessary? Answers to this question lie in the analysis of the kinds of process in which students engage when they are engaged in

interesting and challenging tasks. A clearer understanding of effective processing of complex tasks may make it possible to find alternatives to rewards.

The tension between theories of group processes and classroom practices of group learning presents the richest source of future research questions. Teachers' beliefs about learning lie at the fulcrum of their abilities to implement effective group learning in the classroom. The relationships between teacher beliefs, implementation practices, and student achievement remain untouched as research questions. Research on this set of relationships would provide valuable insights that could inform teacher preparation and in-service professional development, and also improve classroom practice of group learning.

See also: Cooperation and Competition, Psychology of; Cooperative Learning in Schools; Group Processes in Organizations; Group Processes, Social Psychology of; Groups, Sociology of; Hidden Curriculum in the Classroom; Open Classroom: Learning and Teaching

Bibliography

- Antil L R, Jenkins J R, Wayne S K, Vadasy P F 1998 Cooperative learning: Prevalence, conceptualizations, and the relationships between research and practice. *American Educational Research Journal* **3**: 419–55
- Brown A L, Campione J C 1990 Communities of learning and thinking, or a context by any other name. *Developmental Perspectives on Teaching and Learning Thinking Skills* **21**: 108–26
- Cohen E G 1994 Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research* **64**: 1–35
- De Lisi R, Golbeck S 1999 Implications of Piagetian theory for peer learning. In: O'Donnell A M, King A (eds.) *Cognitive Perspectives on Peer Learning*. Lawrence Erlbaum, Mahwah, NJ, pp. 3–37
- Hogan D, Tudge J H R 1999 Implications for Vygotsky's theory for peer learning. In: O'Donnell A M, King A (eds.) *Cognitive Perspectives on Peer Learning*. Lawrence Erlbaum, Mahwah, NJ, pp. 39–65
- Johnson D W, Johnson R 1994 *Learning Together and Alone: Cooperative, Competitive, and Individualistic Learning*, 4th edn, Allyn and Bacon, Boston
- King A 1999 Discourse patterns for mediating peer learning. In: O'Donnell A M, King A (eds.) *Cognitive Perspectives on Peer Learning*. Lawrence Erlbaum, Mahwah, NJ, pp. 87–115
- O'Donnell A M, Dansereau D F 1992 Scripted cooperation in student dyads: A method for analyzing and enhancing academic learning and performance. In: Lazarowitz R, Miller N (eds.) *Interaction in Cooperative Groups: The Theoretical Anatomy of Group Learning*. Cambridge University Press, New York, pp. 120–41
- O'Donnell A M, O'Kelly J 1994 Learning from peers: Beyond the rhetoric of positive results. *Educational Psychology Review* **6**: 321–49
- Palinscar A S, Brown A 1984 Reciprocal teaching of comprehension monitoring activities. *Cognition and Instruction* **2**: 117–75

Schultz S E 2000 *Are Four Heads Better than One? Effects of Group Work on Students' Achievement in Science*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA

Slavin R E 1995 *Cooperative Learning: Theory, Research, and Practice*, 2nd edn. Allyn and Bacon, Boston

Slavin R E 1996 Research on cooperative learning and achievement: What we know, what we need to know. *Contemporary Educational Psychology* **21**: 43–69

Van Meter P, Stevens R J 2000 The role of theory in the study of peer collaboration. *Journal of Experimental Education* **69**: 113–27

Webb N M 1992 Testing a theoretical model of student interaction and learning in small groups. In: Lazarowitz R, Miller N (eds.) *Interaction in Cooperative Groups: The Theoretical Anatomy of Group Learning*. Cambridge University Press, New York, pp. 102–19

Webb N M, Palinscar A S 1996 Group processes in the classroom. In: Berliner D, Calfee R C (eds.) *Handbook of Educational Psychology*. Macmillan, New York

A. M. O'Donnell

Group Processes, Social Psychology of

Group processes encompass such a wide range of phenomena that the social psychological study of group processes has many different focuses. It can focus on individual cognitive and motivational processes that produce group behaviors, interpersonal processes among more than two people, cognitive and social processes that cause people to conceive of themselves as group members and behave accordingly, intergroup relations that affect intragroup processes, the interrelationship of individual, interpersonal, and social processes, the behavior of specific groups or types of group, and so forth (e.g., Baron et al. 1992, Brown 2000, Hogg and Tindale 2001, Levine and Moreland 1998, Moreland and Levine 1994). What unites this diversity is a focus on the group as a social psychological entity.

1. What is a Group?

There is, however, little agreement on a social psychological definition of the group (see Hogg and Abrams 1988). The dominant view (since Floyd Allport in the early 1920s) is that a group is a collection of individuals, and that group processes are actually individual or interpersonal processes among a number of people. For instance, Latané's social impact theory attributes differences between interpersonal and group behavior to, among other things, the effects of increasing numbers of people. An alternative perspective (see Hogg and Abrams 1988) is that groups and group behavior are qualitatively distinct from individuals and interpersonal behavior, and that different or

additional concepts are needed to properly analyze groups—for example, groups produce norms that regulate members' behavior and people reconfigure their self-concept so that it is defined in terms of properties of the group as a collective. This perspective is represented by the early work of Wundt and Durkheim (late nineteenth century), and McDougall and Mead (early twentieth century), the more recent work of Sherif and Tajfel, and presently by social identity theory (see *Social Identity, Psychology of*) and intergroup approaches (see *Intergroup Relations, Social Psychology of*). It is argued that the analysis of groups in terms only of interpersonal processes is reductionist—it does not adequately address questions to do with group processes, but instead deals only with interpersonal processes among more than two people.

The issue of 'what is a group' continues to simmer as an important, and unresolved, metatheoretical debate in social psychology. Many researchers adhere to one view or the other, for example, perspectives that focus on the interdependence of individuals in small face-to-face aggregates vs. perspectives that focus on the collective self-concept that derives from membership of groups of all sizes. Others opt out of the formal controversy—this is particularly true of an increasing number of scholars who research group processes from more applied perspectives (e.g., sport psychology). In recent years, however, there have been promising attempts to integrate interactive small group perspectives, intergroup perspectives, and cognitive self-concept perspectives (see Brown 2000).

2. Studying Group Processes

Traditionally the study of group processes has lain at the heart of social psychology. It was in many ways the explosion in the late 1940s and early 1950s of research into what was then called group dynamics that set experimental social psychology off on its current expansion. Even at the beginning of the twenty-first century, early group dynamics researchers like Kurt Lewin and Leon Festinger have significant impact on the discipline. Group processes research thrived until about the mid-1960s when it suffered a fairly rapid decline in popularity (replaced by attribution theories, then social cognition), from which it is showing strong signs of recovery, mainly through cognitive and intergroup perspectives on group processes (Moreland et al. 1994). The apparent demise of traditional group processes research may be due to a variety of factors (Steiner 1974). For instance: (a) it is time consuming and costly to conduct research with interactive groups; (b) data from group research are rarely neat, tidy, and easily explained; (c) the computer revolution shifted social psychological interest from the study of groups to the development of theories of individual human behavior that resembled computer software; and (d)

the logical end-point of reductionist perspectives on group processes is to focus exclusively on individual psychology.

3. Group Formation, Commitment, and Identity

A central concept in the study of group processes is 'group cohesiveness,' which refers to the overall solidarity of groups, their 'groupness,' and the process that transforms an aggregate of unrelated individuals into a psychological group. Through cooperative interdependence in the pursuit of shared goals, people become attracted to one another, to the group, and to the group's goals. In this way a group forms, and its overall cohesiveness is an arithmetic combination of individual members' attraction to one another and the group. This attraction-to-group—or interpersonal attraction—analysis has not gone unchallenged, and recent perspectives place an emphasis on some form of commitment to group, or self-definition in terms of group membership (Hogg 1992). For instance, social identity theory does not consider attraction-to-group to be the basic process of group formation, but rather one of a range of consequences of a process of identification with the group (i.e., self-categorization in terms of the group's defining features—see *Social Identity, Psychology of*). Moreland and Levine (e.g. 1994) have developed a 'group socialization model' to address what they consider to be a lack of diachronic perspective in many models of group formation and development. Commitment is considered the most basic process. People become committed to a group, and a group to its members, through bilateral cost-benefit analyses of membership in this group relative to other groups. As mutual commitment rises or ebbs, members move in and out of different general roles within the group (e.g., new member, 'oldtimer')—a process characterized by role transitions that are quite marked, and which can involve public rites of passage.

4. Group Influence

Groups influence the attitudes and behaviors of their members in many different ways (e.g., Turner 1991—see *Social Influence, Psychology of; Attitude Change: Psychological*).

4.1 Direct Influence and Persuasion

Direct influence attempts involve individuals, subgroups, or the group as a whole trying to persuade members (e.g., deviants, dissenters) to go along with or come into line with the group. Persuasion attempts may range from relatively polite requests all the way to direct orders (see *Obedience: Social Psychological*

Perspectives) backed up by threats and ultimately force. Persuasion is generally more effective if the source of influence has a degree of legitimate authority and power within the group (see *Leadership, Psychology of*; Chemers 2001). Direct influence generally produces behavioral compliance which is not necessarily matched by underlying cognitive or attitudinal change—people go along with the group for instrumental reasons which may include a desire to be evaluated favorably or to avoid disapproval or punishment.

4.2 Indirect Influence and Group Norms

Groups can also influence members indirectly through the power of social norms that describe and prescribe appropriate behavior for group members in that context (Brown 2000, Turner 1991). In the absence of direct social pressure people conform to group norms. This may happen through ‘normative influence,’ in which people simply publicly comply in order to obtain social approval or avoid disapproval from others—there is no underlying cognitive change. Another process is ‘informational influence’ which produces true conformity—underlying cognitive change which may go along with behavioral change (see *Attitudes and Behavior*). Informational influence is effective where people are uncertain about the objective correctness of their beliefs, and in the absence of objective criteria they use group members’ behaviors as social criteria (Festinger 1954; see *Youth to Adulthood across Cultures, Transition from*). A third process is ‘referent informational influence’ (see *Social Identity, Psychology of*) in which people who identify with a group construct a contextually appropriate norm that defines the ingroup in contrast to relevant outgroups, and then enact the norm as an automatic consequence of self-categorization.

Although indirect group influence usually comes from the ingroup, there are at least two ways in which an outgroup can have influence: (a) outgroups influence the nature of the ingroup norm that people conform to, and (b) minority outgroups can adopt strategies to convert people to their viewpoint (e.g., by promulgating a consistent and consensual viewpoint, and by presenting themselves as actually part of the ingroup)—see *Social Movements: Psychological Perspectives*.

5. Group Structure

Whereas norms and social influence processes tend to produce uniformity within groups, other factors introduce unevenness and variability across the group that structures the group into different regions. Aspects of group structure can surface through intragroup patterns of who prefers or likes whom, known as ‘sociometric choice’ (see Hogg 1992).

5.1 Roles

Groups usually embrace a number of different roles that can be occupied by different people, or by subgroups of people. Roles prescribe behavior, and they emerge to assist the group to function better as an integrated entity. According to Moreland and Levine’s (1994) group socialization model, the group socialization process involves people moving in and out of quite distinct general roles within the group as a function of commitment to the group. Roles are rarely equal—some are more desirable than others, and have more power, prestige, and status attached to them. In this way, groups are often internally structured into role hierarchies. High status roles tend to be occupied by people who have specific status characteristics (abilities and skills directly related to the group’s function) and/or diffuse status characteristics (they are members of social categories that have high status in society at large)—see Ridgeway 2001.

5.2 Communication Structures

Another feature of group structure is the ease with which members or role occupants can communicate with one another (see *Communication and Social Psychology*; see Baron et al. 1992). Depending on the group’s task, different communication structures can emerge—highly centralized structures channel all communication through a single communication hub (often the leader), while less centralized networks allow freer communication among all members. The nature of the communication network affects group functioning, task performance, member satisfaction, and group solidarity. For instance, highly centralized networks can, under certain circumstances, reduce efficiency, cause peripheral members to feel excluded and dissatisfied, and even reduce group cohesion. Communication networks can be traced and described using Bales’s ‘interaction process analysis’ method, which documents who communicates what with whom, and how often.

5.3 Leadership

One of the most important, and certainly the highest status, role in a group is the role of leader (Chemers 2001, see *Leadership, Psychology of*). Research on leadership suggests that effective leadership (often defined in terms of the ability to motivate members to internalize and act on one’s own innovative vision for the group) is an interaction between leadership style and the type of leader that is needed in a particular situation. Another factor that may be important in some circumstances is that the leader is perceived to be a good group member who closely fits the defining features of membership. Because leadership is a role, leadership dynamics involve followers—if followers will not be led then the leader cannot lead.

Generally, followers tend to endorse leaders who have good leadership qualities, who have abilities that benefit and promote the group as a whole, and who are good group members or have been good group members in the past. These attributes encourage followers to endow the leader with status and to allow the leader to be innovative.

5.4 Deviance

Whereas leaders are at the pinnacle of the intragroup role hierarchy, deviants are at the bottom (see Hogg and Tindale 2001). Research suggests that groups try very hard to socialize marginal members by making them conform or by finding them useful and valued roles within the group. If these attempts fail, people evaluatively reject deviants, often pathologizing their behavior and attempting to classify them as outgroup members. Research on the 'black sheep effect' suggests that because marginal members pose a threat to the normative integrity of the group and thus threaten members' group membership-based self-concept, marginal ingroup members are actually disliked even more than an identical person who belongs to an outgroup.

6. Group Performance

Since groups often exist to perform specific tasks, an important question to ask is whether groups perform better or worse than individuals (see Baron et al. 1992). This is an extension of an even more fundamental question about how individual task performance is affected by the presence of others. Coactors or a passive audience tend to improve performance of a well-learned or easy task, but impede performance of a poorly learned or difficult task (see *Social Facilitation, Psychology of*). In general, however, groups are interactive entities, and the question becomes one of whether the task is better accomplished by interaction in a group, or simply by individuals working alone. The answer to this question is that it depends on the specific nature of the task—for instance, more cars can be produced by a group effort based on a coordinated division of labour, than by individuals building them from scratch (see *Group Productivity, Social Psychology of*).

6.1 Social Loafing, Free Riding, and Social Dilemmas

Groups often perform less well than would be suggested by the sum of the abilities or performances of individual members (see Moreland and Levine 1994). This can be because of coordination losses, that is, interference between individuals trying to coordinate their actions within the group. Coordination

losses may be one explanation of why group brainstorming is often a suboptimal way to generate novel ideas in a group. Groups can also perform less well because of motivation losses, that is, people in groups put in less effort because they are less motivated. This latter 'social loafing' effect seems to be partly due to a reduction in perceived personal identifiability and responsibility in the group (see *Deindividuation, Psychology of*), and can be explained in terms of 'social impact theory.' Social loafing can be reduced by, among other things, task importance, enjoyment, and relevance, and increased group cohesiveness and personal identifiability. Related to loafing is a tendency for individuals in a group confronted by a scarce resource to free ride, that is, maximize their own short-term gains with little regard for the group as a whole and thus their long-term gains (see *Social Dilemmas, Psychology of*). Many of the environmental problems faced by the world today are examples of social dilemmas. Social dilemmas can be (partly) avoided by having a leader manage a resource rather than letting the members compete directly, or by having members identify strongly with the group that accesses the resource.

7. Deindividuation, Collective Protest, and the Crowd

A common assumption about groups is that they cause people to lose their sense of individuality and that this leads to low motivation, lack of responsibility, and antisocial conduct—all captured by the notion of 'deindividuation' (Zimbardo 1970, see *Deindividuation, Psychology of*). Historically, this perspective, which has its critics (see Postmes and Spears 1998), is most evident in research on a particular type of group—the crowd. Starting with LeBon in the late nineteenth century, social psychologists have generally considered crowds to be irrational and antisocial because members of the crowd are deindividuated. A contrasting perspective is one that views crowd behavior as social protest (see *Social Movements: Psychological Perspectives*). Crowds have a purpose that is tied to a specific identity, and so crowd behavior has a logic to it because it is tightly regulated by the context relevant norms of the group. Where a crowd engages in antisocial acts, this is because the event is an expression of protest against a dominant group—the protest takes the form of violation of the dominant group's norms and rules.

8. Group Decision Making

One of the most common tasks that groups perform is decision making (Baron et al. 1992, see *Group Decision Making, Social Psychology of*). Group decision making is potentially superior to individual decision

making because in a group there is a larger pool of expertise, viewpoints, and so forth. Groups may come to decisions slowly, but this may be a good thing, particularly for complex issues, because it indicates proper consideration of all facets of the issue. However, groups are also subject to an array of group processes that may impair decision making, for example, factional conflict, biased leadership, and pressures to be obedient and compliant.

8.1 Groupthink and Group Polarization

One potential problem is the tendency for highly cohesive decision-making groups, that are under stress and do not have impartial leadership, to engage in 'groupthink,' a mode of thinking in which the desire to reach unanimous agreement overrides the motivation to adopt proper, rational group decision-making procedures. Groupthink involves defective decision-making procedures, which generally produce poor decisions that can have disastrous consequences. Many poor-quality crisis decisions made by governments have been attributed to groupthink, for example, the 1961 Bay of Pigs fiasco. Another potential bias in group decision making is 'group polarization' (see Turner 1991). This is a tendency for group discussion to cause groups to come to a decision that is more extreme than the mean of the individual members' prediscussion opinions. Polarization occurs only when the prediscussion mean already tends to favor one pole of the decision-making dimension—the group decision is then more extreme in this direction.

8.2 Social Combination Models

A major focus of research on group decision making concerns how one can predict a group decision from the initial distribution of members' views (Baron et al. 1992, Hogg and Tindale 2001). A number of social combination models have been derived to relate the distribution of initial opinions in a decision-making group to the final group decision. The most enduring of these is Davis's social decision schemes model (Davis 1973, which identifies at least five implicit (or explicit) decision rules that can exist in groups (e.g., unanimity, majority wins, truth wins). Different rules tend to prevail for tasks in which there is a demonstrably correct solution ('intellective tasks') as opposed to tasks which are more a matter of taste and preference ('judgemental tasks'). Rules also differ in terms of how strict they are (i.e., how much intragroup agreement is called for), and how much power concentration they have (i.e., how much power is located in one individual). Unanimity is very strict but low on power concentration, while two-thirds majority is less strict but has greater power concentration. Strictness and power concentration have effects on satisfaction

within the group, and on the speed with which a group makes a decision. Social combination models are often tested using a simulated jury paradigm.

8.3 Group Remembering

Group decision making usually involves collaborative recall of information, or group remembering (Baron et al. 1992, Hogg and Tindale 2001). Group remembering is not simply a retrieval activity in the pursuit of veridical recall, rather it is a constructive process where social comparisons among group members shape the resultant group memory. As such, group remembering may be subject to implicit or even intentional bias. Another feature of group remembering is 'transactive memory.' Groups often need to bring a large amount of information to bear upon a decision, and thus they develop a differentiated memory structure where individuals or subgroups specialize in different memory domains. Where this occurs people not only have to remember their own part of the puzzle, but they also need to know who specializes in what memories—they develop a transactive memory structure. Research shows that group functioning and decision making is compromised where people have not acquired a transactive memory, or where such a system has not developed in the group.

8.4 Intergroup Decision Making

Decision-making groups can contain factional rifts, or can be considered to be interactive decision-making contexts in which opposing groups or their representatives come together to reach an agreement. Under these circumstances the group provides a context for intergroup decision making (see *Intergroup Relations, Social Psychology of*). Because individuals are acting as group representatives rather than individuals, face-to-face intergroup bargaining has a relatively low probability of success—intransigence and stalemate is more likely. Third party mediation can help, and as a last resort, so can arbitration (in which a settlement is imposed by a third party).

9. The Intergroup Dimension

The study of group processes has tended to focus on what happens among members of small interactive groups (e.g., Levine and Moreland 1998, Moreland and Levine 1994), and does not fully consider how relations between groups may influence these processes. There are, however, indications that an intergroup perspective may be useful in helping to understand intragroup processes (Brown 2000, Hogg and Abrams 1988, see *Social Identity, Psychology of; Intergroup Relations, Social Psychology of*). For inst-

ance, group polarization may be an intergroup phenomenon in which a relatively extreme ingroup differentiates itself from a less extreme outgroup by becoming even more extreme. Group decision making may sometimes actually be intergroup decision making, and the commitment process that underlies group socialization involves an assessment not only of ingroup membership but also outgroup memberships. There is also an array of group behaviors that are more properly termed intergroup behaviors: these may include a more favorable evaluation of the ingroup, its members, and all it stands for, in relation to outgroups (i.e., ethnocentrism); discriminatory behavior against outgroups and in favor of the ingroup; and stereotypic perceptions and assumptions about ingroup and outgroup members (see *Stereotypes*). Also, the extreme forms of collective behavior that occur in crowds are sometimes best explained in intergroup terms—the crowd is a group with a passion that is guided and fuelled by the group's opposition to an outgroup (e.g., football supporter groups). Finally, group influence is not purely one-way, with the group bringing dissenters and deviates into line. It is often an intergroup affair with active minorities within the group trying to sway the majority to its own position.

10. Specific Groups

The study of group processes has tended to have numerically small, face-to-face, demographically homogenous, short-lived, task-oriented groups in mind (Steiner 1974). As such, a great deal of research has used four to six-person laboratory groups of university (often psychology) students. In many cases, experimental studies of group processes have used the dyad as the prototypical group. This is based on an assumption that group processes are merely aggregated interpersonal processes. More naturalistic studies have dealt with a great variety of groups—gangs, housing programs, sports teams, summer camps, scout troops, school classes, military groups, therapy groups, juries, government and organizational decision-making groups, and so forth. Studies in the intergroup tradition have tended to be interested in cognitive processes and so have often, though not exclusively, used rather abstract laboratory experiments where groups are often only nominal and there is no interaction.

11. Future Directions

There has been a marked revival of interest among social psychologists in group processes (Moreland et al. 1994). The key feature of this revival is an integration of intergroup, small group, and social cognitive perspectives in social psychology (see Brown

2000, Hogg and Tindale 2001) that sets a clear course for future developments. Recent developments have, however, tended to focus on generic group processes rather than the phenomenology of specific types of group. Although the study of larger, naturally occurring groups in more naturalistic contexts is time consuming, methodologically problematic, and involves weaker experimental control, future developments in group processes may be assisted by more of this type of research. Indeed there is evidence that although certain group processes may be common to all groups, a full understanding of groups also needs to recognize differences between small interactive groups and large social categories, between groups that vary in the degree to which they are a coherent entity, and between groups as diverse as a family, a sect, a therapy group, an army, a nation, a virtual group, and so forth (Deaux et al. 1995).

See also: Group Decision Making, Cognitive Psychology of; Group Decision Making, Social Psychology of; Group Processes in Organizations; Group Processes in the Classroom; Group Productivity, Social Psychology of; Stereotypes, Social Psychology of

Bibliography

- Baron R S, Kerr N L, Miller N 1992 *Group Process, Group Decision, Group Action*. Brooks/Cole, Pacific Grove, CA
- Brown R 2000 *Group Processes: Dynamics Within and Between Groups*, 2nd edn. Blackwell, Oxford
- Chemers M M 2001 Leadership effectiveness: An integrative review. In: Hogg M A, Tindale S T (eds.) *Blackwell Handbook of Social Psychology: Group Processes*. Blackwell, Oxford, UK, pp. 376–99
- Davis J H 1973 Group decision and social interaction: A theory of social decision schemes. *Psychological Review* **80**: 97–125
- Deaux K, Reid A, Mizrahi K, Ethier K A 1995 Parameters of social identity. *Journal of Personality and Social Psychology* **68**: 280–91
- Festinger L 1954 A theory of social comparison processes. *Human Relations* **7**: 117–40
- Hogg M A 1992 *The Social Psychology of Group Cohesiveness: From Attraction to Social Identity*. Harvester Wheatsheaf, London
- Hogg M A, Abrams D 1988 *Social Identifications: A Social Psychology of Intergroup Relations and Group Processes*. Routledge, London
- Hogg M A, Tindale R S (eds.) 2001 *Blackwell Handbook of Social Psychology: Group Processes*. Blackwell, Oxford
- Levine J M, Moreland R M 1998 Small groups. In: Gilbert D T, Fiske S T, Lindzey G (eds.) *The Handbook of Social Psychology*. McGraw-Hill, New York, Vol. 2, pp. 415–69
- Moreland R L, Hogg M A, Hains S C 1994 Back to the future: Social psychological research on groups. *Journal of Experimental Social Psychology* **30**: 527–55
- Moreland R L, Levine J M 1994 *Understanding Small Groups*. Allyn and Bacon, Boston
- Postmes T, Spears R 1998 Deindividuation and anti-normative behavior: A meta-analysis. *Psychological Bulletin* **123**: 238–59

- Ridgeway C L 2001 Social status and group structure. In: Hogg M A, Tindale S T (eds.) *Blackwell Handbook of Social Psychology: Group Processes*. Blackwell, Oxford, pp. 352–75
- Steiner I D 1974 Whatever happened to the group in social psychology? *Journal of Experimental Social Psychology* **10**: 94–108
- Turner J C 1991 *Social Influence*. Brooks/Cole, Pacific Grove, CA
- Zimbardo P G 1970 The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. In: Arnold W J, Levine D (eds.) *Nebraska Symposium on Motivation* 1969. University of Nebraska Press, Lincoln, NE, Vol. 17, pp. 237–307

M. A. Hogg

Group Productivity, Social Psychology of

Group productivity refers to the degree to which working in a group context either enhances or reduces individual motivation and performance. Group productivity is one of social psychology's cornerstones. The earliest documented studies in social psychology examined the effects of groups on individuals, and hundreds of laboratory and field studies have been conducted on this issue.

1. Historical Background

The earliest studies revealed two seemingly opposing consequences of group work. In the 1880s, Max Ringelmann observed that oxen yoked together did not pull plows as fast as would be expected from their individual efforts. In his laboratory, he asked male volunteers to pull on a rope as hard as they could, both alone and as members of groups ranging in size from two to six. He found that as group size increased, group performance was increasingly lower than would be expected from the simple addition of individual performances, suggesting that working in a group could reduce performance (see Kravitz and Martin 1986). In 1898, Norman Triplett noted that competitive cyclists were faster when racing against other cyclists than when racing against the clock. In his laboratory, he demonstrated that children worked harder on a physical task when they were in the presence of other children who were working on the same task compared to when they worked alone, suggesting that working in a group could enhance performance.

Triplett's work helped stimulate a flurry of studies on what came to be known as 'social facilitation,' revolving around the idea that the presence of others could facilitate individual performance. Meanwhile, Ringelmann's findings were essentially ignored, regarded with skepticism, or interpreted solely in terms of poor coordination of group members' activities. It was not until 1974 that Ringelmann's findings were

replicated, and not until even later that the motivational component of these findings was examined as an important and reliable phenomenon that came to be known as 'social loafing.'

2. Social Facilitation

Though Triplett's findings were replicated early and often, inconsistencies began to emerge across studies throughout the first half of the twentieth century. Specifically, the presence of others, either as an audience or as coactors (i.e., individuals who work on the same task in each other's presence but who do not combine their outputs into a group product), did not always enhance performance and sometimes reduced performance. Zajonc (1965) devised an elegant solution to this discrepancy. He suggested that the presence of other people increases generalized drive, thereby facilitating dominant responses. On simple or well-learned tasks, dominant responses are likely correct and the presence of others should enhance performance. On complex or unfamiliar tasks, dominant responses are likely to be in error and the presence of others should reduce performance.

Later research has largely supported Zajonc's view (for a review, see Bond and Titus 1983). Indeed, most researchers now define social facilitation as the tendency for the presence of others to facilitate performance on simple or well-learned tasks but to impair performance on complex or unfamiliar tasks. However, there is an active controversy among researchers and theorists as to why and when the presence of others facilitates dominant responses. The mere presence perspective of Zajonc (1965) and others focuses on arousal and assumes that the presence of other people introduces an element of uncertainty to a situation that leads people to become more aroused and alert, regardless of whether or not the other people present have any interest in what the individual is doing. Cottrell's (1972) evaluation apprehension theory states that mere presence is not sufficient to produce arousal, but rather that it is the ability of other people to create apprehension about being evaluated that leads to social facilitation. If others are present but are paying no attention to the individual and have no regard for how well the individual performs, performance should not be affected. Finally, distraction-conflict theory (Baron 1986) focuses on attention rather than arousal as the primary mediator of social facilitation. Research has documented repeatedly that the presence of others can be distracting and can divide an individual's attention between working on the task at hand and monitoring the behavior and reactions of the audience or coactors. Distraction-conflict theory suggests that individuals try to resolve this cognitive and attentional conflict by increasing their effort at the task to overcome the effects of distraction, thereby enhancing performance on simple tasks and reducing it on complex tasks.

Research has found some support for all of these theories. It appears as though mere presence, evaluation apprehension, and distraction are all capable of creating and contributing to social facilitation effects. Arousal levels have been documented to increase in the mere presence of others, and support has been found for social facilitation even when the others present are very unlikely to create any apprehension about evaluation. Arousal does become more pronounced, however, when observers are clearly evaluating the individual. Similarly, studies have shown that social facilitation effects can emerge from situations in which the presence of others creates distraction. Thus, arousal, evaluation apprehension, and distraction can all potentially contribute to social facilitation effects, both alone and in combination with one another (for a detailed discussion of these factors, see Geen 1991).

Research has also documented social facilitation effects across a wide array of settings, tasks, and subject populations. The effect also appears to hold for a wide variety of animals in addition to humans, including armadillos, chickens, monkeys, and rats. In one rather unusual study, the ability of cockroaches to run either a simple runway or a maze was examined either alone, in coacting pairs, or in the presence of other cockroaches who observed from behind a plastic barrier (Zajonc et al. 1969). When in the presence of coactors or observers, cockroaches ran the simple runway more rapidly but took longer to navigate the maze, replicating prior social facilitation research on humans and other animals. Social facilitation has also been widely documented in field settings. As just two examples, social facilitation effects have been documented for gymnastics performance and billiards playing in public settings.

However, although social facilitation effects are well-established, they appear to be somewhat small in magnitude in comparison with other prominent social psychological phenomena and the impact of facilitation on performance sometimes varies across performance measures. Based on their meta-analysis of 241 studies, Bond and Titus (1983) concluded that, on simple tasks, the presence of others increased the rate of productivity but did not improve performance quality. They also concluded that, on complex tasks, the presence of others reduced both the quantity and quality of performance. Finally, they noted that the productivity gains associated with simple tasks were weaker in magnitude than the productivity losses associated with complex tasks performed in the presence of others.

3. Social Loafing

Formally, social loafing is a reduction in motivation and effort when individuals work collectively compared with when they work individually or coactively.

In order to study the effects of group context on individual motivation while controlling for mere presence effects, social loafing researchers usually compare individual performance on collective tasks to individual performance on coactive tasks, while holding group size constant across these conditions. In addition, in order to separate motivation loss from mere coordination loss (i.e., poor coordination of group member's actions), social loafing researchers often employ pseudogroups in which individuals work alone or produce outputs that can be individually monitored by the researcher, but believe that their outputs are being combined with those of others into a group product.

These methodological features were prominent in a highly influential experiment by Latané et al. (1979). Participants shouted and clapped as loudly as they could both individually and with others. While blindfolded and wearing headphones to mask the noise, participants shouted both in actual groups, and in pseudogroups in which they shouted alone but believed they were shouting with others. Participants still reduced their efforts on the pseudogroup trials, even though they only thought they were working with others. Thus, Latané and colleagues (1979) demonstrated that a substantial proportion of the decreased performance of groups was attributable to reduced individual effort, as distinct from coordination loss. They also coined the term 'social loafing' for the demotivating effects of working collectively.

Since the late 1970s, more than 90 studies on social loafing have been conducted. Both laboratory experiments and field studies have been conducted, using a wide variety of tasks and examining a range of subject populations. The bulk of the available evidence suggests that social loafing is moderate in magnitude and consistently found across studies. A meta-analysis of 78 studies by Karau and Williams (1993) found that the magnitude of social loafing was comparable in size to that of a number of prominent social psychological effects. Social loafing has been found for a wide variety of tasks, including evaluative tasks such as rating the quality of editorials and poems, cognitive tasks such as navigating computer mazes and detecting visual signals, creative tasks such as song-writing and thought-listing, physical tasks such as pulling on a rope and swimming, and work-related tasks such as typing. Furthermore, although national differences are occasionally found in some studies, a number of studies have documented social loafing regardless of participants' gender, nationality, or age. Thus, social loafing is robust and generalizes across tasks, as well as across most subject populations (Karau and Williams 1993).

However, social loafing is not inevitable and a number of factors have been found to moderate the effect. Thus, studies have found that social loafing can be reduced or eliminated by making individual inputs identifiable, enhancing task meaningfulness or personal involvement with the task, providing individual

or group comparison standards, providing incentives for good performance, threatening punishment for poor performance, increasing personal perceptions of efficacy at the task, and increasing the uniqueness of individual contributions (Karau and Williams 1993). Finally, gender and culture both appear to moderate social loafing. In their meta-analysis, Karau and Williams (1993) found that the magnitude of social loafing across studies was lower for women and for people in Eastern cultures such as Japan, Taiwan, and China, than it was for men and for people in Western cultures such as the USA and Canada, although significant effort reductions were still found within each of these groups.

A number of theories have been offered to explain social loafing, of which four perspectives are most prominent in the literature. First, social impact theory (Latané 1981) views people as either sources or targets of social influence and states that the amount of social impact experienced in a situation is a function of the strength (status or power), immediacy (physical or psychological closeness), and number of sources and targets present. On collective tasks, the influence of the outside influence source (the manager or the experimenter) is diffused across all of the group members, making it less potent on collective tasks than on individual tasks.

Second, the arousal reduction viewpoint (Jackson and Williams 1985) builds on social impact theory to suggest that the presence of others actually reduces arousal when those others are co-workers rather than coactors. According to this view, social loafing is the flip side of social facilitation because the presence of other co-workers lead to reduced drive, and therefore to reduced performance on simple tasks but enhanced performance on complex tasks.

Third, the evaluation potential perspective (Harkins 1987) suggests that social loafing arises because individuals can often be evaluated when working coactively, but are harder to evaluate when working collectively. Thus, individuals can 'hide in the crowd' on groups tasks and avoid taking the blame for a poor group performance. Similarly, they may also feel 'lost in the crowd' and unable to take credit for a good group performance.

Fourth, expectancy-value perspectives (Karau and Williams 1993, Kerr 1983, Sheppard 1993) suggest that individuals work less hard on a collective task because they feel their efforts will not be instrumental in obtaining an outcome that they personally value. As delineated in detail in a model by Karau and Williams (1993), collective tasks pose unique barriers to motivation both because they create more obstacles to the perceptions of efficacy and instrumentality, and because they may undermine the value attached to the task. All of these theories appear to have some usefulness, with the expectancy-value approach representing a more comprehensive or integrative approach to understanding individual motivation in

groups, and the other viewpoints explaining specific mediators of social loafing quite well, albeit in a more restricted fashion.

4. The Relationship Between Social Facilitation and Social Loafing

Both social loafing and social facilitation are concerned with the effects of group contexts on motivation and performance and yet social loafing suggests that this presence decreases performance whereas social facilitation suggests that this presence increases performance for some tasks but decreases performance for other tasks. Indeed, the very first empirical studies of group productivity by Ringelmann and Triplett documented these seemingly contradictory effects of groups.

The key to resolving this discrepancy appears to be in specifying whether the other people present are co-workers or coactors. In social loafing studies, groups are composed of co-workers who pool their efforts together. In social facilitation studies, however, groups are composed of coactors who work in one another's presence but do not pool their efforts. Therefore, while both types of studies explore individual motivation in social contexts, they examine two very distinct aspects of group life. Taken together, they suggest that the impact that other people have on an individual's effort depends on whether or not the individual is actually working with those others on a collective task. Collective tasks have potentially demotivating properties, whereas coactive tasks tend to increase drive, thereby enhancing performance on simple tasks but reducing it on complex tasks. In other words, social loafing research is targeted to collective group work, whereas social facilitation research is targeted to the mere presence of coactors or observers.

5. Future Directions

Existing research on social facilitation and social loafing has revealed many important dynamics of group life, highlighted the conditions under which groups are likely to enhance or reduce individual performance, and provided a number of theoretical perspectives for understanding the dynamics of social motivation. Additional research is likely to provide an even richer understanding and greater guidance as to how groups can be designed and managed to bring out the best in individuals.

See also: Group Decision Making, Cognitive Psychology of; Group Decision Making, Social Psychology of; Group Processes in Organizations; Group Processes in the Classroom; Group Processes, Social Psychology of

Bibliography

- Baron R S 1986 Distraction-conflict theory: Progress and problems. In: Berkowitz L (ed.) *Advances in Experimental Social Psychology*. Academic Press, New York, Vol. 19, pp. 1–40
- Bond C F, Titus L J 1983 Social facilitation: A meta-analysis of 241 studies. *Psychological Bulletin* **94**: 265–92
- Cottrell N B 1972 Social facilitation. In: McClintock C G (ed.) *Experimental Social Psychology*. Holt, Rinehart, and Winston, New York, pp. 185–236
- Geen R G 1991 Social motivation. *Annual Review of Psychology* **42**: 377–99
- Harkins S G 1987 Social loafing and social facilitation. *Journal of Experimental Social Psychology* **23**: 1–18
- Jackson J M, Williams K D 1985 Social loafing on difficult tasks: Working collectively can improve performance. *Journal of Personality and Social Psychology* **49**: 937–42
- Karau S J, Williams K D 1993 Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology* **65**: 681–706
- Kerr N L 1983 Motivation losses in small groups: A social dilemma analysis. *Journal of Personality and Social Psychology* **45**: 819–28
- Kravitz D A, Martin B 1986 Ringelmann rediscovered: The original article. *Journal of Personality and Social Psychology* **50**: 936–41
- Latané B 1981 The psychology of social impact. *American Psychologist* **36**: 343–56
- Latané B, Williams K, Harkins S 1979 Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology* **37**: 822–32
- Sheppard J A 1993 Productivity loss in performance groups: A motivation analysis. *Psychological Bulletin* **113**: 67–81
- Triplett N 1898 The dynamogenic factors in pacemaking and competition. *American Journal of Psychology* **9**: 507–33
- Zajonc R B 1965 Social facilitation. *Science* **149**: 269–74
- Zajonc R B, Heingartner A, Herman E M 1969 Social enhancement and impairment of performance in the cockroach. *Journal of Personality and Social Psychology* **13**: 83–92

K. D. Williams and S. J. Karau

Group Psychotherapy, Clinical Psychology of

Group psychotherapy is the treatment of psychological, behavioral, emotional, and interpersonal problems in a group context. This article traces the development of group procedures in clinical psychology from its roots as a didactic technique to contemporary psychoanalytic, behavioral and cognitive-behavioral, and interpersonal treatment procedures. These approaches stress different therapeutic mechanisms and methods, but nearly all trace the health-promoting effects of groups to a common set of factors that instill a sense of hope, promote social learning, and induce self-disclosure, emotional ex-

pression, and self-insight. Research suggests that group psychotherapy, when properly administered, is as effective as individualized treatment techniques.

1. Approaches to Group Psychotherapy

Group psychotherapy emerged as a means of helping people deal with mental and emotional problems in the early years of the twentieth century. Psychotherapy itself was developing during this period, but most pioneering psychotherapists worked with individual patients only. Some innovators, however, applied their basic change principles in group settings, as did some therapists who turned to groups for didactic purposes or because their patient load demanded a massed approach to treatment. These group applications were often confined to a particular psychiatric facility or clinic, but in time they became more dispersed as other mental health professionals heard of and made use of the group methods. These early psychotherapy sessions involved treating individuals 'in groups, with the group itself constituting an important element in the therapeutic process' (Slavson 1950, p. 42).

Group psychotherapy is used currently to treat a variety of problems, including addiction, thought disorders, depression, eating disorders, and personality disorder. Group psychotherapists are usually credentialed mental health professionals and their patients tend to be suffering from diagnosed clinical conditions. The methods used by group psychotherapists are as varied as individual approaches to therapy. Psychoanalysis, psychodrama, systems-theory, object relations, existential, Gestalt, and humanistic are just a few of the general orientations adopted by group psychotherapists, who tend to be more eclectic than nongroup therapists in their approaches to treatment. Most approaches, however, fall into one of three basic categories: psychoanalytic, behavioral and cognitive-behavioral, and interpersonal (Forsyth and Corazzini 2000).

1.1 Psychoanalytic Group Psychotherapy

Psychoanalytic group psychotherapy draws on Freud's (1922) conceptualization of the causes of dysfunction and his recommendations regarding the most effective way to treat those dysfunctions (see *Psychoanalysis in Clinical Psychology*). The analyst, through such traditional analytic methods as free association, interpretation, and the dissection of transference processes, creates a powerful relationship with the client who then gains insight into unresolved conflicts. The psychoanalytic group psychotherapist, however, exploits the group situation to stimulate self-insight, transference, and the cathartic release of unconscious anxieties. In most cases the therapist acts

as the authority in the group and works with each patient in turn. This rotation gives each patient the opportunity to experience the benefits of one-on-one psychotherapy, but also gives them the opportunity to gain from the analysis of co-members' psychological and emotional difficulties. The group setting, as a recapitulation of a family grouping, also provides more opportunities to work through problems that result from early family conflicts. Just as individual therapy usually stimulates parental transference, group psychoanalysis stimulates sibling transference. Indeed, some experts suggest that Freud's Vienna Circle—ostensibly a study group formed by Freud and his students to discuss theories of adjustment—was a form of group psychotherapy (Roth 1993).

Psychoanalytic group psychotherapists diverge on one key aspect of the process. Some stress the importance of the individual in the group. This approach argues that the therapist should stress individual-level processes in treatment, and avoid spending valuable treatment time exploring the dynamics of the therapeutic group. Others, in contrast, integrate the treatment of the individual with the analysis of the group itself. The group-as-a-whole approach, which can be traced to Bion, Rickman, and Foulkes's work at the Northfield Military Hospital in England, is a sterling example of this approach. Although group-as-a-whole therapists embrace psychoanalytic assumptions of unconscious motivations, personality conflicts, and transference, they make use of the tension between the group and the individual to promote growth and development. Bion, for example, stressed the analysis of defense mechanisms used by groups to cope with ego threats, uncertainties, and anxieties. He believed that these strategies are rooted in dependence, for the group members engage in collective projective identification as they struggle to avoid blame for their problems by shifting responsibility from themselves to the therapist. Bion assumed that rapid and lasting adaptive change could be achieved by helping group members understand both individual and collective defensive processes that unfold in their group session, and his work provided the basis for the Tavistock Institute of Human Relations (Harrison 2000).

1.2 Cognitive–Behavioral Therapy Groups

Cognitive and cognitive–behavior psychotherapists, rather than searching for the cause of the problematic behavior in unseen unconscious conflicts or interpersonal transactions, base their interventions on principles derived from learning theories (Dobson et al. 2000) (see *Dialectical Behavior Therapy*; *Cognitive Therapy*). They identify and quantify desirable cognitions and behaviors that will be encouraged and undesirable cognitions and behaviors that will be extinguished, and then modify these behaviors thro-

ugh cognitive restructuring, self-instructional training, and problem-solving instruction. Cognitive–behavioral therapy groups use these principles with two or more individuals (Rose 1993). Cognitive–behavioral group therapy is generally both structured and problem-focused. Pretreatment assessments identify a patient's current level of functioning, and interventions are calibrated to fit the patient's level of functioning and targeted outcomes. The therapeutic interventions usually use a constellation of behavioral procedures, including modeling, rehearsal, and feedback.

1.3 Interpersonal Group Psychotherapy

Many group psychotherapists not only exploit the social processes that occur in groups during treatment, but they also trace psychological disturbances back to social sources—particularly interactions with friends, relatives, and acquaintances. Interpersonal theorists assume that behavior results from individuals' failures to take note of, and correct, cognitive and behavioral tendencies that disrupt their relationships with other people (see *Interpersonal Psychotherapy*).

Many group therapists, recognizing the interpersonal bases of psychological problems, use the group setting to help members examine their interpersonal behavior. Yalom's (1995) interpersonal group psychotherapy (also called interactive group psychotherapy), for example, stresses the 'here-and-now' of the group's interaction over the analysis of events and experiences that occur outside of the group (the 'then-and-there'). Members may spend some time, especially early in the group's life span, discussing problems they are facing at home or work, but Yalom feels that much of this discussion will be futile since members' characterize such events inaccurately. Yalom instead recommends that members focus on interactions within the group, with each member eventually responding to one another as they do to others in their lives. The group thus becomes a 'social microcosm,' giving members true insight into the causes of the personal problems.

2. Therapeutic Factors in Groups

Group psychotherapy is based on the Lewin's law of psychological change: 'It is easier to change individuals formed into a group than to change any of them separately' (Lewin 1951, p. 228). But what aspects of the group promote positive change? A common-factors perspective on group therapy suggests that various types of group psychotherapies, despite their many differences, share some common properties, and these commonalities may be responsible for their effectiveness. A psychodynamic therapist uses different techniques than a cognitive-

behavior or interpersonal therapist, but these various formats share mechanisms that promote change and sustain well-being. Yalom (1995) terms these shared qualities curative factors, and his list includes the installation of hope, universality, the imparting of information, altruism, the corrective recapitulation of the primary family group, the development of socializing techniques, imitative behavior, interpersonal learning, group cohesiveness, catharsis, and existential factors.

2.1 Social Motivation in Group Psychotherapy

Groups, by their very nature, satisfy a range of basic human social motives. Although individuals who are suffering from some behavioral or psychological problem may not find 'safety in numbers,' they can find 'reassurance in numbers.' When they assemble in a therapeutic group they quickly realize that the situation they face is shared by others. Yalom (1995) calls this realization universality, for patients realize that they have not been singled out to bear some extraordinary misfortune; rather, the problem they face is a universal one.

Therapeutic groups also capitalize on social comparison processes to stimulate growth and adjustment. Social comparison occurs when individuals spontaneously compare themselves to other people around them. This comparative process does not occur in individualistic therapies, for patients can only imagine how they are coping relative to others. In groups, in contrast, patients learn, both directly and indirectly, about one another's functioning. These comparative data yield two useful consequences. First, comparison provides patients with a baseline to calibrate their own level of functioning and determine where they stand along the 'coping vs. collapsing' continuum (Yalom 1995). They may learn, through downward social comparison—comparing themselves to those in the group who are performing poorly—that their own situation is not so bleak as they originally assumed. Second, this comparison can increase their sense of optimism about reaching an improved level of functioning. Yalom (1995) terms this process the installation of hope. Whereas individuals facing unpleasant circumstances alone may feel discouraged and pessimistic, when group members recognize that others are improving their motivational levels.

2.2 Interpersonal Learning

Much of the learning that occurs during psychotherapy is social in nature. Under the guidance of the therapist, patients explore and gain insight into sources of anxiety, they identify and practice behaviors that they can use to cope with sources of stress, and they acquire the skills they need to interact with others. The therapist models the behaviors they want

clients to adopt, and shape their behaviors through feedback and selective reinforcement. When therapy is delivered in a group context, however, the sources of social learning are multiplied. Group psychotherapy necessarily involves patients interacting with other patients, and so it sets up opportunities for interpersonal learning among patients. Some therapists purposely minimize the impact of the group on the individual patients, but most group therapists exploit these social processes to promote change.

Yalom (1995) argues that much of the curative power of a group springs from such social processes as imparting of information from one person to another (guidance), the development of socializing techniques, imitative behavior, and interpersonal learning. For example, the exchange of information among participants was one of the first purposes of groups, for the earliest group therapies were primarily didactic in nature, with psychotherapists lecturing to large groups of patients about mental and physical health. Interpersonal learning also occurs indirectly, as group members monitor their impact on the other people within their group, and the psychotherapist helps them draw conclusions about their maladaptive interpersonal tendencies. Group members also gain from simple observation of others. By watching others' emotional displays, problem-solving behaviors, coping strategies, and interpersonal maneuverings, patients acquire adaptive social skills.

2.3 Emotional Experiences and Group Development

Groups set the stage for strong emotional experiences for members. As therapeutic groups develop over time, members become closely connected to each other emotionally, developing bonds of friendship that are as strong as any they experience in nontherapeutic relationships. Members also report a strong level of identification with the group, with the result that they are more willing to accept the group's influence.

The range of emotions experienced depends on the group's stage of development. Early in the group interactions tend to be superficial, but in time conflicts tend to surface. In most groups, these conflicts between members and the group therapist give way to the development of cohesion. This progression of emotional experience, although characterized in a variety of ways, generally follows the forming, storming, norming, performing, and adjourning pattern identified by Tuckman (Tuckman and Jensen 1977). During the forming stage members are uncertain how to behave, and do not trust one another enough to disclose personal information or to admit they need help. During the storming stage group members often find themselves in conflict over the purpose of the group, and they may withdraw from the group and the therapist. In groups where the leader takes a non-directive orientation, this conflict stage focuses on the leader's failure to 'take charge' of the interactions.

During the next phase, norming, the group becomes more cohesive, as norms and standards emerge that dictate how each session will unfold. This stage makes the performing stage possible, where group members engage earnestly in the exchange of information that will help them achieve their therapeutic goals. In time, members must deal with issues of termination as the group adjourns. This movement from one stage to another is therapeutic, so much so that a group that does not move through these stages may not benefit its members (Yalom 1995).

2.4 Self Processes in Therapeutic Groups

All group psychotherapies stress the value of groups as vehicles of self-understanding and self-efficacy. Although group members may initially be reluctant to disclose any information about themselves in the relatively public group setting, as they develop stronger bonds of trust their disclosures change from superficial to substantial. Group members also report feeling gains in self-efficacy as they help others in their group with their problems.

Members particularly value the self-insights they achieve in groups. In interpersonal therapy, in particular, patients often act in the therapeutic group as they do in other social contexts, so much so that the behaviors that caused them interpersonal difficulties erode their relationships within the group. When their fellow group members and therapist call their attention to the parallel, patients experience greater insights into their own motives and emotions. In many cases, as Freud's analysis of groups suggests, these tendencies can be traced to early childhood experiences in the family of socialization.

3. The Effectiveness of Group Psychotherapy

When they were initially proposed, skeptics questioned the wisdom of putting people who are suffering from psychological problems together in one group. How, they asked, could troubled individuals be expected to cope in a group when they had failed individually? And how could the therapist guide the therapeutic process in a group, given the subtle nuances that pervade one-to-one therapies? In time, however, the method gained widespread acceptance as therapists added group sessions, either supplementing or completely replacing their individual sessions.

But do members leave the group better adjusted than when they entered? Meltzoff and Kornreich (1970) were guardedly optimistic about the utility of group therapies because 80 percent of the methodologically sound studies reported either major or minor benefits for clients, whereas nearly all of the studies that reported no benefit were methodologically flawed. Bednar and Kaul (1994) were guardedly

positive, although they questioned the quality of the outcome research in group psychotherapy. Meta-analytic reviews also suggest that group approaches are as effective as individual methods (e.g., Fuhrman and Burlingame 1994). Fuhrman and Burlingame (1994), for example, reviewed 700 group therapy studies and seven meta-analytic studies before concluding that group methods are effective treatments for a wide variety of psychological problems. Similarly, Faith et al. (1995), in a meta-analytic review of 63 studies of sensitivity training, concluded that these groups generally led to increases in self-actualization and self-esteem and improved interpersonal relations. They noted that these effects increased in larger groups, when the groups met for longer periods of time, and when the measures focused on behavioral outcomes rather than self-reported ones. McRoberts et al. (1998) found that group therapies were more effective with clients (a) who were not diagnosed clinically, (b) suffering from substance abuse problems and/or chemical dependencies, and (c) who attended 10 or fewer sessions. Older studies—those conducted prior to 1980—were more likely to favor group over individual approaches.

This positive conclusion, however, requires some qualification. First, the empirical evidence is not definitive. Whereas a number of reviews are positive, others conclude that group therapy is not as potent as individual therapy. Second, the changes brought about by group experiences may be more perceptual than behavioral. Group members frequently report achieving major insights into their behavior and are positive about the gains they have achieved, but objective indices of outcome are more contradictory. Self-reported insight is also related to actual improvements in adjustment. Third, as Bednar and Kaul (1994) note, a participant may decide to leave the group before they have benefited in any way, and in rare cases patients are significantly harmed by the group experience. Last, far more data of better quality are needed to settle the question of efficacy. As Forsyth and Corazzini (2000, p. 325) conclude, the 'use of varied and undocumented therapeutic methods, with different types of clients, by therapists who differ in skills and experiences, in studies that too frequently lack valid measures and adequate controls, make it difficult to draw firm conclusions.'

Group psychotherapy's empirically supported usefulness is reason enough for its continued growth, but the approach will also prosper as health maintenance organizations require, and patients increasingly seek, group-level treatment methods. Group psychotherapy is an efficient therapy, for a single therapist can treat more than one individual at a time—a characteristic that is considered very valuable in healthcare systems that seek to minimize costs without cutting quality. But group approaches may also become more popular as clients seek out supportive, stimulating environments for achieving personal change. Indeed, the

increasing popularity of self-help groups signals the need for treatments that take advantage of groups for promoting adjustment. Although the use of groups in business and educational settings was once considered innovative, individuals now expect to learn and work in group contexts. These experiences, combined with a growing collectivism in cultures that were once primarily individualistic, should heighten awareness of groups as a means of treatment.

See also: Cognitive Therapy; Group Processes, Social Psychology of; Groups in Special Environments; Interpersonal Psychotherapy; Psychoanalysis in Clinical Psychology; Psychological Treatment, Effectiveness of; Therapist–Patient Relationship

Bibliography

- Bednar R L, Kaul T 1994 Experiential group research: Can the canon fire? In: Garfield S L, Bergin A E (eds.) *Handbook of Psychotherapy and Behavior Change*, 4th edn. Wiley, New York, pp. 631–63
- Dobson K S, Backs-Dermott B J, Dozois D J A 2000 Cognitive and cognitive-behavioral therapies. In: Snyder C R, Ingram R E (eds.) *Handbook of Psychological Change*. Wiley, New York, pp. 409–28
- Faith M S, Wong F Y, Carpenter K M 1995 Group sensitivity training: Update, meta-analysis, and recommendations. *Journal of Counseling Psychology* **42**: 390–9
- Forsyth D R, Corazzini J G 2000 Groups as change agents. In: Snyder C R, Ingram R E (eds.) *Handbook of Psychological Change*. Wiley, New York, pp. 309–36
- Freud S 1922 *Group Psychology and the Analysis of the Ego*. Hogarth, London
- Fuhrman A, Burlingame G M 1994 Group psychotherapy: Research and practice. In: Fuhrman A, Burlingame G M (eds.) *Handbook of Group Psychotherapy: An Empirical and Clinical Synthesis*. Wiley, New York, pp. 3–40
- Harrison T 2000 *Bion, Richman, Foulkes, and the Northfield Experiments*. Kingsley, London
- Lewin K 1951 *Field Theory in Social Science*. Harper, New York
- McRoberts C, Burlingame G M, Hoag M J 1998 Comparative efficacy of individual and group psychotherapy: A meta-analytic perspective. *Group Dynamics: Theory, Research, and Practice* **2**: 101–17
- Meltzoff J, Kornreich M 1970 *Research in Psychotherapy*. Atherton, New York
- Rose S D 1993 Cognitive-behavioral group psychotherapy. In: Kaplan H I, Sadock M J (eds.) *Comprehensive Group Psychotherapy*, 3rd edn. Williams and Wilkins, Baltimore, MD, pp. 205–14
- Roth B E 1993 Freud: The group psychologist and group leader. In: Kaplan H I, Sadock M J (eds.) *Comprehensive Group Psychotherapy*, 3rd edn. Williams and Wilkins, Baltimore, MD, pp. 10–21
- Slavson S R 1950 Group psychotherapy. *Scientific American* **183**(6): 42–5
- Tuckman B W, Jensen M A C 1977 Stages of small group development revisited. *Group and Organizational Studies* **2**: 419–27

Yalom I D 1995 *The Theory and Practice of Group Psychotherapy*, 4th edn. Basic Books, New York

D. R. Forsyth

Groups in Special Environments

1. Introduction

‘Special’ environments include those that require extraordinary facilities for human survival, as well as those that temporarily pose dangers because of disaster, accident, or violent conflict. Also known as ‘extreme and unusual environments’ (EUEs), they have always engaged high public interest. Since the 1950s, social scientists have begun the systematic study of how people react in and are affected by experiences in hazardous habitats. These include life-support capsules (under water, in space, near the Poles, on offshore oilrigs), natural and industrial cataclysms, combat zones, refugee and concentration camps.

To date, most of the research emphasis has been on the ways in which people react to their experience (Wilson et al. 1988), on methods for preventing, minimizing, or treating negative outcomes (Holland and Curtis 1998), and on how individuals are selected and trained for some special environments, such as space and submarine missions (e.g., Santy 1994). Methods used in these efforts have included some experiments; but the emphasis has been on simulation and analogue studies, questionnaires, interviews, and the content analysis of archival materials.

Although researchers have focused on the individual, the behavior of human beings in special environments is to a great extent the story of groups. Solitary individuals have found themselves in such situations, sometimes by accident (shipwrecked mariners, sole survivors of other disasters) and sometimes on purpose (aboriginal youths on spirit quests, solo mountaineers, hermits, single-handed sailors, early astronauts, and adventurers of various sorts); but the modal inhabitants of EUEs are not alone.

Many aspects of the environment itself strongly influence group relations and behavior. Some of these are: the dangers involved, life-support equipment, duration of the stay, deviation from accustomed physical parameters (temperature, humidity, air quality and pressure, gravity, noise, crowding, comfort, the availability of food and water, and shelter), remoteness from home, the possibility of communication with one’s family and the parent organization, medical care, and rescue if needed. Whether the group is caught up in an extreme environment unwillingly, by accident or superior force, or has volunteered, and its motives for volunteering (excitement, prestige, patriotism, curiosity, money, altruism) are also significant. Last, it is useful to distinguish between organized

groups (teams, crews, units) and random conglomerations of individuals who happen to be thrown together in an unusual environment.

Many issues apply only to a subset of EUE situations; for example, polar, space, undersea, and similar missions require a relatively small group to be chosen from a larger number of applicants. This is obviously less true of the recruitment of combat soldiers, and not at all of groups caught up in accident or disaster. Nevertheless, there are concerns that are common to many kinds of groups, in a range of extreme environments. It is these general aspects of group functioning in EUEs that will be covered in the subsequent examination of the existing literature.

2. Cohesiveness

One striking pattern is the deep interpersonal bonding that many groups experience, existing alongside other groups' very high levels of friction and hostility. In extreme situations, self-sacrifice is not unusual, from Captain Lawrence Oates of Scott's South Pole team who, when supplies were running too low to sustain the whole group, left their shelter and went off to certain death in a blizzard to the real-life hero of the World War II ballad, Rodger Young, who 'fought and died for the men he marched among.' Studies of combat soldiers show that their primary motivation for risking their own safety is identification with their leader and unit, and comradeship with their fellows. Soldiers who have not developed such ties—for example, recent replacements—are much more likely to fail the test of battle (Manning 1991). Even in less organized groups, such as civilians in cities undergoing aerial bombardment or a natural disaster, high levels of mutual helpfulness and feelings of solidarity are commonplace.

In capsule EUEs, being confined in a small space with the same people for a long time, under conditions of danger, discomfort, fatigue, lack of privacy, and other stressors, is a potential source of hostility. Admiral Byrd (1938), among others, reflected on the time when a colleague's 'pet ideas become a meaningless drool, and the way he blows out a pressure lamp or drops his boots on the floor or eats his food becomes a rasping annoyance.' Intimate disclosures, encouraged earlier by loneliness, proximity, and boredom, in retrospect seem embarrassing and regrettable. Violations of personal space or possessions seem contemptuous provocations, irritating habits are perceived as deliberate affronts, and minor disagreements can become infuriating. The results range from vehement arguments through ostracism and physical attacks, to murderous assaults (Suedfeld and Steel 2000).

One of the most pervasive findings in isolated groups is that of dislike and disrespect between subgroups, such as polar scientists and support personnel or people from different cultural or national backgrounds

(Kanas 1998). This is a growing problem, as many projects involving EUEs are becoming international and/or multicultural.

3. Workload

One important aspect of EUE groups is that they are often overworked as well as stressed. In many extreme environments, living and working space is limited, facilities are functional and Spartan, and duties are both important and demanding. Survival as well as mission success may depend on everyone's performance, so that there are superordinate goals—important to all, and unreachable without total effort and cooperation. Social loafing or shirking is, while certainly not unknown, relatively rare and easily detected. In fact, studies have shown that members of submarine, space, and polar crews spend much of their supposed recreational or resting time working, sometimes leading to severe fatigue. In other situations, such as disaster rescue and combat, exhaustion may be the result of extreme physical exertion and danger. Insomnia, which is endemic in many EUEs, exacerbates the problem, as do physical discomforts including noise and duties that interrupt sleep (Stuster 1996).

Even the most dedicated crews may come to resent the demands made on them by the situation or by the organization to which they belong. There have been minirebellions or simple refusals to comply with work schedules by groups in both the American and Soviet space programs, as well as in many other EUE settings (Harrison et al. 1991). Typically, these have been brief and moderate, consisting of a self-declared and temporary 'time out,' although in some more extreme cases in wartime and extended sailing voyages they became full-blown mutinies.

4. Boredom

There is a quite opposite situation as well: in many special environments, there may be long stretches of unstructured time when there is little if any important work to be done. This has been true of groups of prisoners of war, ships' crews on multiyear voyages or frozen into the polar ice, personnel in remote stations, civilians in air raid shelters and refugee camps, troops between active duty deployments, resource extraction workers waiting for equipment repairs, resupplies, or further exploration, emergency workers (firefighters, paramedics) between calls, and so on. Spacefarers spending a year or more en route to their destination and then again on the way back will experience this (in contrast to astronauts in short-term, overprogrammed missions such as those of the Space Shuttle) (Stuster 1996).

Under physical and social monotony, free time becomes a time of boredom, lassitude, and apathy.

The desire to explore and manipulate—play with—elements of the environment becomes important. The group may manufacture artificial challenges, sometimes taking unnecessary risks either unthinkingly or deliberately. As individuals lose interest and motivation, they may neglect personal hygiene and withdraw from the group. As a result, and because hostile exchanges at least break the monotony, interpersonal arguments and conflict may intensify.

Alert leaders of organized groups make sure that no one has too much free time: initiating systematic recreational activities, getting people to put in extra work on cleaning and maintenance, and introducing novel work assignments such as cross-training crew members. In Arctic whaling and exploring ships of the eighteenth and nineteenth centuries, sailors and officers put on elaborate theatrical and musical performances, making their own scenery and costumes; competed in athletic events; and published shipboard 'newspapers' filled mostly with stories and poetry, since there was very little real news (Johnson and Suedfeld 1996). In the ghettos of Nazi-dominated Europe, Jewish authorities arranged elaborate cultural, scientific, religious, and educational programs. Even in lifeboats drifting after shipwreck, officers have organized duty watches, group sing-songs, story-telling, and the like. In all of these environments, such strategies help to maintain morale, ensure that necessary work gets done, and divert group members from their danger, deprivation, and fear.

5. Group Culture

Group culture represents an identity, made up of the characteristics of the members and what emerges from their interactions, which identifies the in-group as distinct from all others. This phenomenon is particularly marked in EUE groups: nicknames, jargon, practical jokes, initiation rites, and so on, weld the group together and simultaneously express and improve its morale. So do adornments such as bizarre hair-styles or beards, and special symbols such as badges, patches, pins, special articles of clothing, and other unique identifiers.

This pattern of values, activities, norms, jokes, and general lifestyle characteristics becomes particularly meaningful to isolated groups and frequently incomprehensible to outsiders. Such groups have an awareness of shared hardships and triumphs, and feel removed from the outside. As an example, a group of aquanauts living in an undersea environment came to refer to the rest of humanity as 'Earth people' (Radloff and Helmreich 1968), and of course the designation 'the world' is widely used among polar personnel and foreign-service soldiers when referring to home. The use of in-group jargon is also a way to deliberately stigmatize newcomers and outsiders (Taylor 1987).

Rituals, special foods, and the celebration of major

occasions are important aspects of group culture. For centuries, these have been used to strengthen a feeling of solidarity and continuity, and to exercise everyone's ingenuity, have fun, and reduce boredom and they continue to do so in the Space Age.

6. Leadership

Leaders of EUE groups have a role that is in many ways more crucial to the group's success than those normally encountered in most modern organizations. They may be cut off from the parent organization, so that they can not receive advice or instructions from their superiors; even when communication is possible, it may be long delayed. In any case, because of the unique circumstances, supervisors or consultants 'back home' may not fully understand all of the relevant factors—or the group may think that they do not, and therefore may be reluctant to follow instructions or advice (Connors et al. 1985). In such cases, the leader on the spot must decide how to handle the uncertainty and conflict.

The leader's formal rank confers the power of reward and punishment; but the impact of these can be attenuated in long-duration remote work, where any material reinforcers may be trivial or far in the future, and in extremely dangerous environments where any 'punishment' may in fact increase the transgressor's safety (e.g., a court-martial that would involve a transfer to the rear echelon). Even worse, the leader may come to be seen as dictatorial, petty, and unfair, with consequences that can range from the unpleasant to the disastrous. The best EUE leaders are those who use formal authority sparingly but are admired, perceived as holding authority legitimately, seen as a role model to be emulated, and believed to be expert in solving the problems that the group faces (Stuster 1996).

Good interpersonal relations are made even more important by the fact that in EUEs the trappings that normally keep leaders remote and aloof—secretaries and aides, imposing offices, separate facilities, and so on—are drastically attenuated (Kanas 1998). Social leveling occurs even in traditionally hierarchical groups, such as military units. The leader is revealed as having human foibles and faults. If the crew dislikes or disrespects the leader, it has many opportunities to express its feelings. Conflict between the leader and a key subordinate or a clique of group members, is common under stress and requires wise strategies balancing tact and forcefulness. Excessive consciousness of rank differences may also lead to disaster for the group: social psychologists have identified a number of airline crashes that could have been avoided if a crew member who noticed that something was wrong had dared to assertively interrupt the captain while the latter was paying attention to something else (Helmreich 1984).

How the leader should function varies with the circumstances. EUE groups have indicated that on matters of a technical nature, like navigation or equipment repair, the leader is expected to make decisions in consultation with the relevant technical experts; on issues of routine policy, like work assignments and recreational schedules, the leader should consult with the entire group in an open way, even if the group's preferences are not followed exactly; and in emergencies, the leader's clear, unhesitating, and unilateral decision is appreciated even by those who otherwise resent direction (Stuster 1996, Suedfeld and Steel 2000).

The difficulties related to EUEs are especially severe for the leader: resolving other people's environmentally caused stresses while one is experiencing the same ones is a true challenge to leadership. One development that arises only too frequently is the leader's becoming a source of stress for the rest of the group. In an EUE, attributes that can be tolerated in a normal environment become unbearable. Some leaders are too unstable, demanding, threatening, perfectionistic, or inconsiderate—traits that may be part of their personality, or that emerge because they too are under stress. These traits lead to decreased leader influence, and to increased resistance on the part of other group members. In long-duration special environments, it is probably not a good idea to rely too much on a crew's ability to tolerate the commander (Kanas 1998).

7. Sex and Gender Issues

The presence of growing numbers of women in formerly all-male EUEs has affected morale and performance, and raises some interesting issues. Almost all Western military forces, national Antarctic programs, and space programs have by now included female members and some leaders. These developments have changed group dynamics, in as yet undetermined ways. If a recent all-female Antarctic traverse expedition is typical, a shift toward more emphasis on group cooperation and solidarity may be occurring, perhaps with some decrease in goal orientation, and it certainly seems that women in mostly male groups both provide and are offered more interpersonal support and nurturance than has been usual (Rothblum et al. 1998).

There is a lively controversy among military and polar officials as to whether sexual integration has resulted in additional or reduced sexual tension, and as to just how pair relationships affect the participants and the rest of the group. Some people are less sexually frustrated, but others are not; resentment, jealousy, and envy have surfaced. In at least one Antarctic case, there was a serious drop in station morale when the cook's romantic partner abandoned him and began a relationship with someone else, and meals became unappetizing. One can only speculate what would

happen on a long-duration EUE mission (or in an emergency) if, for example, the group leader were involved in this kind of triangle (Suedfeld and Steel 2000).

8. Selection and Training

This topic applies to only a subset of special environments, but is important in those. There are two phases of selection: selecting out (eliminating applicants below the threshold) and selecting in (picking the best possible crews). The former is done on an individual basis, using biodata, interviews, and tests; in the second stage, both individual and group selection become relevant. The interaction and performance of potential crewmates are observed in analogue or simulation situations (field and laboratory settings respectively, both incorporating some aspects similar to the eventual EUE). For example, candidates may go through a wilderness survival course or (in the ex-Soviet space program) undertake a long cross-country journey in a small automobile. Many programs choose and train crews as a unit; if for some reason one member must be struck from a mission, the entire crew is replaced (Santy 1994).

9. Future Directions

The number of people exposed to EUEs, fortuitously or deliberately, is not likely to diminish in the near future; and as the sheer technical and physical difficulties are solved, more attention will be paid to psychosocial issues. An emphasis on understanding group reactions, not only individual ones, is emerging. As the connections between EUEs and 'general' social science areas become clearer, research will become less atheoretical and applied; EUEs provide unique opportunities for both gathering new knowledge and having an impact on serious real-life problems, which may tempt more social scientists into the field (both literally and metaphorically).

See also: Culture and Emotion; Group Decision Making, Cognitive Psychology of; Group Decision Making, Social Psychology of; Group Processes, Social Psychology of; Group Productivity, Social Psychology of; Groups, Sociology of; Leadership in Organizations, Psychology of; Military and Disaster Psychiatry; Military Psychology: United States; Violence and Effects on Children

Bibliography

- Byrd R E 1938 *Alone*. Putnam's, New York
 Connors M M, Harrison A A, Akins F R 1985 *Living Aloft: Human Requirements for Extended Spaceflight*. NASA, Washington, DC

- Harrison A A, Clearwater Y A, McKay C P (eds.) 1991 *From Antarctica to Outer Space: Life in Isolation and Confinement*. Springer-Verlag, New York
- Helmreich R L 1984 Cockpit management attitudes. *Human Factors* 26: 583–89
- Holland A W, Curtis K 1998 Operational psychology countermeasures during the Lunar–Mars life support test project. *Life Support and Biosphere Science* 5: 445–52
- Johnson P J, Suedfeld P 1996 Coping with stress through creating microcosms of home and family among Arctic whalers and explorers. *History of the Family* 1: 41–62
- Kanas N 1998 Psychosocial issues affecting crews during long-duration international space missions. *Acta Astronautica* 42: 339–61
- Manning F J 1991 Morale, cohesion, and esprit de corps. In: Gal R, Mangelsdorff A D (eds.) *Handbook of Military Psychology*. Wiley, New York
- Radloff R W, Helmreich R L 1968 *Groups Under Stress: Psychological Research in SEALAB II*. Appleton-Century-Crofts, New York
- Rothblum E D, Weinstock J S, Morris J F (eds.) 1998 *Women in the Antarctic*. Harrington Park Press, Binghamton, NY
- Santy P A 1994 *Choosing the Right Stuff: The Psychological Selection of Astronauts and Cosmonauts*. Praeger, London
- Stuster J 1996 *Bold Endeavors: Lessons from Space and Polar Exploration*. Naval Institute Press, Annapolis, MD
- Suedfeld P, Steel G D 2000 *The Environmental Psychology of Capsule Habitats*. *Annual Review of Psychology*. Annual Reviews Press, Chippewa Falls, WI
- Taylor A J W 1987 *Antarctic Psychology*. DSIR, Wellington, NZ
- Wilson J P, Harel Z, Kahana B (eds.) 1988 *Human Adaptation to Extreme Stress: From the Holocaust to Vietnam*. Plenum, New York

P. Suedfeld

Groups, Sociology of

A group is a bounded collection of interacting individuals or a consensual social categorization. It is difficult to distinguish sociological and psychological studies of groups and, in the past, there has been a

considerable overlap. Still, one feature may help to make a reasonably accurate distinction. The sociological study of groups always involves attention to collective phenomena in addition to the concept of the group itself. The psychological study of groups sometimes does, but often does not, involve attention to these phenomena. In the work of sociologists, processes of interaction and of individual behavior in social contexts are seen as giving rise to durable or semidurable collective phenomena (such as cohesion; relational structures pertaining to status, communication, friendship, power, control; structural similarity; distribution of information; collective goods provision; social norms; collective representations; styles of conflict resolution; patterns of group decision-making).

These small-scale collective phenomena, in turn, influence the behavior of individuals and their interactions (see Fig. 1). Some sociological work concentrates on relation 1, some on relation 2, and some on 1 and 2, and other combinations. Relation 3b allows explicit feedback mechanisms involving both levels (as in expectation states theory, see below). Psychological studies often focus on the direct influence of interaction (or the presence of others) on attitude change, productivity, satisfaction, self-esteem, depression, etc., occasionally allowing for feedback mechanisms (3a and 3b). The extra arrows in Fig. 1(i) indicate exogeneous factors (such as genetic dispositions, rules, established categories, status differences).

1. Historical Background of Small Group Research

The scientifically serious investigation of groups in the sociological tradition started with anthropological studies of tribes, clans, and families in the nineteenth and early twentieth centuries. Sumner, Simmel, and Cooley were early codifiers of that kind of research. The most important turn in this development occurred

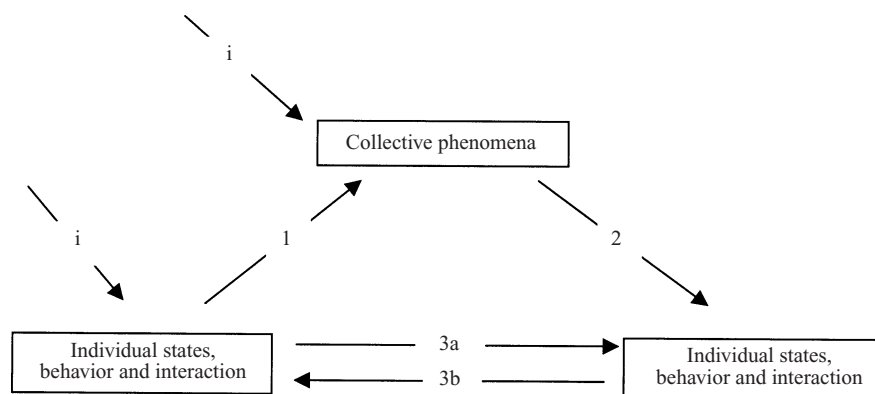


Figure 1
Schematic depiction of the explanation of groups

in the 1930s. Within the span of a few years, four great books appeared which had two very important messages in common: that groups are constituted by a cognitively mediated functional interdependence between individuals, i.e., people who are dependent on each other for achieving a goal interact, and something important happens to their minds and the way their minds operate when they interact. In addition, the actual situation in which their action is placed matters most for behavior (as opposed to traits or other cross-situational constants). In 1934 Mead's (then already posthumous) *Mind, Self and Society* appeared. Two years later, Lewin's *Principles of Topological Psychology* and Sherif's *The Psychology of Social Norms* came out. A year later, Parsons' *The Structure of Social Action* made its appearance. Merton's early work on reference groups was also quite influential.

From this time, group studies continued in two traditions—experimental and field studies—and both traditions bore the mark of this convergent conception of groups. Under the leadership of Lewin and Sherif, 'group dynamics' as the experimental study of groups was then called, achieved great heights. Field studies had Mayo's Hawthorne studies and Whyte's famous *Street Corner Society* as models. Both traditions owe their success to the simultaneous consideration of the three fundamental interdependencies in groups: *functional* (related to goal achievement), *cognitive* (related to common categorizations, beliefs, and knowledge) and *structural* (related to the configurations of social relations) (see Lindenberg 1997 for more detail on this and other aspects of this article).

However, the precise nature of the interdependencies was not yet well investigated and in the late 1950s and early 1960s, specialized groups of studies on each of these interdependencies became dominant, with interesting crossovers between psychology and sociology. Homans (1950, 1958), a sociologist, combined the field and the experimental traditions in the study of functional interdependencies in exchange relations. Inspired by Homans, two psychologists, Thibaut and Kelley (1959), worked out the study of small groups as the study of functional interdependencies. Their book is to this day one of the best on the subject. It influenced Homans' further work on the subject (Homans 1961). Cognitive interdependencies were elaborated by Sherif and turned into a major strand of group research called 'social identity theory' by Tajfel (1982) and Turner (1982), both psychologists but very influential among sociologists and in turn influenced by sociological symbolic interactionism. In a related development, Moscovici, a psychologist, elaborated Durkheim's theory of collective representations (see Flick 1998). Structural interdependencies were more formally elaborated by the psychologist Cartwright and the mathematician Harary and turned into a full-blown network approach by sociologist White and his followers (see White et al. 1976).

2. Bringing Different Kinds of Interdependencies Back Together Again

This section deals briefly with the field study approach and in some more detail with experimental studies. No attempt is made to be complete. Rather, the overview is selective. The selection criteria are first, that the studies should be substantively important. Second, they must mainly focus on group processes comprising small groups (i.e. face-to-face interaction with at least one common goal), small networks, and face-to-face encounters. Third, the studies must be seriously concerned with the integration of at least two of the three fundamental interdependencies. Fourth, they must cluster around a program of investigation with some continuity.

With regard to field studies, work inspired by Goffman was chosen. Goffman's studies provided a new integration for field studies in sociology (excluding Mayo-type studies). With regard to experimental (and simulation) studies, expectation states theories, various groupings of social exchange theories, and studies of friendship and cognitive networks were selected. Studies of social dilemmas are covered elsewhere (see *Cooperation: Sociological Aspects*). Helpful discussions of a broader scope can be found in Cook et al. (1995) and Foschi and Lawler (1994). Due to space restrictions, the reader will at times be referred to bibliographies and collections rather than individual works.

2.1 Fieldwork Studies

From the beginning of the 1960s, fieldwork studies on face-to-face groups have received their main theoretical guidance from Goffman (1961). He also put short-lived groups, or 'encounters', on the map. There is a relatively small but very productive circle of contributors to this kind of research (a few important examples are Anderson 1978, Cahill and Lofland 1994, Fine 1987, MacLeod 1995). These studies are concerned with conformity, status, group divisions, identities, and their combination in impression management. They are rich in descriptive detail. Aspects of functional and cognitive interdependencies are clearly present in Goffman's work.

2.2 Experimental Studies

2.2.1 Expectation states theories. Generic expectation states theory is concerned with the formation and maintenance of power and prestige structures in face-to-face task groups (see literature in Berger and Zelditch 1993). Following the tradition of Bales in the 1950s, power and prestige are conflated and their structure is conceived in terms of behavioral aspects

(opportunities and influence) rather than the distribution of scarce resources. People are higher on the power and prestige structure, the more opportunities to act they receive, the more often they take advantage of these opportunities, and the more often they win disagreements if and when they occur.

The basic mechanism that creates this structure is pairwise interaction (between, say, Ego and Alter) that is governed by performance expectations. It is the (unobservable) structure of these performance expectations which keeps generating the observable power and prestige structure. When Ego has positive performance expectations towards Alter, he or she will give more action opportunities to Alter and Alter will more likely take up each opportunity offered. Alter has influence over Ego. It is assumed that there is an evaluation bias. Ego is more likely to judge Alter's performance to be positive if the performance expectation was positive and vice versa. And if Ego nonetheless voices disagreement, he or she is more likely to eventually yield to Alter if the performance expectation was positive to begin with. Such evaluation biases greatly enhance the stability of the power and prestige structure (provided the task endures).

The theory is meant to apply to task groups. For this reason, functional interdependence is by definition the starting point. Here, a joint task and face-to-face interaction is what makes a collection of individuals into a group. Each group member's contribution influences a group score (mostly outcome interdependence, say, linking bonus to agreement in answers). There are also cognitive aspects. For the theory to apply, group task achievement must be a salient goal in the interaction with other members ('collective orientation'). The central cognitive element in the theory is a performance expectation. It presupposes other cognitive elements, such as goal criteria and causal beliefs about how likely a particular characteristic or behavior is to lead to success or to failure. Cognitive interdependence comes about when the performance expectations (and, prior to that, the goal criteria and causal beliefs) are being shared. Given the evaluation bias, the interdependence means that performance expectations of different group members interlock (see Moore 1985).

Functional and cognitive interdependencies are closely interlinked in this theory. This link is made even stronger through extensions of the theory to include external status characteristics (see Berger and Zelditch 1993) and the generation of status characteristics from nominal distinctions (see Webster and Hysom 1998). Task groups are embedded in a wider culture in which such status and competence beliefs exist (such as 'men are better at mathematics'). People seem to follow the maxim 'when in doubt, assume that a salient characteristic is task-relevant' (the so-called 'burden of proof principle'). With regard to gender, this favors men in general tasks as well as in 'masculine' tasks, and it favors women only in stereotypically

'feminine' tasks. Through the link of performance expectations with worthiness, there is also a link of expectation structures to conceptions of just rewards, which is worked out in the 'status value theory' of distributive justice (see Berger and Zelditch 1993).

Ridgeway has shown that, in turn, such cultural beliefs can be explained by the diffusion of *status-belief-generating mechanisms* in small groups. When interaction partners in task groups are doubly dissimilar in that they differ in valuable resources (say high pay level) and in a salient nominal distinction (such as race or gender), then a process of misattribution will causally link having the valuable resource with the nominal distinction and eventually with assumptions of competence, worthiness, and esteem. The group consensus on this attribution generalizes to the idea that 'most people' believe it to be so (Ridgeway et al. 1998). Group members are likely to transfer such beliefs to other group contexts, thereby spreading and reinforcing them across groups and turning them into cultural status beliefs. Similar mechanisms are now invoked to explain legitimation and delegitimation of power and prestige structures in groups (see Berger et al. 1998).

2.2.2 Theories of exchange in networks. Theories of social exchange by Blau (1964), Homans (1950), and Thibaut and Kelley (1959) have given rise to a number of theories on the influence of social structure on the distribution of exchange ratios or benefits within groups. The results of these theories are also relevant to the explanation of network effects on power advantages on a macro level (see, e.g., Burt 1992). Paradigmatic for these theories is Emerson's (1976) concept of 'network connection,' which points to externalities between exchange relations. Two exchange relations are connected if and only if the exchange in one relation affects the exchange in the other (in terms of frequency and value). Thus, in an exchange network, the specific distribution of ties can have a strong impact on the distribution of benefits, but this influence depends on the nature of the network connections. For example, if there is a rule that one may exchange only with one other partner then the network is negatively connected. Division of labor creates positively connected networks in which exchange with one stimulates exchange with the other(s). Negative connections are easiest to trace both theoretically and empirically, and thus the bulk of research in this area is on negatively connected networks (an important exception is Burke (1997), who has developed an interesting link between network exchange and identities).

One prominent version of such a theory is *exchange network theory* (Cook and Yamagishi 1992). The distribution of benefits (or rates of exchange) within a network is explained on the basis of power ('power/dependence'). In positively connected networks, the power process is determined by brokerage, i.e., by the

ability to connect resources of various actors, and by the value of the focal resource to the actors. In a negatively connected network, power is based on the value and degree of availability of a focal resource from alternative sources. The more value Ego gets from an exchange with Alter compared to the best available exchange alternative, the more dependent Ego is on Alter and the more power Alter has over Ego. With repeated exchange in such negatively connected networks, the exchange ratio is predicted to shift toward the point at which neither of them could improve his or her outcome by exchange with another partner. Power advantages are thus structurally induced, given the 'connectedness' of the network.

A variant of this theory for coercive power has been elaborated by Molm (1997). Interestingly, Molm predicts and finds that power based on the ability to inflict damage on others is not structurally but strategically induced in order to extract rewards from the other. This opens exchange theory to a more explicit consideration of rational strategies, especially with regard to the balance between incentive to coerce and fear of retaliation. Lawler's work is also relevant to this topic (see, e.g., Lawler 1992a).

Elementary theory (ET, sometimes also called 'network exchange theory') is also rooted in the Homans/Emerson tradition and it also explains exchange ratios on the basis of structurally induced power (see Berger and Zelditch 1993, pp. 323). The major difference with exchange network theory is twofold. First, in reference to Weber, ET considers resistance to the exercise of power more explicitly than exchange network theory. Second, ET uses experiments in which the degree of 'exclusion' (i.e., the degree of negative connectiveness) varies. This second possibility gives it more flexibility to approximate various real life phenomena. However, the way resistance and the representation of exchange are worked out imposes strong restrictions, which reduces real life application to some degree (see Van Assen 2001). More recently, interesting links of ET with expectation states theory (Willer et al. 1997) and with game theory (see Willer and Skvoretz 1997) have been developed.

An additional grouping of theories on exchange should be mentioned. It is also related to the Homans/Emerson tradition but uses different elaborations. There have been long-standing efforts to explain structural network effects on power and exchange rates by means of game theory (especially using the concept of the 'core') by Bonacich and co-workers (see, e.g., Bonacich 1998). See Friedkin (1995) for yet a different model and Fararo and Skvoretz (see Berger and Zelditch 1993, pp. 416) for a comparison of various models. There is also a productive development of models that use exchange in networks for the explanation of power processes in collective decision making, mostly in small groups (see Stokman et al. 2000). Because these models are not tied to a particular experimental paradigm, they are more open

to the consideration of other exchange models and strategies (such as Coleman's model; see Coleman 1990) and to an explicit consideration of strategies of persuasion and challenges that accompany exchange strategies.

2.2.3 Friendship and cognitive networks. In recent years, dynamic friendship networks have received an increasing amount of attention. A new line of research has opened up by taking relatively simple behavioral assumptions and tracing their structure-generating consequences through simulation and, at times, also empirical research. What is new is that the interdependence of simultaneous friendship choices is modeled. If choices go on simultaneously in a group of strangers, what structure will emerge from the interdependence of choices under various choice preferences and constraints? How does 'group' size affect the resulting structure(s)? How does heterogeneity of the choosing individuals affect the resulting structure(s)? Preferences have been made more complex by stepwise adding status preferences to similarity preferences, and by explicitly introducing friendship dissolution to friendship formation. As to influence, friends do not just have a preference for similarity, they also become more similar. Paradigmatic for this kind of work is Stokman and Zeggelink (1996) and work presented in Doreian and Stokman (1997). See Van der Bunt (1999) for empirical research. Work linking this kind of dynamic approach to issues of cooperation is exemplified by Flache and Macy (1996) and Flache and Hegselmann (1999).

Networks are also cognitively represented which, in turn, contributes to the generation of joint cognitive and social structures. Research dealing with these aspects of networks link explicitly cognitive and structural interdependencies (see Carley 1986, Freeman 1994, Krackhardt 1987, Pattison 1994). Related research deals with the effects of cognitive interdependencies on learning, search, and general problem-solving strategies in groups and organizations (see Carley 1995, Edmonds and Dautenhahn 1999). Here functional interdependencies in terms of joint production are added explicitly (see also Troyer and Silver 1999). In the future, one can expect innovative impulses from these lines of research for sociological group research in general.

3. Evaluation

Sociological research on groups is lively, productive and, in the experimental fields, also cumulative. There is also a slow tendency to open up across the narrow borders of the research paradigms. However, there are a number of important shortcomings (see Lindenberg 1997 for more detail and for suggestions for new directions). First, field and experimental studies generally ignore each other. This is not a

desirable state. Crossovers in the past have been very productive and should be encouraged. Second, the experimental paradigms carry such a strong weight that they hamper new theory formation. For example, in expectation states theories, the functional interdependence is experimentally restricted to outcome interdependence with regard to agreement. There is little or no task interdependence. In the exchange theories, exchanges are often limited to divisions of resource pools. Such exchanges lack many of the properties of the exchange of goods, information, and emotions. On the whole, there is little attention to the effects of joint production on expectations and exchange.

Third, functional interdependencies should have analytical primacy. This primacy derives from the fact that cognitive processes are strongly influenced by goals (see Gollwitzer and Moskowitz 1996) and structural effects only occur in interaction with goals (externalities) as has been seen in exchange network theory. Yet, in most exchange theories this primacy is lacking. Fourth, present research rarely includes 'groupishness' (i.e., categorization and framing effects) in its consideration of cognitive interdependencies (see Lawler 1992b as an exception). For example, in expectation states theory, 'collective orientation' is introduced as a scope condition, not as a variable. Related to this point is, fifth, that the distinction between small groups and small networks is non-existent or ad hoc. There is little or no theory on the influence of groups on networks or vice versa, or on networks nested in groups. Sixth, the study of inter-group relations for the explanation of what goes on inside the group needs much more attention. A positive example of what is needed is Coleman's classic study of the influence of interscholastic athletic competition on the local status structure in schools (see Coleman 1961).

See also: Collective Behavior, Sociology of; Exchange: Social; Networks: Social; Status and Role, Social Psychology of; Status and Role: Structural Aspects

Bibliography

- Anderson E 1978 *A Place on the Corner*. University of Chicago Press, Chicago
- Berger J, Ridgeway C L, Fisek M H, Norman R 1998 The legitimation and delegitimation of power and prestige orders. *American Sociological Review* **63**: 379–405
- Berger J, Zelditch M (eds.) 1985 *Status, Rewards, and Influence*. Jossey-Bass, San Francisco
- Berger J, Zelditch M (eds.) 1993 *Theoretical Research Programs*. Stanford University Press, Stanford, CA
- Blau P 1964 *Exchange and Power in Social Life*. Wiley, New York
- Bonacich P 1998 A behavioral foundation for a structural theory of power in exchange networks *Social Psychology Quarterly* **61**: 185–98

- Burke P J 1997 An identity model for network exchange. *American Sociological Review* **62**: 134–60
- Burt R S 1992 *Structural Holes. The Social Structure of Competition*. Harvard University Press, Cambridge, MA
- Cahill S E, Lofland H (eds.) 1994 *The Community of Streets*. JAI Press, Greenwich, CT
- Carley K 1986 An approach for relating social structure to cognitive structure. *Journal of Mathematical Sociology* **12**: 137–89
- Carley K 1995 Computational and mathematical organization theory: Perspectives and directions. *Computational and Mathematical Organization Theory* **1**: 39–56
- Coleman J S 1961 *The Adolescent Society*. Free Press, Glencoe, IL
- Coleman J S 1990 *Foundation of Social Theory*. Harvard University Press, Cambridge, MA
- Cook K S, Fine G A, House J S (eds.) 1995 *Sociological Perspectives on Social Psychology*. Allyn and Bacon, Boston
- Cook K S, Yamagishi T 1992 Power in exchange networks: A power-dependence formulation. *Social Networks* **14**: 245–66
- Doreian P, Stokman F N 1997 *Evolution of Social Networks*. Gordon and Breach, Amsterdam
- Edmonds B, Dautenhahn K (eds.) 1999 Special issue on social intelligence. *Computational and Mathematical Organization Theory* **5**
- Emerson R M 1976 Social exchange theory. *Annual Review of Sociology* **2**: 335–62
- Fine G A 1987 *With the Boys*. University of Chicago Press, Chicago
- Flache A, Hegselmann R 1999 Rationality vs. learning in the evolution of solidarity networks. *Computational and Mathematical Organization Theory* **5**: 97–127
- Flache A, Macy M 1997 The weakness of strong ties: Collective action failure in a highly cohesive group. *Journal of Mathematical Sociology* **21**: 3–28
- Flick U (ed.) 1998 *The Psychology of the Social*. Cambridge University Press, Cambridge, UK
- Foschi M, Lawler E J (eds.) 1994 *Group Processes: Sociological Analyses*. Nelson-Hall, Chicago
- Freeman L C 1994 Interpersonal proximity in social and cognitive space. *Social Cognition* **12**(3): 223–47
- Friedkin N E 1995 The incidence of exchange networks. *Social Psychology Quarterly* **58**: 213–22
- Goffman I 1961 *Encounters*. Bobbs-Merrill, Indianapolis, IN
- Gollwitzer P M, Moskowitz G B 1996 Goal effects on action and cognition. In: Higgins E T, Kruglanski A W (eds.) *Social Psychology. Handbook of Basic Principles*. The Guilford Press, London, pp. 361–99
- Homans G C 1950 *The Human Group*. Harcourt, Brace & World, New York
- Homans G C 1958 Social behavior as exchange. *American Journal of Sociology* **63**: 597–606
- Homans G C 1961 *Social Behavior: Its Elementary Forms*. Harcourt, Brace & World, New York
- Krackhardt D 1987 Cognitive social structures. *Social Networks* **9**: 109–34
- Lawler E J 1992a Power processes in bargaining. *The Sociological Quarterly* **33**: 17–34
- Lawler E J 1992b Affective attachments to nested groups. *American Sociological Review* **57**: 327–39
- Lindenberg S 1997 Grounding groups in theory: Functional, cognitive, and structural interdependencies. *Advances in Group Processes* **14**: 281–331

- MacLeod J 1995 *Ain't No Making It*. Westview, Boulder, CO
- Molm L 1997 *Coercive Power in Social Exchange*. Cambridge University Press, Cambridge, UK
- Moore J C 1985 Role enactment and self-identity. In: Berger J, Zelditch M (eds.) *Status, Rewards, and Influence*. Jossey-Bass, San Francisco, pp. 262–316
- Pattison P 1994 Social cognition in context: Some applications of network analysis. In: Wasserman S, Galaskiewicz J (eds.) *Advances in Social Network Analysis*. Sage, Thousand Oaks, CA, pp. 79–109
- Ridgeway C L, Heger Boyle E, Kuipers K J, Robinson D T 1998 Resources and interaction in the development of status beliefs. *American Sociological Review* 63: 331–50
- Stokman F N, Van Assen M, Van Der Knoop J, Van Oosten R C H 2000 Strategic decision making. *Advances in Group Processes* 17: 131–53
- Stokman F N, Zeggelink E P H 1996 'Self-organizing' friendship networks. In: Liebrand W B G, Messick D M (eds.) *Frontiers of Social Dilemma Research*. Springer, New York, pp. 385–418
- Tajfel H 1982 Instrumentality, identity and social comparisons. In: Tajfel H (ed.) *Social Identity and Intergroup Relations*. Cambridge University Press, Cambridge, UK, pp. 483–507
- Thibaut J, Kelley H H 1959 *The Social Psychology of Groups*. Wiley, New York
- Troyer L, Silver S D 1999 Institutional logics and group environments. *Advances in Group Processes* 16: 219–52
- Turner J C 1982 Toward a cognitive redefinition of the social group. In: Tajfel H (ed.) *Social Identity and Intergroup Relations*. Cambridge University Press, Cambridge, UK, pp. 15–40
- Van Assen M 2001 Two representations of negotiated exchange: A review and comparison. In: Van Assen M (ed.) *Essays on Actor Models in Exchange Networks and Social Dilemmas*. Thela Thesis, Amsterdam
- Van der Bunt G 1999 *Friends by Choice: An Actor Oriented Statistical Network Model for Friendship Networks Through Time*. Thela Thesis, Amsterdam
- Webster M, Hysom S J 1998 Creating status characteristics. *American Sociological Review* 63: 351–78
- White H C, Boorman S A, Breiger R L 1976 Social structure from multiple networks: I. Blockmodels of roles and positions. *American Journal of Sociology* 81: 730–80
- Willer D, Lovaglia M J, Markovsky B 1997 Power and influence: A theoretical bridge. *Social Forces* 76: 571–603
- Willer D, Skvoretz J 1997 Games and structures. *Rationality and Society* 9: 5–35

S. Lindenberg

Growth Curve Analysis

The phrase 'growth curve analysis' denotes the processes of describing, testing hypotheses and making scientific inferences about the growth and change patterns in a wide range of time-related phenomena. The term 'growth curve' was used originally to describe a graphic display of physical stature (e.g., the height or weight) of an individual over consecutive ages. In contrast to some longitudinal data, growth curves have unique features: (a) the same entities are repeatedly observed, (b) the same procedures of measurement and scaling of observations are used,

and (c) the timing of the observations is known. Formal models for the analysis of growth curves were developed in many different substantive domains with a common goal—to examine and uncover a fundamental set of regularity conditions or 'basic functions' responsible for the manifest growth and change. Techniques for the analysis of growth curves were initiated in the physical sciences, more fully developed in the biological sciences, and then used in studies of the size and health of plants, animals, and humans. In the behavioral sciences, growth curve analyses have been applied routinely to a wide range of phenomena—from experimental learning curves, to the growth and decline of intellectual abilities and academic achievements, and to changes in other psychological traits over the life span.

Growth curve analyses are among the most widely studied and well-developed mathematical and statistical techniques in all scientific research. These analyses have roots in the seventeenth- and eighteenth-century calculus of Newton and probability of Pascal, but this article is limited to more recent historical developments and other related techniques, such as time-series and dynamical systems analyses are not covered.

1. Growth Curve Data Collections

The first reported growth curve data were collected by the French Count de Montbeillard between 1759 and 1777. These data consist of semi-annual measurements on the growth of the height of his son over nearly 18

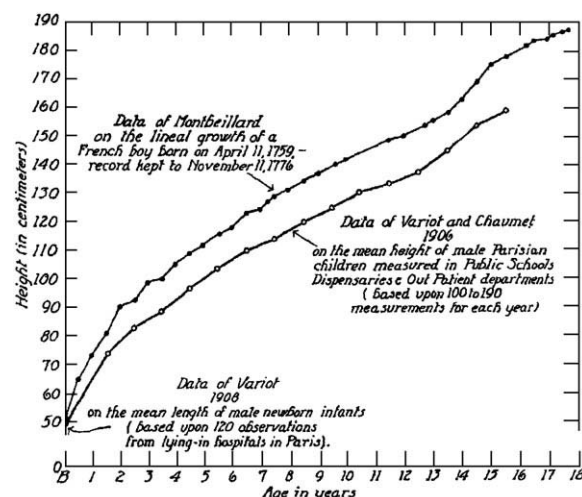


Figure 1

The initial 'growth curves' of human height data from Scammon (1927, p. 334). Note: The vertical (y) axis represents the height in centimeters and the horizontal (x) axis represents the age in years from birth

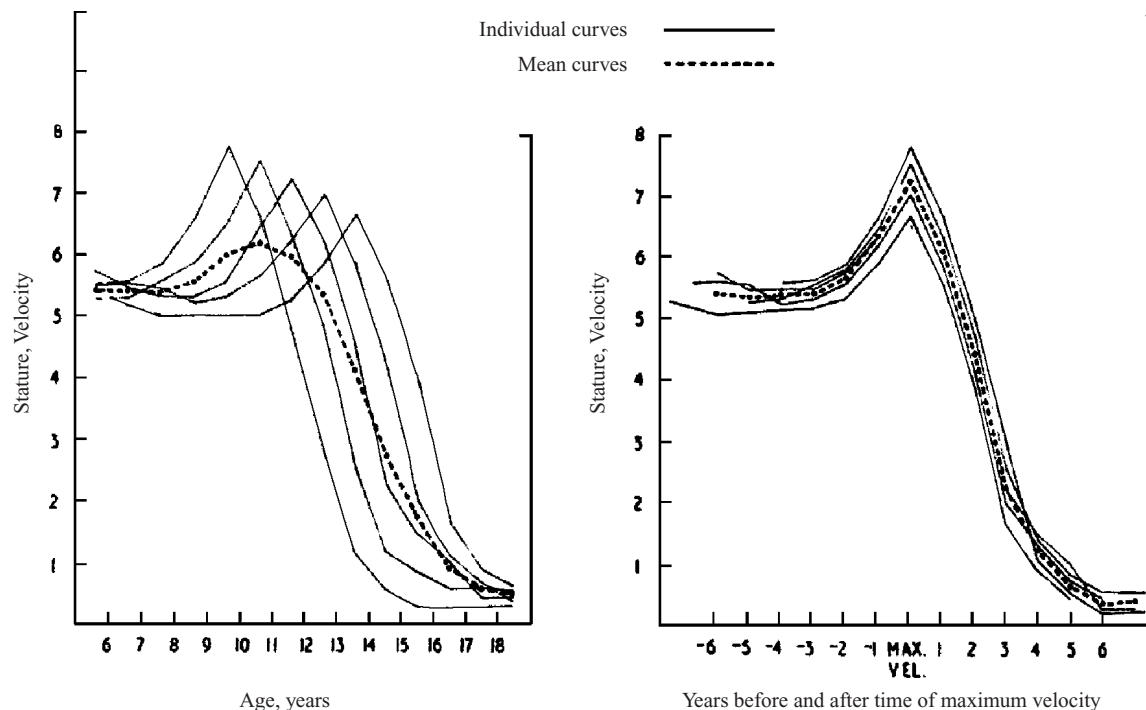


Figure 2

Alternative 'velocity curves' of physical stature from Tanner (1960, p. 22). Note: the left-hand graph shows the change in height as a function of the change in time (y) versus the age in years (x), while the right-hand graph shows the same curves (y) plotted around the time of maximum change in scores (x) for each individual

years. The data were plotted by Scammon (1927) and are presented here in the upper curve of Fig. 1. As Scammon reported

it will be noted that the curve shows the typical four phases which most modern students have observed in the postnatal growth in stature of man, and which are characteristic of the growth of so many parts of the body (p. 331).

The first analysis of these data was by the naturalist Buffon in 1799 and

should be given full credit for the discovery of seasonal differences in growth a full hundred years before the modern investigation of this work (p. 334).

The lower growth curve in Fig. 1 is based on group averages obtained by Variot in 1908 and

it is interesting to note that, while the absolute values of the two series are quite different, the general form of the curve is essentially the same in both instances (p. 335).

These early growth curves were the precursor to an enormous collection of biological data on growth and change. A more recent illustration comes from the important work of Tanner and his colleagues (see Tanner 1960). In the individual plots of height in Fig. 2: (a) the 'velocity' or rate-of-change is plotted at each

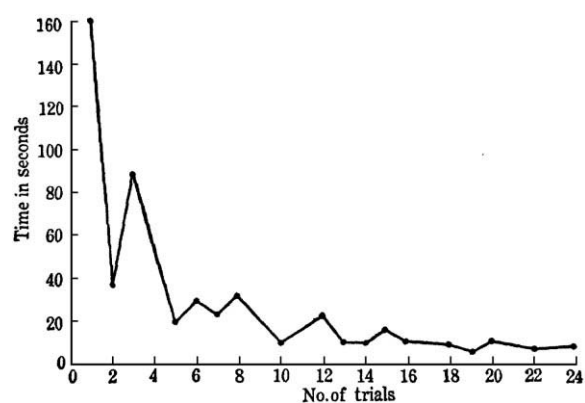


Figure 3

The 'learning curve' of a cat from Thorndike (1911) as reported by Garrett (1951, p. 44). Note: the response time (y) required for a cat to escape from a closed box by pulling on a wire as a function of consecutive number of trials (x)

age and (b) these curves are plotted against their own 'highest peak velocity.' This display demonstrated two interesting features of physical growth: (a) persons

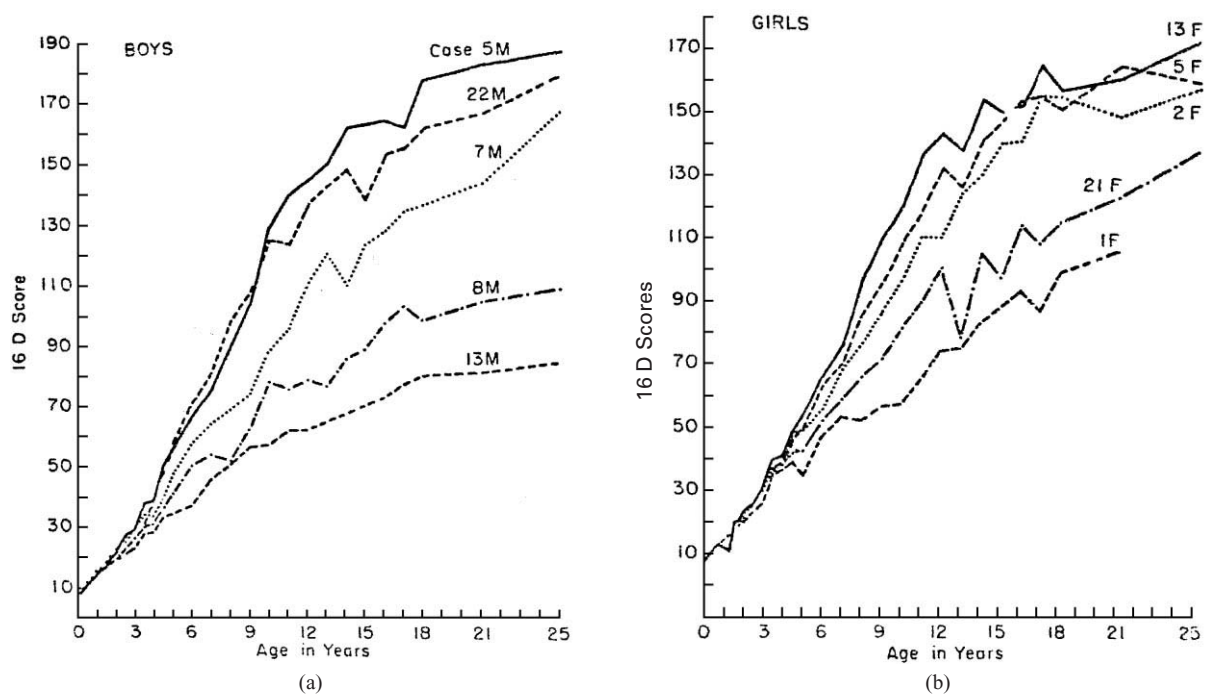


Figure 4

Growth curves of intellectual abilities in selected boys and girls from Bayley (1956, p. 67). Note: Age 16 D scores (y) plotted as a function of age at measurement (x) for (a) five boys and (b) five girls

who start growth at the earliest ages also grow the most but (b) all individuals share a remarkably similar shape in their 'adolescent growth spurt.' Extensions of these basic findings on physical growth systems and the contrast of chronological time vs. biological time, are still used in contemporary research.

Experimental psychologists have routinely collected different kinds of growth curves. Among the first here were the classical 'forgetting' curves collected by Ebbinghaus in the 1880s, whose work introduced the use of quantitative methods in the study of learning and memory and stimulated many experimental data collections. Included here are the 'curves of improvement' in studies on telegraphic language by Bryan and Harter in 1897. These principles were also used in the animal learning curve experiments of Thorndike in 1911, and one example of such data appears in Fig. 3. This plot illustrates an example of 'trial and error' learning: learning was objectively defined by decreasing response time and the lack of smooth function over trials was considered as error. Thorndike used these growth curves to illustrate several classical principles of learning, including the 'law of exercise' and the 'law of effect' (see Garrett 1951).

Differential psychologists have also contributed growth data in many different substantive areas. One good example of this tradition is given in the plots of Fig. 4 from work of Bayley (1956). Individual growth

curves of mental abilities from birth to age 25 are plotted for a selected set of boys and girls from the well-known Berkeley Growth Study. Because mental ability was not easily measured in exactly the same way at each age, these individual curves were created by adjusting the means and standard deviations of different mental ability tests (i.e., Stanford-Binet, Terman-Merrill) at different ages into a common metric. As Bayley says

they are not in 'absolute' units, but they do give a general picture of growth relative to the status of this group at 16 years ... These curves, too, are less regular than the height curves, but perhaps no less regular than the weight curves. One gets the impression both of differences in rates of maturing and of differences in inherent capacity (1956, p. 66).

This use of 'linked' measurement scales created a novel set of growth data and permitted the use of growth analyses initially derived in other scientific areas.

2. Linear Models for Growth Curve Analyses

The simplest growth models are based on the simplest mathematical functions. The fitting of a straight-line to a set of measurements, a standard procedure in scientific research, was an initial model used by

Wishart (1938) in a classic paper on growth analysis. These linear models were quite practical because the calculations used standard statistical assumptions (e.g., independence and normality of the residuals), standard forms of estimation (e.g., Gauss-Markov least squares, maximum likelihood), and standard statistical tests of goodness-of-fit (F-tests, chi-square tests, and likelihood ratio tests).

Another key idea was also introduced by Wishart (1938) when he showed how power polynomials could be used to fit the nonlinear curvatures apparent in growth data. The individual growth curve (consisting of $t = 1, T$ occasions) is summarized into a small set of linear orthogonal polynomial coefficients based on a power-series of time ($t, t^2, t^3, \dots, t^p$) describing the general nonlinear shape of the growth curve. A model of growth data might require a second-order (quadratic), third-order (cubic), or even higher-order polynomial model fitted to the data, but the basic shape of each individual curve could be captured with a small number of parameters.

The organization of individual differences is an important consideration in any contemporary growth model. In the linear model, each individual is assumed to be following a straight line over time, describable by an intercept term and a slope term. The average of these individual scores can be used to describe the overall growth curve for the group, and the individual differences can be considered in the context of other substantive questions, e.g., do the groups under observation differ in intercepts or slopes? In the more complex polynomial models, Wishart suggested a second stage of analysis where these polynomial coefficients describing each individual curve were considered as the underlying scores that co-vary among different individuals. In a quadratic polynomial, for example, a person's trajectory over time could be summarized using an individual intercept, linear slope, acceleration, and a residual term, and secondary questions could be addressed, e.g., do the groups under observation also differ in acceleration?

This 'curve-fitting' or 'two-stage' approach was relatively easy to implement using a standard statistical basis for hypothesis testing (e.g., Bock 1975). The simplicity of this approach allowed linear growth models to enjoy widespread use in behavioral studies and become a useful starting point for understanding individual differences in change (e.g., Rogosa and Willett 1985). The persistent critiques of this polynomial approach include some key issues:

- (a) the individual parameters can be difficult to estimate, especially for higher-order terms and incomplete data,
- (b) even high-order polynomials (i.e., $p > 4$) cannot capture 'seasonality' or 'phase-shifts,'
- (c) the substantive meaning of these individual parameters is often in doubt, and
- (d) the individual level parameters can be transformed (i.e., rotated) to yield alternative meaning.

These standard critiques are appropriate, and they still apply to many other contemporary forms of growth curve analyses.

More complex linear models have been used to represent growth. A linear model applied to different age segments but with connected critical 'knot points' permitted a 'conjoined' or linear or polynomial 'spline' model (e.g., Bryk and Raudenbush 1992). This kind of spline model is not often used, partly because it is atheoretical, and partly because it may require a relatively large number of parameters to achieve adequate fit. A more parsimonious organization of individual differences was proposed in the form of 'latent curves,' independently defined by both Rao and Tucker in 1958 (cited in Meredith and Tisak 1990). Principal components of the repeated observations were used to reorganize these data into the sum of a small number of unspecified linear functions are fitted to growth data, the shape of the latent curves is determined by the component loadings, and the individual curve parameters are the component scores. The summation of latent curves has roots in classical nineteenth-century work of Fourier as now widely used. In a recent innovation, Meredith and Tisak (1990) showed how these 'Tuckerized curve' models could be represented using confirmatory common factor analysis models, used to represent a wide range of alternative models (including most of the models discussed herein), and fitted using structural equation modeling techniques (also see McArdle 1988, McArdle and Bell 2000, McArdle and Woodcock 1997).

3. Nonlinear Models for Growth Curve Analysis

The linear models can be used to describe a variety of nonlinear shapes, but many growth curve models for specific substantive problems have included nonlinear functions of the parameters. For example, Ebbinghaus in 1880 described his forgetting curves using a form of the classic 'exponential growth' model. Here the rate of change is a linear function of the percentage of initial size (e.g., compound interest), and individual trajectories rise or fall in nonlinear fashion towards some asymptotic values. In contrast, the nineteenth-century Velhurst curve of population growth, an S-shaped 'logistic curve,' was used in the early twentieth century by Pearl for many forms of growth. At about the same time Peters advocated an 'ogival' curve of growth in ideational learning, Thurstone found that a hyperbolic curve of learning best to fit the norms of many different tests, and both Ettlinger and Valentine demonstrated direct relationships among these functions (see Seber and Wild 1989, Bock and Thissen 1980).

More complex nonlinear models have been used extensively. In about 1825 Gompertz described the growth curve in terms of two exponential accumulations of different rates towards different asymptotes.

This flexible model was studied in the first half of the twentieth century by Winsor, used by Medwar to study the growth of chicken hearts, and by Deming for human physical growth, and in studies of learning curves (see Browne and du Toit 1991). Another popular nonlinear model, introduced around 1938 by von Bertalanffy, suggested that an individual's change in weight was the direct result of the difference in the forces of anabolism and catabolism, with the exact relationship specified by known empirical properties. In subsequent work, Richards challenged the empirical basis of this model, and formally showed how many prior models can be seen as specific solutions of a 'family' of deterministic differential equations; specific restrictions led to the exponential, logistic, Gompertz, and von Bertalanffy equations. This work was more recently extended by Nelder and Sandland and MacGhilcrest (1979). (See also Seber and Wild 1989, Zeger and Harlow 1987.)

Attempts to fit a single growth model to observations over a wide range of ages with a minimal set of parameters led researchers to combine aspects of other models. Jenss and Bayley in 1937 proposed a simple composite where a linear part was used to fit the early rapid growth phase and this was added to an exponential part to fit negative acceleration of the later slowing-down phase. A more complex composite model, the sum of multiple logistic curves, was suggested by Robertson in 1908 and Burt in 1937 but only made practical and fully developed by Bock and his colleagues (see Bock and Thissen 1980). In each segment the rate of growth (a) increases early, (b) reaches a maximum (peak growth velocity), (c) decreases towards the asymptote, and (d) the final value of one segment is used as the starting value of the next segment. A novel composite model based on multiple logistic functions was developed by Preece and Baines in 1978 and here the model parameters were supposed to be linearly related to meaningful biological parameters. Simple versions of this new family of curves have proven useful in contemporary work (see Hauspie et al. 1991).

The mathematical elegance of these nonlinear formulations often make it seem that the parameters estimated are related directly to the underlying phenomena and there is a tendency to treat these estimates with direct substantive interpretation. Prior research has shown this to be a difficult goal and now it is known that all models permit alternative formulations that may fit equally well. Another important property of growth curves comes when the fixed parameters of the nonlinear curves, the 'parameters of the means' also represents the 'means of the parameters' of the individual curves. However, classical research by Merrill in 1931, Estes in 1957, Tucker in 1966, and Keats in 1983 has shown this form of 'dynamic consistency' does not always hold, even for some of the most popular growth models (e.g., a group exponential). Contemporary statistical models are

now constructed carefully to clarify the unique substantive interpretation of nonlinear parameters (see Browne and du Toit 1991, Bock and Thissen 1980).

4. *Multivariate Growth Curve Analyses*

The collection of multiple variables at each occasion of measurement leads naturally to questions about relationships among growth processes and multivariate growth models. In biological research, early efforts were directed at characterizing the parallel properties of different growth variables. Models were developed originally to deal with the size of two different organs, and early nineteenth-century work was used by Huxley to form a classical 'allometric' model where two variables have a constant ratio of growth rates throughout the growth period. Many physical processes were subsequently found to grow in parallel, or in an ordered time-sequence (e.g., Tanner 1960). This work led to more sophisticated models based on systems of differential equations for the size of multiple variables. In one comprehensive multivariate model, Turner in 1978 extended the simple growth principles to more variables and permitted an examination of biologically important interactions based on the size and sign of the estimated parameters (see Griffiths and Sandland 1984).

Multivariate research in the behavioral sciences has relied on advanced versions of the linear growth models formalized by Rao, and Pothoff and Roy (see Bock 1975). The structural models developed within this framework have added measurement models and this includes work on multivariate models of latent curves (McArdle 1988). These structural models have proven useful in testing hypotheses about parallel growth processes across time in multiple psychological variables. In recent examples, McArdle and Woodcock (1997) showed how latent growth of a single common factor of intellectual abilities is a poor representation of multiple abilities data and McArdle (2000) showed how a multivariate dynamic sequence could be fitted using standard structural equations. Current research in this area is focused on an improved understanding of the time-sequences and interactions among multiple behavioral and biological variables.

5. *Recent Issues in Growth Curve Analysis*

The prior developments in growth curve analysis represent an important but limited class of longitudinal data analyses (e.g., Nesselroade and Baltes 1979, Bock 1975, Pinheiro and Bates 2000). These developments have been distinguished by features of the models themselves: (a) the degree of mathematical specification, (b) the way these statistical models are fitted, and (c) the clarity and substantive meaning of

the results. In future work, most of these problems are likely to remain under active investigation.

Most contemporary growth curve models can be written in a common symbolic form (Seber and Wild 1989). A general model for a change in the scores over time (often using derivatives dy/dt or differences $\Delta y/\Delta t$) can be based on some mathematical functional form ($f\{x\}$) with unobserved scores ($x[t]$) and with unknown parameters ($\beta[t]$) to be estimated. Additional statistical forms not discussed here can be included, including autoregressive residual error structures, Markov chains, and Poisson processes. The growth functions described here are not an exhaustive listing and future extensions to more generalized functions, and to dynamic and chaotic forms are likely.

Recent research in statistical model fitting is important here too. It is now clear that growth curve models of arbitrary complexity can be fitted to the observed trajectory over time (i.e., the integral) and the unknown parameters can be estimated to minimize some statistical function (e.g., weighted least squares, maximum likelihood), and parameter estimates obtained using nonlinear programming. This generic statistical approach avoids some problems in older techniques, such as fitting a model to a log-scale or directly to the velocity or the analysis of difference score data (Fig. 2). These new techniques make it possible to address the critical problems of forecasting future observations, and further research on Bayesian estimation is the focus of current research here (for details, see Seber and Wild 1989).

This generic model fitting approach also permits a wide range of new possibilities for dealing directly with unbalanced, incomplete or missing data. In early work, linear polynomials were used extensively to deal with these kinds of problems. But recent work on linear and nonlinear mixed and multilevel models have surfaced, making it possible to estimate growth curves and test hypotheses by collecting only small segments of data on each individual (Bryk and Raudenbush 1992, Pinherio and Bates 2000, Verbeke and Molenberghs 2000). These statistical models are being used in many longitudinal studies to deal with self-selection and subject attrition, multivariate changes in dynamic patterns of development, and the tradeoffs between statistical power and costs of person and variable sampling. The statistical power questions of the future are not 'how many occasions do we need?' but 'how few occasions are adequate?' and 'which persons should we measure at which occasions?' (McArdle and Woodcock 1997, McArdle and Bell 2000).

Some of the most difficult problems for future work on growth curves are related to the elusive meaning of the growth model parameters. Growth curve analyses also rely on basic measurement properties including (a) interval scaling, (b) age-and-time equivalence, and (c) substantively meaningful. These issues are key problems in the behavioral sciences and future

research is needed to address these fundamental concerns. Multivariate models may be useful here, especially as these models are further extended to a fully dynamic time-dependent form. In any case, empirical information and experimentation will be needed to judge the utility of any growth curve model.

Given the long history of elegant formulations from mathematics and statistics in this area, it is somewhat humbling to note that many of the most insightful growth curve analyses have come from careful visual inspection of the growth curves (e.g., in Figs. 1–4). The insight gained from visual inspection of a set of growth curves is not in dispute now and in fact obvious visual features should be highlighted and emphasized in future research (e.g., Pinherio and Bates 2000). As in the past, the best future growth curve analyses are likely to be the ones we can see most clearly.

See also: Hierarchical Models: Random and Fixed Effects; Latent Structure and Casual Variables; Learning Curve, The; Longitudinal Data; Statistical Analysis: Multilevel Methods; Time Series: General

Bibliography

- Bayley N 1956 Individual patterns of development. *Child Development* 27(1): 45–74
- Bock R D 1975 *Multivariate Statistical Methods in Behavioral Research*. McGraw-Hill, New York
- Bock R D, Thissen D 1980 Statistical problems of fitting individual growth curves. In: Johnston F E, Roche A F, Susanne C (eds.) *Human Physical Growth and Maturation: Methodologies and Factors*. Plenum, New York, pp. 265–290
- Browne M, Du Toit S H C 1991 Models for learning data. In: Collins L, Horn J L (eds.) *Best Methods for the Analysis of Change*. APA Press, Washington, DC, pp. 47–68
- Bryk A S, Raudenbush S W 1992 *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA
- Fischer G H, Molenaar I W 1995 *Rasch Models: Foundations, Recent Developments, and Applications*. Springer, New York
- Garrett H E 1951 *Great Experiments in Psychology*. Appleton-Century-Crofts, New York
- Griffiths D, Sandland R 1984 Fitting generalized allometric models to multivariate growth data. *Biometrics* 40: 139–150
- Hauspie R C, Lindgren G W, Tanner J M, Chrzastek-Spruch H 1991 Modeling individual and average growth data from childhood to adulthood. In: Magnusson D, Bergman L R, Rudinger G, Törestad B (eds.) *Problems and Methods in Longitudinal Research: Stability and Change*. Cambridge University Press, Cambridge, GB
- McArdle J J 1988 Dynamic but structural equation modeling of repeated measures data. In: Nesselroade J R, Cattell R B (eds.) *The Handbook of Multivariate Experimental Psychology*, 2. Plenum Press, New York, pp. 561–614
- McArdle J J 2000 A latent difference score approach to longitudinal dynamic structural analyses. In: Cudeck R, du Toit S, Sorbom D (eds.) *Structural Equation Modeling: Present and Future*. Scientific Software International, Lincolnwood, IL, pp. 342–380

- McArdle J J, Bell R Q 2000 Recent trends in modeling longitudinal data by latent growth curve methods. In: Little T D, Schnabel K U, Baumert J (eds.) *Modeling Longitudinal and Multiple-group Data: Practical Issues, Applied Approaches, and Scientific Examples*. Erlbaum, Mahwah, NJ, pp. 69–107
- McArdle J J, Woodcock J R 1997 Expanding test-retest designs to include developmental time-lag components. *Psychological Methods* 2(4): 403–35
- Meredith W, Tisak J 1990 Latent curve analysis. *Psychometrika* 55: 107–22
- Nesselroade J R, Baltes P B (eds.) 1979 *Longitudinal Research in the Study of Behavior and Development*. Academic Press, New York
- Pinheiro J C, Bates D M 2000 *Mixed-effects Models in S and S-PLUS*. Springer, New York
- Rogosa D, Willett J B 1985 Understanding correlates of change by modeling individual differences in growth. *Psychometrika* 50: 203–228
- Sandland R L, McGilchrist C A 1979 Stochastic growth curve analysis. *Biometrics* 35: 255–271
- Scammon R E 1927 The first serial study of human growth. *American Journal of Physical Anthropology* 10: 329–36
- Seber G A F, Wild C J 1989 *Nonlinear Models*. Wiley, New York
- Tanner J M (ed.) 1960 *Human Growth*. Pergamon Press, New York
- Verbeke G, Molenberghs G 2000 *Linear Mixed Models for Longitudinal Data*. Springer, New York
- Wishart J 1938 Growth rate determinations in nutrition studies with the bacon pig, and their analyses. *Biometrika* 30: 16–28
- Zeger S L, Harlow S D 1987 Mathematical models from laws of growth to tools for biologic analysis: Fifty years of growth. *Growth* 51: 1–21

J. J. McArdle

Guilt

Guilt has a number of related but distinguishable meanings. The word generally signifies having committed or being responsible for a wrongdoing, sin, or offence. This definition easily translates into the legal use of the word which means having committed a crime or tort as indicated by the offender's plea or the court's decision. A third usage of the word describes the feelings or emotion connected with recognition of having done wrong. Guilt, in other words, has a descriptive sense, a legal sense, and an affective sense.

1. Descriptive and Legal Meanings of Guilt

The Oxford English Dictionary traces the word guilt back to Old Teutonic words such as *geld*, meaning to pay, and Old English words (*gylt*, *gielt*, etc.) meaning delinquency, the fact of having committed an offence, being legally liable to a penalty, and a penalty for a crime. Modern definitions, such as are found in the seventh edition of the *Concise Oxford Dictionary*,

focus more simply on having 'committed a specific or implied offence.' Thus, in its broadest sense guilt simply signals that someone is responsible for an action or omission that is considered wrong. Implicit to this is the assumption that the guilty person was causally involved, though it has been argued that the person's intentions were not considered important in many times and places (Wierzbicka 1999).

As is evident from the origin and definitions of guilt, the concept is strongly linked to legal, especially criminal frameworks. Within a legal setting, the word guilt is used either by the defendant in pleading to a charge, or by the jury in convicting (a guilty verdict), to indicate responsibility for having committed a crime or tort (see Lely 1902). It often also indicates, as a consequence, liability for some penalty (Walker 1980).

To be guilty of a crime, one needs to commit the *actus reus* (the wrongful act) and to have *mens rea* (a guilty mind). Intention to commit the *actus reus* is the paradigm guilty mind. But over the centuries a variety of defences against a finding of *mens rea* have been refined in the common law: coercion, duress, mistake, self defence, provocation, insanity, necessity, accident and automatism being the main ones. The most sweeping defence against criminal responsibility is infancy, even though the *actus reus* has been committed with intent. Children under eight in most nations cannot be found guilty of a crime and there are generally qualifications on the capacity to hold older children criminally responsible. This is a major distinction between legal guilt and descriptive and affective guilt (which children can experience or at least are learning to experience).

In some contexts, recklessness—a weaker standard than intention but a stronger one than negligence—will be found to justify criminal guilt. In others the criminal law will countenance strict liability—guilt without *mens rea*. Generally this is reserved for contexts in which proving a guilty mind is difficult or when public policy regards the act itself to be sufficiently dangerous. Proving beyond reasonable doubt that the possessor of the illicit item had any criminal intent in relation to it is difficult, so the law resorts to the expedient of the *actus reus* of possession as sufficient for guilt. A second set of circumstances where the criminal law sometimes allows guilt without *mens rea* is where the *actus reus* involves a very great danger to the community and where the criminal actor has a capacity to take extraordinary steps to protect the community from that danger. Manufacturing pharmaceutical products or operating nuclear power plants are examples of where such strict liability crimes are countenanced in many nations.

Increasingly the world's legal systems accept the idea that a collectivity, such as a corporation, can be guilty of a crime. The reading of corporate criminal guilt initially became more expansive in the US; it then spread throughout the common-law world and in the final decades of the twentieth century took hold in the

civil-law world. Corporate criminal guilt can be based on a notion of corporate intention, or it can be determined that the corporation is vicariously responsible for the guilty intention of one of its members. Or corporate guilt can be based on a combination of a harmful corporate act that was enabled by irresponsible corporate policies or procedures (the corporate equivalent of the individual mental element of recklessness) (Bucy 1991, French 1984, May 1987, Fisse and Braithwaite 1993). In some ways the law imposes higher standards on corporations which are able to implement corporate compliance systems than it does on individuals. Arguably this spectrum of expectations is necessary for a 'risk society' (Beck 1992) where large corporations can cause aircraft to crash and nuclear power plants to melt down. At some point during the twentieth century in all industrialized societies, most litigants in the higher courts were corporations, rather than individuals.

Prior to the Middle Ages, when the distinction between tort and crime evolved, guilt was routinely used in civil law. Whether one burnt down the house of another intentionally (arson) or accidentally, one would be guilty of the act. In most legal systems guilt came to be used almost exclusively to refer to criminal guilt. However, when a civil wrong involved some element of fault, it remained appropriate to refer to the defendant as guilty of that civil wrong. While guilt was decreasingly so used when the fault element was only negligence, in US law it remained more common than in other systems to use the term guilt when exemplary damages were set in civil cases. When treble damages are imposed in a tort case, the existence of a punitive as well as a compensatory element in the court order means that it is appropriate to speak of the wrongdoer as guilty of that wrong. In equity and contract, if the conduct of a defendant is 'unconscionable' or shows bad faith, we can refer to the person as being guilty of unconscionable conduct. Finally, in US regulatory law there is the widespread phenomenon, less common in other nations, of civil penalties which can be imposed on evidence of guilt on the balance of probabilities, as opposed to the criminal test of beyond reasonable doubt. Americans speak of such defendants as being guilty of a breach of the law, even though there may not be criminal intent and defendants may not be able to avail themselves of excuses which would be available to them were criminal charges to be laid. Today, in a world where most of the things that can be done for good or ill in the world are done by organizations, most such civil penalty defendants are corporations. (See *Remedies and Damages: Legal Aspects*.)

At an even more macro level than the corporation, international law debates about crimes against humanity, such as genocide, resonate today with concern about 'impunity' as the world moves to establish an International Criminal Court. In this literature collective guilt is often supplanted by the concept of

collective shame as something nations might transcend through a process such as the South African Truth and Reconciliation Commission.

At both the macro level of national reconciliation and in the most micro context of juvenile justice and child protection, part of the impact of the growing social movement for restorative justice has been to shift the emphasis in debates around legal institutions from debates about guilt to debates about shame (Braithwaite 1999). We now turn to the affective foundations of this contest.

2. Feelings of Guilt

In addition to communicating both general and legal responsibility for wrongdoing, guilt has also come to be discussed as a feeling or emotion. At its core, a guilty feeling simply refers to the knowledge of one's guilt. However, the concept assumes not just awareness of wrongdoing but also discomfort with it. Thus, acceptance of having done wrong seems to be an important precondition of feeling guilty, although it is not when someone accepts or is assigned guilt in a descriptive or legal sense. The feeling of guilt is characterized by negative affect and may include behavioural changes. So, for example, it is sometimes said that someone is acting guilty or even that someone was consumed by guilt. Indeed, much classical literature, such as Shakespeare's *Macbeth* and Dostoevsky's *Crime and Punishment*, has followed the deterioration of characters haunted by past deeds. In such examples of guilt or guiltiness, the emotion is usually endured in secret, as the individual tries to hide his or her transgressions from the view of others. In fact, as will be seen, guilt is described by some theoretical traditions as a distinctively private emotion, especially in contrast to other related emotions such as shame.

Perhaps the most influential, or at least the most widely recognised, use of guilt in the social sciences has occurred within psychoanalytic theory. Freud (1949) argues that a primitive basis for guilt is the fear of losing another's love. In this primitive form feelings of guilt occur when one's transgression is discovered by an authority figure, or when the individual fears its discovery. This, according to psychoanalytic theory, is the only form of guilt young children can feel. However, individuals become able to feel what might be called true guilt with the formation of a super-ego or conscience, which internalizes the values that might have otherwise been enforced by an external authority. In this way individuals come to punish themselves through feeling bad (guilty) for transgressions or, because of the internal nature of the superego, even for offensive thoughts. Within this framework, conscious feelings of guilt, which Freud suggests should be called remorse, occur after committing a transgression. However, Freud also suggests that unconscious feelings of guilt play a much more malevolent role and are

the underlying cause of neuroticism. This is because the superego is characterized as taking on the aggressive and punishing characteristics of the external authority as well as the individual's own feelings of aggression towards the external authority. Ultimately the superego redirects this aggressiveness towards the ego, or conscious self. Ignoring the complexities raised by Freud's analysis of neuroticism and unconscious forms of guilt, what is most significant about this psychoanalytic perspective is that it describes guilt as an internally generated sanction based upon the individual's recognition of wrongdoing. For further detail on Freud's conception of guilt see his *Civilisation and its Discontents* (1949).

Guilt is also assigned an important role in the comparison of cultures by anthropologists such as Benedict (1946) and Mead (1937). In comparing the way societies maintain social control, a distinction is drawn between those that rely more heavily on internalized feelings of guilt and those that rely to a greater extent on social mechanisms of shame and shaming to maintain conformity. In Benedict's famous comparison, United States society is classified as a guilt culture, while Japanese society is classified as a shame culture. This distinction reflects Freud's conception of guilt as an internally generated sanction which develops as a result of socialization practices. However, it also highlights the alternative emotion of shame, which is described as a reaction to others' criticism and is similar to Freud's conception of 'primitive guilt.' This contrasting of shame and guilt as social versus internal control has been an influential one, with much of the recent research on guilt and shame examining the social vs. non-social dimension (Cheek and Hogan 1983, Gibbons 1990, Wallbott and Scherer 1995, Tangney et al. 1996). Indeed, as will be seen, a growing interest in shame has meant that research on both emotions has been dominated by an interest in distinguishing them.

While the social distinction between shame and guilt has been influential, it has also been questioned by a number of perspectives. Piers and Singer (1953) argue that there is much evidence from everyday life that individuals experience shame in private and without others' intervention. Building on the psychoanalytic framework they argue that shame and guilt are distinguishable because they result from different psychological tensions. Whereas guilt arises from tension between the ego and superego (conscience), shame is the product of tension between the ego and what they call the ego ideal. The ego ideal might be described as the individual's perception of their ideal self, those attributes and characteristics to which he or she aspires. This structure implies a new distinction: shame is about failure, it represents a shortcoming, whereas guilt is about transgression, it represents breaking a rule. Lewis (1971) argues that a second distinction implied by this psychoanalytic structure concerns the object of the two emotions. When feeling

guilt the individual's attention is focused upon the act or omission itself, but because shame is about failure to achieve an ideal, it is focused upon a global evaluation of the whole self. Thus, in addition to the emphasis upon transgression evident in the legal, Freudian and early anthropological perspectives, this later work has argued that guilt is felt about specific actions rather than the self.

The distinctions between guilt and shame have become a significant focus of recent empirical research. While empirical studies show that subjects' memories of feeling guilt involve less self-conscious feelings than their memories of shame (Wicker et al. 1983, Tangney et al. 1996), evidence that guilt is less focused upon evaluation of the whole self is equivocal. While at least one study (Niedenthal et al. 1994) has suggested that guilt is more specific in focus, a number of other studies (Tangney et al. 1996, Wicker et al. 1983) have found no significant difference between the emotions on this dimension. Indeed, a recent study (Harris 1999) which clustered subjects' reports of feelings using factor analysis did not distinguish between shame and guilt in either dimension. These results suggest that shame and guilt may often be empirically indistinguishable.

The discussion highlights the importance that has been placed in recent time on distinguishing feelings of guilt and shame. It is equally apparent that the issue has not been resolved and continues to be an important focus of research. This is particularly evident in recent research which explores guilt proneness and shame proneness as personality dispositions that are related to individual coping mechanisms and psychopathologies such as depression (Tangney 1991, Harder 1995, Ferguson et al. 1999). Despite disagreement over the dimensionality of guilt, research has highlighted a number of characteristics of guilt: that it is felt as a result of the acceptance that one has committed a wrongdoing; that it is focused upon one's actions or omissions; and that it often involves the perception of having hurt others in some way (Kugler and Jones 1992).

See also: Confessions: Psychological and Forensic Aspects; Crime: Knowledge about and Prevalence; Criminal Defenses; Criminal Law and Crime Policy; Honor and Shame; Law: Defense of Insanity; Personality and Crime; Responsibility: Philosophical Aspects; Shame and the Social Bond

Bibliography

- Beck U 1992 *Risk Society: Towards a New Modernity*. Sage, Newbury Park
- Benedict R 1946 *The Chrysanthemum and the Sword; Patterns of Japanese Culture*. Houghton Mifflin, Boston
- Braithwaite J 1999 Restorative justice: Assessing optimistic and pessimistic accounts. In: Tonry M (ed.) *Crime and Justice: A*

- Review of Research. University of Chicago Press, Chicago, Vol. 25
- Brett P 1963 *An Inquiry into Criminal Guilt*. Law Book Company, Sydney
- Bucy P H 1991 Corporate ethos: A standard for imposing corporate criminal liability. *Minnesota Law Review* 75: 1095–184
- Cheek J M, Hogan R 1983 Self-concepts, self-presentations, and moral judgements. In: Suls, Greenwald A G (eds.) *Psychological Perspectives on the Self*. Lawrence Erlbaum Associates, Hillsdale New Jersey, Vol. 2
- Ferguson T J, Stegge H, Miller E R, Olsen M E 1999 Guilt, shame and symptoms in children. *Developmental Psychology* 35(2): 347–57
- Fisse B, Braithwaite J 1993 *Corporations, Crime and Accountability*. Cambridge University Press, Cambridge
- French P A 1984 *Collective and Corporate Responsibility*. Columbia University, New York
- Freud S 1949 *Civilisation and its Discontents*. Hogarth Press and Institute of Psycho-Analysis, London
- Gibbons F X 1990 The evolution and manifestation of social anxiety. In: Crozier W R (ed.) *Shyness and Embarrassment: Perspectives from Social Psychology*. Cambridge University Press, Cambridge
- Harder D W 1995 Shame and guilt assessment, and relationships of shame- and guilt-proneness to psychopathology. In: Tangney J P, Fischer K W (eds.) *Self Conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride*. Guilford Press, New York, Chap. 17, pp. 368–92
- Harris N 1999 *Shaming and shame: An empirical analysis*. PhD thesis, Australian National University, Canberra
- Kugler K, Jones W H 1992 On conceptualizing and assessing guilt. *Journal of Personality and Social Psychology* 62(2): 18–27
- Lely J M 1902 *Wharton's Law-Lexicon* (10th edn.). Stevens, London
- Lewis H B 1971 *Shame and Guilt in Neurosis*. International Universities Press, New York
- May L 1987 *The Morality of Groups: Collective Responsibility, Group-based Harm and Corporate Rights*. University of Notre Dame Press, Notre Dame, IN
- Mead M 1937 *Cooperation and Competition Among Primitive Peoples*. McGraw-Hill, New York
- Niedenthal P M, Tangney J P, Gavanski I 1994 'If only I weren't' versus 'If only I hadn't': Distinguishing shame and guilt in counterfactual thinking. *Journal of Personality and Social Psychology* 67(4): 585–95
- Piers G, Singer M B 1953 *Shame and Guilt: A Psychoanalytic and a Cultural Study*. Norton, New York
- Tangney J P 1991 Moral affect: The good, the bad, and the ugly. *Journal of Personality and Social Psychology* 61(4): 598–607
- Tangney J P, Miller R S, Flicker L, Barlow D H 1996 Are shame, guilt, and embarrassment distinct emotions? *Journal of Personality and Social Psychology* 70(6): 1256–69
- Walker D M 1980 *The Oxford Companion to Law*. Clarendon Press, Oxford
- Wallbott H G, Scherer K R 1995 Cultural determinants in experiencing shame and guilt. In: Tangney J P, Fischer K W (eds.) *Self Conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride*. Guilford, New York, pp. 465–87
- Wicker F W, Payne G C, Morgan R D 1983 Participant descriptions of guilt and shame. *Motivation and Emotion* 7: 25–39

Wierzbicka A 1999 *Emotions Across Languages and Cultures: Diversity and Universals*. Cambridge University Press, Cambridge

N. Harris and J. Braithwaite

Gynecological Health: Psychosocial Aspects

Most women do not experience menstrually related changes that interfere with their lives to any appreciable extent. Where such changes are identifiable, they do not have a simple biological basis but are embedded in a complex social and cultural context.

1. Menstrually Related Distress

A comprehensive review (Klebanov and Ruble 1994) indicated no relationship between subjective distress and any hormonal measure; there is no consensus on its definition, symptoms, prevalence, effects, or indeed the component of the menstrual cycle (premenstrual, menstrual, or intermenstrual) with which the distress is associated.

Although mood, physical symptoms, and coping with life are claimed to demonstrate menstrually related cyclicity, the empirical evidence is mixed and inconclusive. Decades of research provide evidence for no cyclic variation in any aspect of cognitive functioning, memory, or motor performance (Richardson 1992). Women's reports of symptoms and mood swings are no different from those experienced by men (McFarlane and Williams 1994), and menstrually-related fluctuations in mood are considerably less than those associated with days of the week.

Prevalence estimates of menstrually related distress range from 95 to 5 percent, illustrating a complete lack of definitional consensus. The main standardized self-report survey, the Moos Menstrual Distress Questionnaire, obtains prevalence rates between 55 and 70 percent for physical discomfort, and between 23 and 70 percent for negative affect, with the majority of symptoms extremely mild and causing negligible distress (Logue and Moos 1986).

The American Psychiatric Association included a preliminary definition of premenstrual dysphoric disorder (PDD) in DSM-IV. PDD involves 'symptoms such as markedly depressed mood, marked anxiety, marked affective lability, and decreased interest in activity ... during the last week of the luteal phase in most menstrual cycles during the last year. The symptoms begin to remit within a few days of the onset of menses and are always absent in the week following menses.' PDD is diagnosed if the condition is com-

parable in severity to a major depressive episode, involving marked impairment in either social or occupational functioning. PDD must be confirmed with prospective daily ratings over at least two cycles, since retrospective reports are unreliable and there is little consistency across consecutive cycles. Rivera-Tovar and Frank (1990), using a prospective, blind, study over 90 days, found that 5 percent of US students met criteria for PDD. This suggests that some women experience menstrually related distress, but that it is not the norm, and it should be noted that experiencing distress on a cyclic basis does not necessarily mean that it is hormonally mediated.

2. Postnatal Distress

It is often assumed that negative psychological outcomes of childbirth result from hormonal changes, but three main categories of childbirth-related psychological distress with different causes and prognoses need to be distinguished.

2.1 Baby Blues

Postpartum dysphoria, or 'baby blues,' tearfulness, mood swings, and irritability, occurs in a high proportion of mothers in the few days following childbirth, and normally resolves itself within 10 days. It is not related to any environmental, social, or cultural factor (Kumar 1994), and may be related to changes in estrogen levels, although individuals undergoing surgery experience similar symptoms, suggesting it may be a reaction to physical stress. Prevalence estimates vary from 26 to 85 percent of all mothers. The condition is distressing, especially if women are not expecting it, but resolves quickly and has little long-term impact.

2.2 Postpartum Psychosis

Postpartum psychosis, characterized by hallucinations, delusions, and other schizophrenia-like symptoms, occurs in all societies, at a relatively consistent rate of 0.01 to 0.02 percent of all births (Kumar 1994), and can result in suicide or infanticide. Previous psychotic illness is the strongest risk factor for postpartum psychosis; there is no link with life events, stress, or relationship difficulties, supporting the view that it arises, like other psychotic states, in biologically susceptible individuals as a reaction to substantial life change.

2.3 Postpartum Depression

Nonpsychotic postpartum depression is mainly influenced by psychosocial factors, and there is no

convincing evidence for hormonal causation (Kumar 1994). Postpartum depression is predicted by relationship difficulties, low social support, previous depression, and stress. It is identical in presentation to other major depression, and while DSM-IV allows the specifier 'with postpartum onset' to be applied to depression which occurs within four weeks of the delivery of a child, it has no separate diagnostic category for postpartum depression.

Rates of depression in new mothers, assessed by the Edinburgh Postnatal Depression Scale (Cox et al. 1987), are between 6 and 10 percent. Symptoms may begin at any time up to 12 months after the birth of the baby and may be preceded by normal adaptation. High rates of depression persist as long as the mother has young children at home, and fathers also experience postnatal depression. Looking after a baby is stressful and anxiety provoking, and fatigue and disturbed sleep add to the difficulty. Depression in the postpartum period is best treated as reactive depression, reflecting a need for social and practical support, rather than as unique to childbirth.

3. Fertility-related Problems

3.1 Infertility

Infertility is generally defined as the failure to achieve a pregnancy after 12 months of unprotected sexual intercourse. Research on failure to conceive focuses on women, even though the cause is equally likely to be a characteristic of the male partner (Rowland 1992). The main causes are biological, although stress and anxiety can exacerbate physical problems.

Infertility and its treatment are distressing, but neither is associated with major psychological disturbance. The long-term outcomes of successful *in vitro* fertilization for children and their families are positive, but most outcomes are less benign. Involvement in *in vitro* fertilization is disruptive, expensive, and stressful. The probability of bearing a living child is around 10 percent per treatment cycle, and the risk of gynecological and obstetric complications, malformations, and multiple births is significantly increased. Repeated cycles of treatment are associated with increasing levels of depression (Rowland 1992).

3.2 Fertility Control

Access to adequate contraception and to abortion is important to women's physical and psychological well being. Where abortions are legal, around 2 percent of women aged between 15 and 44 have an abortion in any year. The worldwide annual death rate from illegal abortions has been estimated at 100 000, while the death rate from legal abortions is 0.05 per 10 000, 25 times lower than the rate for childbirth. Provision

of safe abortions and adequate contraception would have a major impact on the well being of women and children in developing countries.

Women may need support in deciding whether to choose abortion, but the majority experience no significant psychological distress. Psychological distress following abortion is considerably lower than that associated with childbirth. Distress is highest during decision making before the abortion, and the most common postabortion reaction is a sense of relief.

Women who cope best with abortion are characterized by optimism, high self-esteem, and a sense of personal control. Social, cultural, and partner support, and confidence in the decision, assist in coping, while women with a history of psychiatric disorder are likely to react negatively (Adler et al. 1992).

3.3 Miscarriage

Miscarriages are relatively common; over 30 percent of 50-year-old women have had at least one miscarriage. Reactions to spontaneous abortion of a wanted pregnancy are generally characterized by short-term grief, anxiety, and depression, normally resolved successfully within a few months (Slade 1994). Reactions to medical abortions follow a similar pattern, with short-term distress which may be severe but which resolves rapidly. Serious negative consequences are rare, and associated with previous psychiatric illness or lack of social support.

4. Menopause

Menopause is a natural physiological process resulting from aging of the ovaries, leading to cessation of ovulation and of estrogen production. Natural menopause occurs over several years. During the perimenopause, estrogen levels decrease and menstruation becomes irregular and ceases. The postmenopause, defined after 12 months without menstruation, is reached at an average age of 50.4 years.

Menopausal symptoms are categorized as: vasomotor (e.g., hot flushes, night sweats), psychosomatic (e.g., headaches, dizziness), and psychological (e.g., tiredness, irritability). The most common symptoms, reported across most cultures, are hot flushes and night sweats, clearly the result of changes in estrogen. However, prevalence of symptoms varies between countries, suggesting that sociocultural factors play a role (Hunt 1994).

Hormone replacement therapy relieves hot flushes and night sweats, but does not influence psychological well being in mid-aged women. Stress, life events, history of depression, and expectations are the strong-

est predictors of psychological well being during menopause. Women in their middle years often experience major life events (McKinlay et al. 1987). Once life changes are accounted for, there is little evidence to suggest that menopausal women are particularly prone to negative psychological states.

5. Conclusion

Distress associated with gynecological events is better explained by poor social support and by negative expectations than by endocrine function. Future research might usefully examine the experiences of women in a wider range of situations, including women in same-sex relationships, women who choose childlessness, and women with disabilities. An exploration of the experiences of women in different ethnic groups would also provide a broader perspective on psychosocial aspects of gynecological health.

See also: Fertility and Culture: Anthropological Insights; Fertility Control: Overview; Fertility of Single and Cohabiting Women; Fertility: Proximate Determinants; Fertility Theory: Embodied-capital Theory of Life History Evolution; Gender and Health Care; Gender and Physical Health; Gender and Reproductive Health; Gender Role Stress and Health; Reproductive Medicine: Ethical Aspects; Reproductive Rights in Developing Nations; Social Networks and Fertility; Women's Health

Bibliography

- Adler N E, David H P, Major B N, Roth S H, Russo N F, Wyatt G E 1992 Psychological factors in abortion: A review. *American Psychologist* **47**: 1194-204
- Cox J L, Holden J M, Sagovsky R 1987 Detection of postnatal depression: Development of the 10-item Edinburgh Postnatal Depression Scale. *British Journal of Psychiatry* **150**: 782-6
- Hunt K 1994 A 'cure for all ills'? Construction of the menopause and the chequered fortunes of hormone replacement therapy. In: Wilkinson S, Kitzinger C (eds.) *Women and Health: Feminist Perspectives*. Taylor & Francis, London
- Klebanov P K, Ruble D N 1994 Toward an understanding of women's experience of menstrual cycle symptoms. In: Adesso V J et al. (eds.) *Psychological Perspectives on Women's Health*. Taylor & Francis, Washington, DC
- Kumar R 1994 Postnatal mental illness: A transcultural perspective. *Social Psychiatry and Psychiatric Epidemiology* **29**: 250-64
- Logue C M, Moos R H 1986 Perimenstrual symptoms: Prevalence and risk factors. *Psychosomatic Medicine* **48**: 388-409
- McFarlane J M, Williams T M 1994 Placing premenstrual syndrome in perspective. *Psychology of Women Quarterly* **18**: 339-73

- McKinlay J B, McKinlay S M, Brambilla D 1987 The relative contributions of endocrine changes and social circumstances to depression in mid-aged women. *Journal of Health and Social Behavior* 28: 345–63
- Richardson J T E 1992 *Cognition and the Menstrual Cycle*. Springer, New York
- Rivera-Tovar A D, Frank E 1990 Late luteal phase dysphoric disorder in young women. *American Journal of Psychiatry* 147: 1634–6
- Rowland R 1992 *Living Laboratories: Women and Reproductive Technologies*. Pan Macmillan, Sydney, Australia
- Slade P 1994 Predicting the psychological impact of miscarriage. *Journal of Reproductive and Infant Psychology* 12: 5–16

C. Lee

Gypsies and Travelers

The label *Gypsy* refers to diverse groups—each with its own history, culture, and name—who live in scattered communities throughout Europe and North America and in parts of Asia, Central and South America, and Africa. (In Europe, the label *Roma* is commonly used.) The English word ‘Gypsy’ is derived from ‘Egyptian,’ based on the false assumption that they had come from Egypt. From linguistic evidence, however, it is apparent that Gypsies originated in India. They began migrating westward in the tenth century, spreading across Europe and absorbing aspects (e.g., religion, language) of the host cultures they lived among. Eventually, they formed the groups found today. The Indic structure of the major dialects of *Romani* spoken by Gypsies has been heavily influenced by European languages, Greek, Armenian, Persian, and possibly Arabic.

Traveler is the term commonly used to refer to indigenous groups who share economic and structural features with Gypsies. They practice (or once practiced) similar peripatetic occupations and occupy a similar minority position vis-à-vis the majority populations of the countries in which they live. In Ireland and Scotland, for example, indigenous groups once known as ‘tinkers’ and ‘tinklers,’ but who refer to themselves as Travelers or Traveling People, have occupied a niche very similar to Gypsies for at least 300 years. Similar indigenous Travelers found in other European countries include the *Resande* in Sweden and *Woonwagenbewoners* in The Netherlands. Other indigenous groups are found in South Asia, they include the *Kanjar* in Pakistan and northern India and the *Ghorbat* in Afghanistan. Like Gypsies, members of some Traveler groups have since migrated to other countries. Irish and Scottish Travelers, for example, now also live in the United States, having first emigrated there in the mid-1800s. The designation Traveler is also sometimes used as a broad term to encompass both indigenous Travelers and Gypsy populations.

1. Identity

All Gypsies and Travelers see themselves in opposition to outsiders who they typically categorize with a single label. Many Gypsies, for example, use the term *gadje* (non-Roma in the Romani language); Irish Travelers call members of the majority Irish population *buffers*. Most Gypsies follow, to varying degrees, a system of purity rules which govern women’s behavior and how food, eating utensils, clothing, washing containers, and the body are handled. Most believe that the body is ritually unclean from the waist down and, consequently, that anything associated with the lower body is polluting. Individuals and groups who do not follow purity rules are by definition ‘polluted’ or ritually unclean, which is one factor reinforcing the boundaries between Gypsies and non-Gypsies. Of course, Gypsies and Travelers also distinguish their own group from similar groups, usually on the basis of kin ties (or lack thereof), differences in dialect, and other criteria such as socioeconomic status, occupation, and length of settlement (i.e., before 1945 or during the communist era for East European groups).

Language also acts as a marker of ethnic identity. While Gypsies speak the language of the majority population of the countries in which they live, most also speak a form of Romani. Most indigenous Travelers likewise have specialized argots which they typically use to disguise their communications in the presence of outsiders (e.g., during economic transactions). Other markers of group identity, such as items of material culture, skills, or traditions, are also used to create and reinforce a symbolic separation between themselves and outsiders. Roma in Hungary, for example, assert that it is their ‘Gypsy work’—their wit, speech, and ability to hustle—that allows them to survive, rather than the income they make from the wage labor the State forces them to engage in. Irish Travelers pride themselves in their ‘cleverness’—their ability to live by their wits—and see this as something which differentiates them from other Irish. It has been suggested that ‘disreputable’ behavior of Spanish Gitanos—their flaunting of accepted social norms about appropriate public comportment (e.g., loudness, violations of street etiquette)—is a deliberate strategy intended to offend non-Gypsies and keep them at a distance. Other groups use their involvement in horse trading or their distinctive music and celebrations to symbolically draw boundaries between themselves and outsiders.

When possible, Gypsies and Travelers limit their interaction with members of the majority population to economic transactions and brief, necessary encounters with institutional representatives such as legal, social service, or medical personnel. Most academics who have studied these groups have documented their antiestablishment attitudes and efforts to remain outside the direct supervision of the State. Socialization among Gypsies and Travelers instills in

most children a strong sense of ethnic pride, and few express a desire to abandon their identity or group. This pride is reflected in and reinforced by high rates of group endogamy. Nevertheless, attrition in the form of 'passing' into mainstream society as well as intermarriage with non-Gypsies or non-Travelers does occur. Some groups have been subjected to long-term state-sponsored assimilation programs, which have included forced settlement among non-Gypsies, wage labor, and education.

As the quintessential Other, Gypsies and Travelers have historically captured the imagination of artists and writers and figure prominently in the arts and literature of many countries. In Russia, for example, writers such as Pushkin, Tolstoy, and Blok included Roma in their works. Likewise, Travelers ('tinkers') have figured prominently in the work of such literary figures as Irish playwright J. M. Synge.

2. Economic Organization

The traditional economy of Gypsies and Travelers was based on petty trade and the provision of craft products, specialized services, and entertainment to customers and clients in the majority populations they lived among. In many countries, they serviced a scattered rural population, conveniently bringing goods and services to their customers' doors. Others provided services which had to be performed on site. Rom wipe-tinners, for example, traveled the United States retinning expensive mixing bowls and massive, relatively immobile counter tops, holding tanks, tubs, troughs, sinks, and separators used in institutional cafeterias, dairies, and commercial bakeries. Black-topping (or tarmac laying), a common occupation of Gypsies and Travelers in North America, Ireland, and Great Britain today, likewise must be done on site. Sometimes, Gypsies and Travelers provide unique products or services that cannot be acquired from anyone else. The Kanjar of Pakistan, for example, sold handmade terracotta toys, while the Qalandar entertained their clients with trained monkeys and bears. Fortune telling remains a profitable specialization for many American Rom. Because Gypsies and Travelers often accepted payment in kind (e.g., farm produce, fuel), the services they provided were especially attractive to poorer clients.

Occupational flexibility, the ability to do several income-earning activities at the same time, and to change occupations quickly as demand dictates, is a key aspect of the traditional economy. When the Ludar arrived in the United States from Bosnia, for example, they brought trained bears and monkeys, but they also peddled small goods, dealt in horses, told fortunes, and made rustic furniture. Since the 1950s, they have focused on laying blacktop, tarring roofs, and working carnival concessions. Irish Travelers

once swept chimneys, made tin containers, bought and sold donkeys and horses, peddled small goods (e.g., brushes, needles, boot polish, religious pictures), told fortunes, and performed agricultural labor. As these rural-based occupations became obsolete, they moved to urban areas and began laying tarmac, selling carpeting from door-to-door, collecting and recycling scrap metal, and selling second-hand furniture, car parts, and other merchandise from the roadside.

Many Gypsy and Traveler populations—or significant segments of them—are still peripatetic, exploiting narrowly-defined markets within the host economy. Families are successful because their overhead is minimal, they are willing to accept a narrow profit margin from multiple sources, and generally lack interest in material accumulation and capital expansion. They also use wage-free household labor. Typically, household income is produced by all family members—men and women, young and old. When larger work units are needed, they are drawn from the kin network or else the lowest socioeconomic stratum of the host society (e.g., the homeless).

While nomadism remains an important feature of many groups—and is the image instantly evoked by the very words 'Gypsy' and 'Traveler'—substantial numbers today are sedentary for all or much of the year. Some Western European groups are also sedentary. Spanish Gitanos, for example, have been sedentary for centuries. Although many Gypsies and Travelers in Ireland and Great Britain remain nomadic, the pressure from government to settle is strong. In Eastern Europe, Gypsies were long ago forced into wage labor—generally low-status, unskilled jobs—and have been sedentary for generations. In North America today, only a few families in each Gypsy and Traveler group are predominantly nomadic. Most own or rent homes, and maintain a mixed pattern of settlement and travel.

3. Social and Political Organization

It is difficult to generalize about the social organization of Gypsy and Traveler groups. Many groups reckon descent bilaterally and have no formalized or corporate kin groups. Individuals have equal rights and obligations towards both maternal and paternal kin. At marriage, a couple can choose to affiliate with either spouse's family, and later, may periodically live on their own or affiliate with nonkin. Other groups—such as the Rom and Ludar in the United States—trace descent through the male line, have named patrilineages, and practice preferential patrilocal residence at marriage, at least until the birth of a young couple's first child. Descent in most South Asian groups is also patrilineal.

Residential groupings larger than the household, whether they are travel clusters or stationary enclaves

within cities, are fluid. Kinship is the fundamental basis for association, but group composition and size is continually being renegotiated in response to changing economic opportunities and a host of social and legal concerns. Groups split apart due to personal animosities, the desire for change, the need to scatter in order to find new clients, or as a result of punitive actions (e.g., evictions) by local authorities. Likewise, they coalesce for many reasons. Scattered kin come together to celebrate life crisis events such as marriages and funerals, to protect a valuable economic territory from the encroachment of other groups, to defend a kinsman from insult or attack, and, sometimes, to make evictions by local authorities more difficult.

Political organization is also diverse. The most formalized mechanisms of social control and political leadership are found among some Gypsy groups and South Asian Travelers. American Rom, for example, have a *Rom Baro* (big man) who speaks for his extended family and controls economic activities in the territory they claim. Important disputes are adjudicated by a *kris*, an *ad hoc* Gypsy court, composed of popularly-nominated, male family heads who act as 'judges' and representatives from each kin grouping within the vicinity who act as witnesses and jury. The *kris* seeks consensus; its proceedings are spontaneous and often volatile; and the fines or punishments it metes out are morally, not legally, binding. Other American Gypsy groups such as the Romnichel, however, lack both the *Rom Baro* and the *kris*. Various South Asian groups have loosely structured *panchayat* councils. These are informal meetings convened primarily to settle disputes, with respected elders acting as arbitrators. Both the *kris* and the *panchayat* are mechanisms for airing grievances, alleviating tension, asserting (and sometimes reassessing) group values, arriving at a compromise, and restoring order.

Researchers working with diverse groups of Gypsies and Travelers in Europe, North America, and South Asia have noted the importance members attach to assembly, both at annual events like seasonal fairs and at periodic events like religious pilgrimages, marriages, and funerals. During assemblies, group solidarity is reinforced through informal social interaction (including music, song, and drinking), marriage negotiations, religious ceremonies, and dispute settlements. Such gatherings demonstrate and reinforce a common identity that overrides, at least temporarily, localized kinship and political alliances. English Gypsies appear to recognize the important role fairs play in their collective life. When authorities tried to close Appleby Fair in 1965 and to prevent them from assembling at Epsom in 1967, they united in protest.

Unless prevented from doing so by the state, Gypsies and Travelers seem to respond eagerly to opportunities for assembly and social engagement. Their gatherings also have an emotional intensity—evident in the total absorption of buleria performances, flamenco, and 'deep song' as performed by

Spanish Gitanos and in the drinking and aggressive interpersonal interactional style of many groups—that is striking to the researchers who have observed them. It is as if what these Gypsy and Traveler groups lack in formal organization or structure they make up for by periodic assembly and intense emotional engagement.

4. Harassment and Assimilation

Gypsies and Travelers have sometimes enjoyed the liberties of other citizens. Under Ottoman rule, for example, Gypsies had the same citizenship rights as anyone else as long as they paid the taxes required of their religious group: Christian Roma paid higher taxes than Muslim Roma. Much more frequently, however, they have been persecuted. In Europe, Gypsies and Travelers have been subject to punitive legislation for many centuries. Individuals as well as entire groups have been denied entry, forcibly settled, had their trades outlawed, forced to adopt wage labor, had their children taken and adopted by non-Gypsy families, been sterilized, imprisoned, enslaved, or deported. In Moldavia and Wallachia—Romania's two historic provinces—Gypsies were sold into slavery as early as the fifteenth century; legal emancipation for Romany slaves or *robi* finally came in 1864. During the Holocaust a half to more than a million Gypsies were exterminated. Deep-seated anti-Gypsy prejudice continues today throughout Eastern Europe and parts of Western Europe, periodically erupting in physical assaults by local residents and the police.

In the second half of the twentieth century, many Western European governments and private agencies became involved in programs ostensibly aimed at providing Gypsies and Travelers with 'choice,' 'options,' or 'self-determination.' Many of these movements came about as modernization in the host society eliminated the need for the rural-based trades and services nomadic groups had provided, forcing them to seek new clients. In Ireland and Great Britain, for example, Travelers and Gypsies converged on cities looking for work, camping along motorways, next to housing estates, and on rapidly diminishing amounts of open land. This led to a public outcry and their definition as a 'social problem.' Public policy has fluctuated over the years. Liberal programs (i.e., providing campsites and mobile classrooms for nomadic families, creating culturally-sensitive curricula) have vied with more straightforward attempts to assimilate Gypsies and Travelers through housing, laws prohibiting camping, and conventional vocational training education. In recent years, however, Gypsies and Travelers have become more politically active. Many organizations have emerged to articulate their concerns and lobby government for the same rights as any other minority. Cooperation across national boundaries with other Gypsy and Traveler groups is also becoming more common.

In Eastern Europe, government policy had focused more on forced assimilation, including aggressively settling Gypsy families that were nomadic, dispersing those who had lived in settled enclaves, enrolling children in school, licensing or banning specific trades, forcing adults to accept wage labor, and officially denying their existence as ethnic minorities. Many Gypsies were able, nevertheless, to preserve their separate identity in spite of official communist policy. The collapse of state socialism in Eastern Europe in 1989 brought new hope to the approximately five million Gypsies in the region who responded by organizing a plethora of cultural groups and political parties to articulate and represent their diverse interests. Some common goals include achieving full recognition and civil rights as distinct ethnic minorities, having schools use Romani as a second language of instruction, and getting state-supported social and economic programs to create opportunities and address poor living conditions. In Macedonia, Gypsies called upon the United Nations to establish a Romany state to be called 'Romanestan.' Most groups efforts, however, have been hampered by internal leadership problems, the multiplicity of competing Gypsy organizations within each country, and, ultimately, their reliance on the State to accord their organizations and demands legitimacy. Anti-Gypsy sentiment in Europe remains strong. Attacks on Gypsies in the 1990s targeted both long-term residents of settled communities in Eastern Europe and recent immigrants fleeing to Western countries from Romania, Bulgaria, Hungary, and the former Yugoslavia.

See also: Counterculture; Diaspora; Ethnic Groups/Ethnicity: Historical Aspects; Ethnic Identity, Psy-

chology of; Ethnicity: Anthropological Aspects; Ethnicity, Sociology of; Nomads/Nomadism in History; Stigma, Social Psychology of; Subculture, Sociology of

Bibliography

- Barany Z 1998 Ethnic mobilization and the state: the Roma in Eastern Europe. *Ethnic and Racial Studies* 21(2): 308–27
- Crowe D 1994 *A History of the Gypsies of Eastern Europe and Russia*. St. Martin's Press, New York
- Gmelch S B 1986 Groups that don't want in: Gypsies and other artisan, trader, and entertainer minorities. *Annual Review of Anthropology* 15: 307–30
- Kendrick D 1993 *Gypsies from India to the Mediterranean*. CRDP, Toulouse, France
- Liegeois J-P 1994 *Roma, Gypsies, Traveller*, rev. edn. Council for Cultural Cooperation, Strasbourg
- Lucassen L, Willems W, Cottaar A 1998 *Gypsies and Other Itinerant Groups: A Socio-Historical Approach*. St Martins Press, New York
- Mayall D 1995 *English Gypsies and State Policies*. University of Hertfordshire Press, Hatfield, UK
- Muller-Hill B 1988 *Murderous Science: Elimination by Scientific Selection of Jews, Gypsies, and Others, Germany 1933–1945*. Oxford University Press, Oxford, UK
- Puxon G 1987 *Roma: Europe's Gypsies*. Minority Rights Report No. 14, Minority Rights Group, London
- Rao A 1986 *The Other Nomads: Peripatetic Minorities in Cross-Cultural Perspective*. Bohlau Verlag, Köln, Germany
- Salo M 1981 *The American Kalderash: Gypsies in the New World*. Gypsy Lore Society, Hackettstown, NJ
- Stewart M 1997 *The Time of the Gypsies*. Westview Press, Boulder, CO

S. B. Gmelch

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Ghetto

It is a paradox that, while the social sciences have made extensive use of the 'ghetto' as a *descriptive term*, they have failed to forge a robust *analytical concept* of the same. In the historiography of the Jewish diaspora in early modern Europe and under Nazism, the sociology of the black American experience in the twentieth-century metropolis, and the anthropology of ethnic outcasts in Africa and East Asia, its three traditional domains of application, the term 'ghetto' variously denotes a bounded urban ward, a web of group-specific institutions, and a cultural and cognitive constellation (values, mindset, or mentality) entailing the sociomoral isolation of a stigmatized category as well as the systematic truncation of the life space and life chances of its members. But none of these strands of research has taken the trouble to specify what makes a ghetto *qua* social form, which of its features are constitutive and which derivative, as they have, at each epoch, taken for granted and adopted the *folk concept* extant in the society under examination—which explains that the notion, appearing self-evident, does not figure in most dictionaries of social science, including previous publications of this encyclopedia.

1. A Woolly and Shifting Notion

Thus the semantic range of the 'ghetto' in American society and social science, which has dominated inquiry into the topic both quantitatively and thematically, has successively expanded and contracted in keeping with how political and intellectual elites have viewed the vexed nexus of ethnicity and poverty in the city (Ward 1989). At first, in the latter half of the nineteenth century, the term designated residential concentrations of European Jews in the Atlantic seaports and was clearly distinguished from the 'slum' as an area of housing blight and social pathology. It dilated during the Progressive era to encompass all inner-city districts wherein exotic newcomers gathered, namely lower-class immigrants from the Southeastern regions of Europe and African Americans fleeing the Jim Crow regime of caste subjugation in the US South. Expressing ruling-class worries over whether these groups could or should assimilate into the predominant Anglo-Saxon pattern of the country, the term then referred to the intersection between the ethnic neighborhood and the slum, where segregation combined with physical disrepair and overcrowding to exacerbate urban ills such as criminality, family breakdown, and pauperism, and thwart participation in national life. This conception was given scientific authority by the ecological paradigm of the Chicago school of sociology. In his classic book *The Ghetto*, Louis Wirth (1928, p. 6) assimilates to the Jewish ghetto of

medieval Europe the 'Little Sicilies, Little Polands, Chinatowns, and Black Belts in our large cities,' along with the 'vice areas' hosting deviant types such as hobos, bohemians, and prostitutes—all of which are said to be 'natural areas' born of the universal desire of different groups to 'preserve their peculiar cultural forms' and each fulfilling a specialized 'function' in the broader urban organism.

The notion contracted after World War II under the press of the Civil Rights movement to signify mainly the compact and congested enclaves to which African Americans were forcibly relegated as they migrated into the industrial centers of the North. The growth of a 'Black Metropolis in the womb of the white,' wherein Negroes evolved distinct and parallel institutions to compensate for and shield themselves from unflinching exclusion by whites (Drake and Cayton 1945), contrasted sharply with the smooth residential dispersal of European Americans of foreign stock. Writing at the acme of the black uprisings of the 1960s, Kenneth Clark (1965, p. 11) made this relationship of ethnoracial subordination epicentral to his dissection of the *Dark Ghetto* and its woes: 'America has contributed to the concept of the ghetto the restriction of persons to a special area and the limiting of their freedom of choice on the basis of skin color. The dark ghetto's invisible walls have been erected by the white society, by those who have power.' This diagnosis was confirmed by the Kerner Commission (1968, p. 2), a bipartisan taskforce appointed by President Johnson whose official report on the 'civil disorders' that rocked the American metropolis famously warned that, due to white racial intransigence, America was 'moving toward two societies, one black, one white—separate and unequal.' But over the ensuing two decades, the dark ghetto collapsed and devolved into a barren territory of dread and dissolution due to deindustrialization and state policies of welfare reduction and urban retrenchment. And, as racial domination grew more diffuse and diffracted through a class prism, the category was displaced by the duet formed by the geographic euphemism of 'inner city' and the neologism of 'underclass,' defined as the substratum of ghetto residents plagued by antisocial behaviors, acute joblessness, and social isolation (Wilson 1987). By the 1990s, the neutralization of the 'ghetto' in policy-oriented research culminated in the expurgation of all mention of race and power to redefine it as any tract of extreme poverty, irrespective of population and institutional makeup, in effect dissolving the ghetto back into the slum.

The extension of the term to the study of the distinctive sociocultural patterns elaborated by homosexuals in the cities of advanced societies 'in response to both stigma and gay liberation' after the Stonewall riots (Levine 1979, p. 31) and its recent resurgence in Western Europe in heated scientific and political debates over the links between postcolonial

immigration, postindustrial economic restructuring, and urban dualization (Mingione 1996) would seem only to further muddle its meaning. Yet one can extract out of these varied literatures common threads and recurrent properties to *construct a relational concept* of the ghetto as an *instrument of closure and control* that clears up most of the confusion surrounding it and makes it a powerful tool for the social analysis of ethnoracial domination and urban inequality. For this it suffices to return to the historical inception of the word and of the phenomenon it depicted in Renaissance Venice.

2. *A Janus-faced Institution of Ethnic Closure and Control*

Coined by derivation from the Italian *giudecca*, *borghetto* or *gietto* (or from the German *gitter* or the Talmudic Hebrew *get*: the etymology is disputed), the word 'ghetto' initially refers to the forced consignment of Jews to special districts by the city's political and religious authorities. In medieval Europe, Jews were commonly allotted quarters wherein they resided, administered their own affairs, and followed their customs. Such quarters were granted or sold as a privilege to attract them into the towns and principalities for which they fulfilled key roles as money-lenders, tax collectors, and long-distance tradesmen. But, between the thirteenth and the sixteenth century, in the wake of the upheavals caused by the Crusades, favor gradually turned into compulsion (Stow 1992). In 1516, the Senate of Venice ordered all Jews rounded up into the *ghetto nuovo*, an abandoned foundry on an isolated island enclosed by two high walls whose outer windows and doors were sealed while watchmen stood guard on its two bridges and patrolled the adjacent canals by boat. Jews were henceforth allowed out to pursue their occupations by day, but they had to wear a distinctive garb and return inside the gates before sunset on pain of serious punishment. These measures were designed as an alternative to expulsion to enable the city-state to reap the economic benefits brought by the presence of Jews (including rents, special taxes, and forced levies) while protecting their Christian residents from contaminating contact with bodies perceived as unclean and dangerously sensual, carriers of syphilis and vectors of heresy, in addition to bearing the taint of money-making through usury, which the Catholic Church equated with prostitution (Sennett 1994, p. 224).

As this Venetian model spread in cities throughout Europe and around the Mediterranean rim (Johnson 1987, pp. 235–245), territorial fixation and seclusion led, on the one hand, to overcrowding, housing deterioration, and impoverishment as well as excess morbidity and mortality, and, on the other, to institutional flowering and cultural consolidation as

urban Jews responded to multiplying civic and occupational restrictions by knitting a dense web of group-specific organizations that served as so many instruments of collective succor and solidarity, from markets and business associations, to charity and mutual aid societies, to places of religious worship and scholarship. The *Judenstadt* of Prague, Europe's largest ghetto in the eighteenth century, even had its own city hall, the *Rathaus*, an emblem of the relative autonomy and communal strength of its residents, and its synagogues were entrusted not only with the spiritual stewardship but also with the administrative and judicial oversight of its population. Social life in the Jewish ghetto was turned inward and verged 'on overorganization' (Wirth 1928, p. 62), so that it reinforced both integration within and isolation from without.

One can detect in this inaugural moment the four constituent elements of the ghetto, viz., *stigma, constraint, spatial confinement, and institutional encasement*. The ghetto is a social-organizational device that employs space to reconcile two antinomic purposes: to maximize the material profits extracted out of a group deemed defiled and defiling and to minimize intimate contact with its members so as to avert the threat of symbolic corrosion and contagion they carry. This same dual rationale of *economic exploitation cum social ostracization* governed the genesis, structure, and functioning of the African-American ghetto in the Fordist metropolis during most of the twentieth century. Blacks were recruited into northern US cities after World War I because their unskilled labor was indispensable to the industries that formed the backbone of the expanding factory economy. But there was no question of them mixing and consorting with whites, who regarded them as inherently vile, congenitally inferior, and shorn of ethnic honor owing to the stain of slavery. As blacks moved in from the South in millions, white hostility increased and patterns of discrimination and segregation that had hitherto been informal and inconsistent hardened in housing, schooling, and public accommodations and were extended to the economy and polity (Spear 1968, Osofsky 1971). African Americans had no choice but to seek refuge inside the bounded perimeter of the Black Belt and to endeavor to develop in it a network of separate institutions to procure the basic needs of the castaway community. Thus arose a parallel city anchored by black churches and newspapers, black block clubs and lodges, black schools and businesses, and black political and civic associations, nested at the core of the white metropolis yet sealed from it by an impassable fence built of custom, legal suasion, economic discrimination (by realtors, banks, and the state), and by violence manifested in the beatings, fire-bombings, and riots that checked those who dared stray across the color line.

This forced institutional parallelism predicated on enveloping and inflexible spatial seclusion—not extreme poverty, housing blight, cultural difference, or mere residential separation—is what has distinguished African Americans from every other group in US history, as noted by leading students of the black urban experience from W.E.B. Du Bois and E. Franklin Frazier to Drake and Cayton to Kenneth Clark and Oliver Cox (Wacquant 1998). It also characterizes the trajectory of the Burakumin in the Japanese city after the close of the Tokugawa era (Hane 1982). As the descendants of the *eta*, the lowest of the four castes forming the estate order of feudal Japan, the Burakumin were untouchables in the eyes of the Buddhist and Shinto religions, and they were legally confined from sundown to sunup in out-of-the-way hamlets (*buraku*), obliged to wear a yellow collar and to walk barefoot, expected to drop on their hands and knees when addressing commoners, and restricted to marrying solely among themselves. Though they were officially emancipated in 1871, as they moved into cities they were funneled against their will into notorious neighborhoods near garbage dumps, crematoria, jails, and slaughterhouses, widely viewed as nests of criminality and immorality. There, they were barred from industrial employment and locked in low-paying and dirty jobs, sent to separate schools, and compelled to remain endogamous by the indelible taint of their blood as traced through ‘family registration records’ (DeVos and Wagatsuma 1966). In the late 1970s, according to the Burakumin Defense League, they were estimated to number 3 million, trapped in 6,000 ghettos in some thousand cities across the main island.

Spread over three continents and five centuries, the Jewish, African-American, and Burakumin cases demonstrate that the ghetto is not, *pace* Wirth (1928, pp. 284–285), a ‘natural area’ arising via environmental adaptation governed by a biotic logic ‘akin to the competitive cooperation that underlies the plant community.’ The error of the early Chicago school here consisted in falsely ‘converting history into natural history’ and passing ghettoization as ‘a manifestation of human nature’ virtually coterminous with ‘the history of migration’ (Wirth 1928, p. 285) when it is a highly peculiar form of urbanization warped by asymmetric relations of power between ethnoracial groupings: a special form of *collective violence concretized in urban space*. That ghettoization is *not* an ‘uncontrolled and undesigned’ process, as Robert E. Park asserted in his preface to *The Ghetto* (Wirth 1928, p. viii), was especially visible after World War II when the black American ghetto was reconstructed from the top down through state policies of public housing, urban renewal, and suburban economic development intended to bolster the rigid separation of blacks from whites (Hirsch 1983). It is even more glaring in the instance of the ‘caste cities’ built by colonial powers to inscribe in

space the hierarchical ethnic organization of their overseas possessions, such as Rabat under French rule over Morocco and Cape Town after the passage of the Group Areas Acts under the apartheid regime of South Africa (Abu-Lughod 1980, Western 1981).

Recognizing that it is a product and instrument of group power makes it possible to appreciate that, in its full-fledged form, the ghetto is a *Janus-faced institution* as it serves opposite functions for the two collectives that it binds in a relation of asymmetric dependency. For the dominant category, its rationale is to *confine and control*, which translates into what Max Weber calls the ‘exclusionary closure’ of the dominated category. For the latter, however, it is an *integrative and protective device* insofar as it relieves its members from constant contact with the dominant and fosters consociation and community building within the constricted sphere of intercourse that it creates. Enforced isolation from the outside leads to the intensification of social exchange and cultural sharing inside. Ghettos are the product of a mobile and tensionful dialectic of external hostility and internal affinity that expresses itself as ambivalence at the level of collective consciousness. Thus, although European Jews consistently protested relegation within their outcast districts, they were also deeply attached to them and appreciative of the relative security they afforded and the special forms of collective life they supported: Frankfurt’s ghetto in the eighteenth century was ‘not just the scene of confinement and persecution but a place where Jews were entirely, supremely, at home’ (Gay 1992, p. 67). Similarly, black Americans took pride in having ‘erected a community in their own image,’ even as they resented the fact that they had done so under duress, as a result of unyielding white exclusion aimed at warding off the specter of ‘social equality,’ that is, sexual mixing (Drake and Cayton 1945, p. 115).

3. *Disentangling Poverty, Segregation, and Ethnic Clustering*

Articulating the concept of ghetto makes it possible to disentangle the relationship between ghettoization, urban poverty, and segregation, and thence to clarify the structural and functional differences between ghettos and ethnic neighborhoods. It also enables us to spotlight on the role of the ghetto as a symbolic incubator and matrix for the production of a spoiled identity.

1. *Poverty is a frequent but derivative and variable characteristic of ghettos*: the fact that most ghettos have historically been places of endemic and often acute misery owing to the paucity of space, the density of settlement, and the economic exploitation and generalized maltreatment of their residents does not imply that a ghetto is necessarily a place of

destitution, nor that it is uniformly deprived. The *Judengasse* of Frankfurt, instituted in 1490 and abolished in 1811, went through periods of prosperity no less than penury and it contained patches of extraordinary opulence as court Jews helped the city become a vibrant center of trade and finance—part of its glamor to this day comes from it being the ancestral home of the Rothschild dynasty (Wirth 1928, Chap. 4). James Weldon Johnson (1930, p. 4) insisted that the Harlem of the 1930s was ‘not a slum or a fringe’ but the ‘cultural capital’ of black America, where ‘the Negro’s advantages and opportunities are greater than in any other place in the country.’ Likewise, Chicago’s ‘Bronzeville’ was far more prosperous at mid-century than southern black communities and harbored the largest and most affluent African-American bourgeoisie of its era (Drake and Cayton 1945). Whether a ghetto is poor or not depends on extraneous factors such as demography, ecology, state policies, and the shape of the surrounding economy.

Conversely, not all dispossessed and dilapidated urban districts are ghettos. Declining white neighborhoods in the deindustrializing cities of the US Midwest and British Midlands, depressed rural towns of the former East Germany and southern Italy, and the disreputable *villas miserias* of the greater Buenos Aires at the close of the twentieth century are territories of working-class demotion and decomposition, not ethnic containers dedicated to maintaining an outcast group in a relationship of seclusive subordination. They are not ghettos other than in a metaphorical sense, no matter how impoverished—if extreme rates of poverty sufficed to make a ghetto, then large chunks of the former Soviet Union and most third-world cities would be gargantuan ghettos. The *favelas* of the Brazilian metropolis often portrayed as segregated dens of dereliction and disorganization turn out to be working-class wards with finely stratified webs of ties to industry and to the wealthy districts for which they supply household service labor. As in the *ranchos* of Venezuela and the *poblaciones* of Chile, the families that dwell in these squatter settlements span the color continuum and have extensive genealogical bonds to higher-income households; they are ‘not socially and culturally marginal, but stigmatized and excluded from a closed class system’ (Perlman 1976, p. 195; also Quijano 1968). Given that not all ghettos are poor and not all poor areas are ghettos, one cannot collapse and confound the analysis of ghettoization with the study of slums and lower-class districts in the city.

2. Similarly, *all ghettos are segregated but not all segregated areas are ghettos*. The select boroughs of the West of Paris, the exclusive upper-class suburbs of Boston or Berlin, and the ‘gated communities’ that have mushroomed in global cities such as São Paulo, Toronto, and Miami are monotonous in terms of wealth, income, occupation, and often ethnicity, but

they are not for all that ghettos. Segregation in them is entirely voluntary and elective, and for that reason it is neither all-inclusive nor perpetual. Fortified enclaves of luxury package ‘security, seclusion, social homogeneity, amenities, and services’ to enable bourgeois families to escape what they perceive as ‘the chaos, dirt, and danger of the city’ (Caldeira 2000, pp. 264–265). These islands of privilege serve to enhance, not curtail, the life chances and protect the lifestyles of their residents, and they radiate a positive aura of distinction, not a sense of infamy and dread.

This suggests that residential segregation is a necessary but not a sufficient condition for ghettoization. For a ghetto to emerge spatial confinement must, first, be imposed and all-encompassing and, second, it must be overlayed with a distinct and duplicative set of institutions enabling the group thus cloistered to reproduce itself within its assigned perimeter. If blacks are the only ethnic group to be ‘hypersegregated’ in American society (Massey and Denton 1992), it is because they are the only community to have combined involuntary segregation with organizational parallelism entrapping them in a separate and inferior social cosmos of their own, which in turn bolstered their residential isolation. That even involuntary segregation at the bottom of the urban order does not *eo ipso* produce ghettos is demonstrated by the fate of the declining French *banlieues* after the 1980s. Although they have been widely disparaged as ‘ghettos’ in public discourse and their inhabitants share a vivid feeling of being cast out in a ‘penalized space’ suffused with boredom, anguish, and despair (Pétonnet 1982), relegation in these depressed concentrations of public housing at the urban periphery is based on class, not ethnicity; as a result they are culturally heterogeneous, typically harboring native French families along with immigrants from two dozen nationalities; and their inhabitants suffer not from institutional duplication but, on the contrary, from the lack of an ingrown organizational structure capable of sustaining them in the absence of gainful employment and adequate public services. Like the British or Dutch inner cities and the immigrant clusters of urban Germany or Italy, the French *banlieues* are, sociologically speaking, *anti-ghettos* (Wacquant to be published).

3. *Ghettos and ethnic neighborhoods have divergent structures and opposite functions*: moving beyond a gradational perspective to scrutinize the peculiar patterning of social relations within the ghetto as well as between it and the surrounding city throws into sharp relief the differences between the ghetto and the ethnic clusters or immigrant neighborhoods such as newcomers to the metropolis have formed in countless countries. The foreign ‘colonies’ of interwar Chicago that Robert Park, Ernest Burgess, and Louis Wirth—and after them the liberal tradition of assimilationist sociology and historiography—mistook for so many white ‘ghettos’ were scattered and

mobile constellations born of cultural affinity and occupational concentration. Segregation in them was partial and porous, a product of immigrant solidarity and ethnic attraction instead of being imposed by outgroup hostility. Consequently, residential separation was neither uniformly nor rigidly visited upon these groups: in 1930, when the all-black Bronzeville harbored 92% of the city's Afro-American population, Chicago's Little Ireland was 'an ethnic hodge-podge' of 25 nationalities composed of only one-third Irish persons and containing a paltry 3% of the city's denizens of Irish ancestry (Philpott 1978, pp. 141–145).

What is more, the distinctive institutions of European immigrant enclaves were turned outward: they operated to facilitate adjustment to the novel environment of the US metropolis. They neither replicated the organizations of the country of origin nor perpetuated social isolation and cultural distinctiveness, and so they typically waned within two generations as their users gained access to their American counterparts and climbed up the class order and the corresponding ladder of places (Nelli 1970; a similar process of spatial diffusion via class incorporation of Belgian, Italian, Polish, and Iberian immigrants in the French industrial city is reported by Noiriel 1988), all of which is in sharp contrast with the immutable racial exclusivity and enduring institutional alterity of the Black Belt. This Chicago illustration dramatizes the fact that immigrant neighborhood and ghetto serve diametrically opposed functions: one is a springboard for *assimilation* via cultural learning and social-cum-spatial mobility, and the other a material and symbolic isolation ward geared toward *dissimilation*. The former is best figured by a bridge, and the latter by a wall.

4. *A Tainted Identity Machine*

The ghetto is not only the concrete means and materialization of ethnoracial domination through the spatial segmentation of the city but also a potent *collective identity machine* in its own right, for it helps incrustate and elaborate the very division of which it is the expression in two complementary and mutually reinforcing ways. First, the ghetto sharpens the boundary between the outcast category and the surrounding population by deepening the socio-cultural chasm between them: it renders its residents objectively and subjectively more dissimilar from other urban dwellers by submitting them to unique conditionings, so that the patterns of cognition and conduct they fashion have every chance of being perceived by outsiders as singular, exotic, even aberrant (Sennett 1994, p. 244; Wilson 1987, pp. 7–8), which feeds prejudicial beliefs about them. Second, the ghetto is a cultural combustion engine that melts divisions among the confined group and

fuels its collective pride even as it entrenches the stigma that hovers over it. Spatial and institutional entrapment deflect class differences and corrode cultural distinctions within the relegated ethnoracial category. Thus Christian ostracism welded Ashkenazic and Sephardic Jews under an overarching Jewish identity such that they evolved a common 'social type' and 'state of mind' across the ghettos of Europe (Wirth 1928, pp. 71–88, 1964). America's dark ghetto similarly accelerated the sociosymbolic amalgamation of mulattos and Negroes into a single 'race' and turned racial consciousness into a mass phenomenon fueling community mobilization against continued caste exclusion (Drake and Cayton 1945, p. 390).

Yet this unified identity cannot but be stamped with ambivalence as it remains tainted by the very fact that ghettoization proclaims what Weber calls the 'negative evaluation of honor' assigned to the group confined. It is therefore wont to foster among its members sentiments of self-doubt and self-hatred, dissimulation of one's origin through 'passing,' the pernicious derogation of one's kind, and even fantastical identification with the dominant (Clark 1965, pp. 63–67). And, because ghettoization is typically bound up tightly with ethnicity, segregation, and poverty, it is difficult to discern empirically which of the properties exhibited by ghetto dwellers are 'ghetto-specific cultural traits' as opposed to properties expressive of class, community, or masculinity (Hannerz 1969, p. 79). Also, cultural forms forged in the ghetto seep across its boundaries and circulate through the surrounding society where they often become outward signs of cultural rebelliousness and social eccentricity—as indicated by the fascination of bourgeois teenagers around the world for black American 'gangster rap.' This makes it difficult to distinguish between cultural forms effectively in currency among ghetto residents and the public imagery of them diffused in the broader society (including via scholarly writings).

It is fruitful to think of ghetto and ethnic cluster as *two ideal typical configurations situated at opposite ends* of a continuum along which different groups can be located or travel over time depending on the intensity with which the forces of stigma, constraint, spatial confinement, and institutional duplication and completeness coalesce with each other and impinge upon them. Ghettoization can then be turned into a *multilevel variable* for comparative analysis and empirical specification. It can become attenuated to the point where, through gradual erosion of its spatial, social, and mental boundaries, the ghetto devolves into an elective ethnic concentration operating as a springboard for structural integration and/or cultural assimilation into the broader social formation. This describes well the trajectory of the Chinatowns of the United States from the early to the late twentieth century (Zhou 1992) and the status of the

Cuban immigrant enclave of Miami that fostered integration through biculturalism after the Mariel exodus of 1980 (Portes and Stepick 1993). It also characterizes the 'Kimchee Towns' in which Koreans have converged in the metropolitan areas of Japan, which sport a blend of features making them a hybrid between ghetto and ethnic cluster (DeVos and Chung 1981): they are places of infamy that first arose through enmity and constraint, but over the years their population has become ethnically mixed and they have enabled Koreans to socialize and intermarry with Japanese neighbors as well as obtain Japanese citizenship through naturalization. This schema also fits the so-called 'gay ghetto,' which is more aptly characterized as a 'quasi-ethnic community' since 'most gay persons can "pass" and need not be confined to interacting with their "own kind"' and none are forced to reside in the areas of visible concentration of gay institutions (Murray 1979, p. 169).

The double-sidedness of the ghetto as a weapon and shield implies that, to the degree that its institutional completeness and autonomy are abridged, its protective role for the subordinate group is diminished and risks being swamped by its exclusionary modality. In cases where its residents cease to be of economic value to the dominant group, ethnoracial encapsulation can escalate to the point where the ghetto serves as an apparatus merely to warehouse the spoiled group or to prepare it for the ultimate form of ostracization, i.e., physical annihilation. The first scenario fits the evolution of the black American 'hyperghetto' in the post-Civil Rights era: having lost its function of reservoir of unskilled labor power, it has become symbolically linked to the hypertrophied carceral system of the United States by a triple relationship of structural homology, functional surrogacy, and cultural fusion (Wacquant 2004). The second scenario is that implemented by Nazi Germany, which revived the *Judenghetto* between 1939 and 1944, first, to impoverish and concentrate Jews with a view toward relocation and later, when mass deportation turned out to be impractical, to funnel them toward extermination camps (Friedman 1980, Browning 1986).

The unchecked intensification of its exclusionary thrust suggests that the ghetto might be most profitably studied not by analogy with urban slums, lower-class neighborhoods, and immigrant enclaves but alongside the reservation, the refugee camp, and the prison, as belonging to a broader class of institutions for the forced confinement of dispossessed and dishonored groups. It is not by happenstance that the Bridewell of London (1555), the *Zuchthaus* of Amsterdam (1654), and the *Hospital général* of Paris (1656), designed to instill the discipline of wage work to able-bodied vagrants, beggars, and criminals via incarceration, were invented around the same time as the Jewish ghetto.

And that today's sprawling refugee camps in Rwanda, Sri Lanka, and in the occupied territories of Palestine look ever more like a cross between the ghettos of late medieval Europe and gigantic gulags.

Bibliography

- Abu-Lughod J L 1980 *Rabat: Urban Apartheid in Morocco*. Princeton University Press, Princeton, NJ
- Browning C R 1986 Nazi ghettoization policy in Poland, 1939–1941. *Central European History* 19(4): 343–68
- Caldeira T 2000 *City of Walls: Crime, Segregation and Citizenship in São Paulo*. University of California Press, Berkeley, CA
- Clark K B 1965 *Dark Ghetto: Dilemmas of Social Power*. Harper, New York
- DeVos G, Chung D 1981 Community life in a Korean ghetto. In: Lee C, De Vos G (eds.) *Koreans in Japan: Ethnic Conflict and Accommodation*. University of California Press, Berkeley, CA, Chap. 10, pp. 225–51
- DeVos G, Wagatsuma H (eds.) 1966 *Japan's Invisible Race: Caste in Culture and Personality*. University of California Press, Berkeley, CA
- Drake St C, Cayton H R, [1945] 1993 *Black Metropolis: A Study of Negro Life in a Northern City*. University of Chicago Press, Chicago
- Friedman P 1980 The Jewish ghettos of the Nazi era. In: *Roads to Extinction: Essays on the Holocaust*. The Jewish Publication Society of America, New York, Chap. 3, pp. 59–87
- Gay R 1992 *The Jews of Germany: A Historical Portrait*. Yale University Press, New Haven, CT
- Hane M 1982 *Peasants, Rebels, and Outcasts: The Underside of Modern Japan*. Pantheon, New York
- Hannerz U 1969 *Soulside: Inquiries into Ghetto Culture and Community*. Columbia University Press, New York
- Hirsch A 1983 *Making the Second Ghetto: Race and Housing in Chicago 1940–1970*. Cambridge University Press, Cambridge, repub. University of Chicago Press
- Johnson J W [1930] 1981 *Black Manhattan*. Da Capo, New York
- Johnson P 1987 Ghetto. In: *A History of the Jews*. HarperPerennial, New York, pp. 230–310
- Kerner Commission [1968] 1989 *The Kerner Report. The 1968 Report of the National Advisory Commission on Civil Disorders*. Pantheon, New York
- Levine M P 1979 Gay Ghetto. *Journal of Homosexuality* 4(4) (Summer). Reprinted in expanded form as "'YMCA": The Social Organization of Gay Male Life,' In: *Gay Macho: The Life and Death of the Homosexual Clone*. New York University Press, New York, pp. 30–54
- Massey D, Denton N 1992 *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press, Cambridge
- Mingione E (ed.) 1996 *Urban Poverty and the 'Underclass'*. Basil Blackwell, Oxford
- Murray S O 1979 The institutional elaboration of a quasi-ethnic community. *International Review of Modern Sociology* 9 (July): 165–77
- Nelli H S 1970 *Italians in Chicago: A Study in Ethnic Mobility*. Oxford University Press, New York
- Noiriel G 1988 *Le Creuset français*. Editions du Seuil, Paris
- Osofsky G 1971 *Harlem: The Making of a Ghetto—Negro New York, 1890–1930*, 2nd edn. Harper and Row, New York

- Perlman J 1976 *The Myth of Marginality: Urban Poverty and Politics in Rio De Janeiro*. University of California Press, Berkeley, CA
- Pétonnet C 1982 *Espaces habités. Ethnologie des banlieues*. Galilée, Paris
- Philpott T L 1978 *The Slum and the Ghetto: Neighborhood Deterioration and Middle-Class Reform, Chicago 1880–1930*. Oxford University Press, New York
- Portes A, Stepick A 1993 *City on the Edge: The Transformation of Miami*. University of California Press, Berkeley
- Quijano A 1968 *Notas sobre el concepto da marginalidad social*. Commission for Latin American Report, Santiago (Chile)
- Sennett R 1994 Fear of touching. In: *Flesh and Stone: The Body and the City in Western Civilization*. W.W. Norton, New York, Chap. 7, pp. 212–51
- Spear A H 1968 *Black Chicago: The Making of a Negro Ghetto, 1890–1920*. University of Chicago Press, Chicago, IL
- Stow K R 1992 *Alienated Minority: The Jews of Medieval Europe*. Harvard University Press, Cambridge, MA
- Wacquant L 1998 A black city within the white: Revisiting america's dark ghetto. *Black Renaissance* 2(1): 141–51
- Wacquant L 2004 *Deadly Symbiosis: Race and The Rise of Neoliberal Penalty*. Polity Press, Cambridge
- Wacquant L *Urban Outcasts: Toward a Sociology of Advanced Marginality*. Polity Press, Cambridge (to be published)
- Ward D 1989 *Poverty, Ethnicity, and the American City 1840–1925*. Cambridge University Press, Cambridge
- Western J 1981 *Outcast Cape Town*. University of Minnesota Press, Minneapolis
- Wilson W J 1987 *The Truly Disadvantaged: The Inner City, The Underclass and Public Policy*. University of Chicago Press, Chicago
- Wirth L 1928 *The Ghetto*. University of Chicago Press, Chicago
- Wirth L [1956] 1964 The ghetto. In: Reiss A J, Jr. (ed.) *On Cities and Social Life*. The University of Chicago Press, Chicago, Chap. 5 pp. 84–98
- Zhou M 1992 *Chinatown: The Socioeconomic Potential of an Urban Enclave*. Temple University Press, Philadelphia

Loïc Wacquant

Copyright © 2004 Elsevier Ltd.

All rights reserved.

H

Habit: History of the Concept

A family of expressions found in many ordinary languages, habit and its linguistic cognates compose a malleable vocabulary that, for over two millennia, has been incorporated into a wide range of intellectual projects in the domains of the social sciences and beyond. Commonly used for the purpose of describing various *recurrent and relatively automatic features* of human conduct, 'the discourse of habit is one of the great, albeit largely unacknowledged, cultural legacies of Western civilization' (J. M. Thomas 1993, p. 8). Until the early twentieth century, the history of this discourse was roughly coextensive with Western intellectual history itself and scarcely more amenable to brief summary, though broad phases may be distinguished.

1. The Concept of Habit

In the course of its long history, habit has been a polysemous term. Common to most of its meanings and usages, however, has been the conceptualization of habit as a more-or-less self-actuating disposition or tendency to engage in previously adopted or acquired forms of action. When these forms of action are differentiated in terms of their content, distinctions then arise between moral habits, cognitive habits, emotional habits, and motor habits.

Of greater historical importance, however, has been another basis of differentiation. The forms of actions whose repetition is designated as habit may range from simple and circumscribed to complex and generalized. Among the social thinkers who have invoked the concept, habit has most frequently been applied to recurrent patterns of activity of moderate complexity and generality, i.e., to habits of economic, political, religious, and domestic conduct, habits of obedience to rules and to rulers, habits of sacrifice, disinterestedness, and restraint, and so on. But, at various points in the history of the concept, habit has referred primarily to certain simpler, or at least more elemental, behavioral regularities, such as habits of talking, walking, eating, dressing, routine problem-solving, etc.

At the opposite extreme, habit has sometimes been applied to the highly generalized and durable disposition that suffuses an individual's mode of action and discernment in entire spheres of activity or even throughout an entire lifetime. in which case the

concept has been largely synonymous with 'character,' a term encapsulating the whole manner or mold of a human personality. When employed in this extended sense, habit has often been spoken of as *habitus*. That habit has historically taken on this full range of meanings serves as a precaution against projecting on to all instances of the concept connotations only applicable in restricted cases. The contemporary supposition that habit is a fixed, mechanical reaction to discrete stimuli and thus devoid of subjective meaning is, for example, a view that has historically been associated mainly with habit in the sense of simple behavioral regularities, not with more complex habitual forms of action.

At no point has the vocabulary of habit been a free-standing one; to a significant degree, its usage has varied because of the different ways in which habit has been opposed to or aligned with other concepts. At a decisive period in its history, habit was counterposed to instinct, such that all *non*-instinctual behavior was regarded as necessarily habitual. Prior to and subsequent to this period, however, habit was ordinarily set over against purposive, reflective types of action that involve the selection of means and ends according to various normative standards; in this case, habit contrasts with those types of action presupposed by most twentieth-century social-scientific *reflective models* of human conduct. At still other moments, however, habit has been taken not as the opposite of, but as the precondition for reflexive action. The same flexibility is evident with regard to moral conduct, which has sometimes been contrasted with habitual action, other times held to depend upon it. Further, conceptual variability has appeared at the aggregate level, as the similar habits of the several members of a social group have sometime been equated with custom, while other usages reserve custom for practices done in compliance with social sanctions and hence distinct from forms of action that have become somewhat more automatic owing to repetition or other habit-formation processes. (For a fuller treatment of the issues in this section, see Funke 1958, Camic 1986.)

2. Historical Course

2.1 Antiquity to the Early 1800s

The discourse of habit came naturally to Greek philosophers as they opened pathways for later social

thinkers. Aristotle's *Ethics* (Book 2, Chapter 1) (Aristotle [1973]) laid it down that 'none of the moral virtues arise in us by nature,' that 'we are adapted by nature to receive them and are made perfect by habit'; and the same discourse molded as well to the varied teachings of Plato, Plotinus, Quintilian, Aquinas, Luther, Montaigne, Pascal, and Wolf, among a great number of others (Funke 1958).

In the Age of the Enlightenment, amid determined efforts to launch a science of man, the concept of habit was subject to more systematic formulations which reaffirmed habit's vast empirical role but appraised this role differently. In France, Helvetius celebrated habit as the wellspring of public and private morality and accepted it as a 'principle by which [humans everywhere] are actuated' (Helvetius [1758] 1807, p. 57); in Scotland, Ferguson declared habit the force 'by which the good or bad actions of men remain with them and become part of their characters' (quoting Hill 1996, p. 210); while from Prussia Kant averred that 'all acquired habits are objectionable'—that 'virtue is moral strength in the pursuit of one's duty, which should never be a matter of habit' (Kant [1798], 1978, pp. 32–34). Hume, Berkeley, Reid, Condillac, Rousseau, Hegel, and the German idealists were among the many thinkers who took part in the discussion. (On the thinkers named here and others, see Funke 1958, Camic 1986.)

2.2 *A Century of Ferment*

From the middle of the nineteenth century until the early decades of the twentieth century, the language of habit was even more ubiquitous in European and American thought. It found expression in the work of a wide range of social thinkers that included James and John Stuart Mill, Spencer, Bagehot, Marshall, Bradley, and Bosanquet in England, Comte and LePlay in France, and Nietzsche, Simmel, Tönnies, and Husserl in Germany. In the United States, the discourse of habit was also drawn upon extensively by, among others, popular reformers, churchmen, and educators, pragmatist philosophers (Peirce, James, and Dewey), and leading thinkers involved in establishing the social sciences as independent academic disciplines (Sumner, Giddens, Veblen, Thomas).

Typical of the plaint ways in which the discourse was used were the views held by Emile Durkheim and Max Weber as they laid the foundations for modern sociology. Throughout his career, Durkheim took the position that 'habits... are the real forces which govern us' (Durkheim [1905–6] 1956, p. 152). For him, as for the American pragmatists, human reflection is a process that generally 'overloads or paralyzes action'; only 'when habit is disrupted,' and the individual or collective 'being is... at a cross-roads situation [facing] a whole range of possible solutions,' does reflection

arise in order to restore 'equilibrium' and allow 'habits of all kinds' to reassert themselves (Durkheim [1913–14] 1983, pp. 38, 79–80). Accordingly, Durkheim's plans for the moral regeneration of modern society relied heavily on habit: on developing in children 'the habit of self control and restraint' and 'habits of group life'; on cultivating in occupational specialists 'an absolute regularity in habits' and interactional 'habits [that], as they grow in strength, are transformed into rules of conduct'; and on instilling in society's leaders 'a habitus of moral being' (Durkheim 1902–3 [1961], pp. 149, 249, 1893 [1984], pp. 187, 302, 1904–5 [1977], p. 29).

Driven by a different intellectual agenda, Weber held similar views. Occupied with the various forms of reflective action, he nonetheless affirmed the importance of habit, holding that 'the further we go back in history, the more we find that conduct, and particularly social action, is determined in an ever more comprehensive sphere exclusively by the disposition toward the purely habitual': that 'individuals are still markedly influenced' by this force 'even today,' with 'the great bulk of all everyday action [approaching an] almost automatic reaction to habitual stimuli which guide behavior in a course which has been repeatedly followed' (Weber [1922] 1978, pp. 320, 337, 25 [modified translation]). The vocabulary of habit thus figured significantly into Weber's treatment of economic, political, and religious traditionalism, into his writings on economic, political, legal, and communal action in the modern world, and even into his analysis of the process by which Protestantism promoted the spirit of modern capitalism. Conceptualizing the spirit of capitalism as a 'particular habitus,' Weber credited the early Calvinists with creating, 'out of their religiously conditioned family traditions and from the religiously influenced life-style of the environment, [a distinctive] 'habitus' among individuals which prepared them in specific ways to live up to the specific demands of early modern capitalism' (Weber [1910a] 1968, [1910b] 1978, p. 1124).

Yet, as social thinkers mobilized the language of habit in these ways, contemporary natural scientists increasingly moved to appropriate it to refer to a far more elementary range of activities, both human and subhuman. This, for example, was the practice of Lamarck, and Darwin followed the model, regularly speaking of the flowering habits of plants, the feeding habits of insects, and the flying habits of pigeons (Darwin [1859] 1964, pp. 11, 183, [1872] 1975, pp. 29–31). This was also the practice of many European physiologists and psychologists as they turned attention to the study of the neural foundations of simple animal and human reflexes, widely using the term habit to describe well-established neural pathways (Fearing 1930). This usage gained great currency among late-nineteenth and early-twentieth century American psychologists, convinced with James that even 'the most complex habits [are] nothing but

concatenated discharges in the nerve-centres, due to the presence there of systems of reflex paths' (James [1890] 1950, p. 108). Extending the same view into their own domain, turn-of-the-century American social scientists ascribed vital significance to habits rooted in 'the physiology of the nervous system,' convinced 'that education and social control are largely dependent upon our ability to establish habits in ourselves and others' (quoting Park [1915] 1969, see also Camic 1986).

2.3 The Decline of Habit

The culmination of these developments came with the behaviorist revolution which John Watson launched in American psychology during World War One and which overran the field in the next two decades. Taking the position that *every* 'mode of acting... not belonging to man's hereditary equipment must be looked upon as a habit,' conceived physiologically in terms of 'muscular and glandular changes which follow upon a given stimulus,' Watson and his followers touted habit-psychology as the way 'to guide society... towards the control of group and individual behavior' (Watson 1919, pp. 270, 14, 2–3). From this point of view, the science required for the study of the social world was 'not [a social science like] sociology, but psychology,' which derives from 'biology, chemistry, and the other natural sciences' (Allport 1927, pp. 167–168).

The challenge to the legitimacy of their own fledgling academic disciplines now manifest, social scientists swiftly recoiled from the language of habit. Countering the behaviorist effort to eliminate the reflective forms of action that had long stood alongside habitual forms, sociologists of the inter-war period led an aggressive campaign to rescue purposive human conduct, deliberately forsaking their previous use of the concept of habit. This campaign was inaugurated by Thomas (a former proponent of the concept) and Znaniecki, whose work attacked the 'behavioristic school' for its 'indistinct [application] of the term 'habit' to [all] uniformities of behavior' and urged that 'habit'... be restricted to the biological field, [since] it involves no conscious, purposeful regulation of [conduct], but merely... is unreflective, [whereas] the uniformity [that constitutes social life] is not a uniformity of organic habits but of consciously followed rules' (W. I. Thomas and Znaniecki [1918] 1958, pp. 1849–1852). Within a generation, this view became dogma, displacing the discourse of habit with the language of 'attitude,' which Thomas and Znaniecki brought to the fore and which subsequently furnished American social scientists with a new vocabulary with which to encompass regular tendencies in human conduct while holding to reflective action models (Fleming 1967, Camic 1986). Parallel movements away from the concept of habit occurred in European thought (as

even Durkheim and Weber excised the concept from sociology) and in European and American literature (Fisher 1973, J. M. Thomas 1993). In the aftermath of this turnabout, at least outside the natural sciences, the language of habit became dormant for the half century from the late 1920s to the late 1970s, with very few exceptions. In the heyday of the behaviorist revolution, Dewey resuscitated a broader conception of habit, and Merleau-Ponty in France subsequently did likewise (Kestenbaum 1977, see also Berger and Luckmann 1966); but such efforts did not stem the behaviorist view of habit, nor the resulting widespread reaction against the concept. Only among psychologists working in the behaviorist tradition, and among a few economists focused on consumer behavior, did the language of habit remain a presence (Deaton and Muelbaur 1980, Smith 1986). Otherwise, social scientists essentially abandoned the discourse, conceiving of all human action as a reflective process of fitting means to ends according to various normative standards.

3. Recent Developments

The last quarter of the twentieth century brought little fundamental alteration in the stance that had been taken toward the language of habit during the previous half century. By and large, reflective conceptualizations of human action have continued to dominate the social sciences. This dominance has been less complete, however, as the concept of habit, understood in non-behaviorist terms that simultaneously accommodate reflective action, has periodically been revived: by social psychologists (Ouellette and Wood 1998), by a wider range of economists (Scitovsky 1976, Campbell and Cochrane 1995), and by sociological theorists, some developing earlier pragmatist and phenomenological approaches (Ostrow 1990, Joas [1992] 1996), others aligned also with the traditions of Durkheim and Weber, most notably Bourdieu ([1980] 1990), in whose writings 'habitus' receives extended attention (see also Giddens 1984). Still largely disconnected from one other, these lines of work open promising future possibilities for the eventual transformation of exclusively reflective models of human conduct in the social world.

See also: Habituation; Tradition, Anthropology of; Traditions in Sociology; Traditions: Social; Weber, Max (1864–1920)

Bibliography

- Allport G W 1927 The nature of institutions. *Social Forces* 6: 167–79
 Aristotle [1973] McKeon R (ed.) *Introduction to Aristotle*, 2nd edn. University of Chicago Press, Chicago

- Berger P L, Luckmann T 1966 *The Social Construction of Reality*. Doubleday, New York
- Bourdieu P [1980] 1990 *The Logic of Practice*. Stanford University Press, Stanford, CA
- Camic C 1986 The matter of habit. *American Journal of Sociology* **91**: 1039–87
- Campbell J Y, Cochrane J H 1995 *By Force of Habit*. National Bureau of Economic Research, Cambridge, MA
- Darwin C [1859] 1964 *On the Origin of Species*. Harvard University Press, Cambridge, MA
- Darwin C [1872] 1975 *The Expression of Emotions in Man and Animals*. University of Chicago Press, Chicago
- Deaton A, Muelbaur J 1980 *Economics and Consumer Behavior*. Cambridge University Press, New York
- Durkheim E [1893] 1984 *The Division of Labor in Society*. Free Press, New York
- Durkheim E [1902–3] 1961 *Moral Education*. Free Press, New York
- Durkheim E [1904–5] 1977 *The Evolution of Educational Thought*. Routledge and Kegan Paul, London
- Durkheim E [1905–6] 1956 *Education and Society*. Free Press, Glencoe, IL
- Durkheim E [1913–4] 1983 *Pragmatism and Sociology*. Cambridge University Press, Cambridge, UK
- Fearing F 1930 *Reflex Action*. Williams and Williams, Baltimore, MD
- Fisher P 1973 The failure of habit. In: Engel M (ed.) *Uses of Literature*. Harvard University Press, Cambridge MA, pp. 3–18
- Fleming D 1967 Attitude: History of a concept. *Perspectives in American History* **1**: 287–365
- Funke G 1958 Gewohnheit. *Archive für Begriffsgeschichte* **3**: 1–606
- Giddens A 1984 *The Constitution of Society*. University of California Press, Berkeley, CA
- Helvetius C A [1758] 1807 *De l'Esprit*. Jones, London
- Hill L 1996 Anticipations of nineteenth and twentieth century social thought in the work of Adam Ferguson. *Archives européennes de sociologie* **37**: 203–28
- James W [1890] 1950 *The Principles of Psychology*. Dover, New York
- Joas H [1992] 1996 *The Creativity of Action*. University of Chicago Press, Chicago
- Kant I [1798] 1978 *Anthropology from a Pragmatic Point of View*. Southern Illinois University Press, Carbondale, IL
- Kestenbaum V 1977 *The Phenomenological Sense of John Dewey*. Humanities Press, Atlantic Highlands, NJ
- Ostrow J M 1990 *Social Sensitivity*. State University of New York Press, Albany, NY
- Ouellette J A, Wood W 1998 Habit and intention in everyday life. *Psychological Bulletin* **124**: 54–74
- Park R E [1915] 1969 Man not born human. In: Park R E, Burgess E W (eds.) *Introduction to the Science of Sociology*. University of Chicago Press, Chicago, pp. 82–4
- Scitovsky T 1976 *The Joyless Economy*. Oxford University Press, New York
- Smith L D 1986 *Behaviorism and Logical Positivism*. Stanford University Press, Stanford, CA
- Thomas J M 1993 Figures of habit in William James. *The New England Quarterly* **66**: 3–26
- Thomas W I, Znaniecki F [1918] 1958 *The Polish Peasant in Europe and America*. University of Chicago Press, Chicago
- Watson J B 1919 *Psychology from the Standpoint of a Behaviorist*. Lippincott, Philadelphia

- Weber M [1910a] 1968 Antikritisches zum 'Geist' des Kapitalismus. In: Winkelmann J (ed.) *Die Protestantische Ethik II*. Siebenstern, Munich, pp. 149–87
- Weber M [1910b] 1978 Anticritical last word on 'The Spirit of Capitalism.' *American Journal of Sociology* **83**: 1105–31
- Weber M [1922] 1978 *Economy and Society*. University of California Press, Berkeley, CA

C. Camic

Habituation

If a drop of water falls on the surface of the sea just over the flower-like disc of a sea anemone, the whole animal contracts vigorously. If, then, a second drop falls within a few minutes of the first, there is less contraction, and finally, on the third or fourth drop, the response disappears altogether. Here in this marine polyp with the primitive nerve net is clearly exhibited one of the most pervasive phenomena of the animal kingdom—decrement of response with repeated stimulation. Almost every species studied, from amoeba to man, exhibits some form of response decrement when the stimulus is frequently repeated or constantly applied. The ubiquity of the phenomenon plus its obvious survival value suggests that this kind of plasticity must be one of the most fundamental properties of animal behaviour. (Sharpless and Jasper 1956, p. 655)

With these words, Sharpless and Jasper introduced their now classic study of habituation of the electroencephalogram (EEG) arousal response. Their study had a profound impact on behavioral neuroscience and served to kindle the interest of neural scientists in brain mechanisms of habituation. *Habituation* is a decrease in response to repeated stimulation, and *sensitization* is an increase in response as a result of stimulation. Virtually all behavioral responses, as well as autonomic and EEG responses and signs of arousal, habituate rapidly and to a profound degree. The processes responsible for habituation of behavioral responses probably occur largely after the primary sensory pathways of the brain and before the motor neurons that move muscles—somewhere between initial processing of sensory information and the execution of decisions to respond. Sensitization, the increase in response that often occurs when strong stimuli are given, may be similarly localized, although some evidence indicates that sensitization may occur in motor systems and motor neurons.

In early work on brain processes of habituation, it was thought that virtually all measures of neural activity might habituate, ranging from first-order sensory nuclei to the cerebral cortex. An early experiment to report evoked response habituation was that of Hernández-Peón et al. (1957); they recorded responses to click stimulation at several levels of the

auditory system. Trains of clicks were delivered once every two seconds for long periods of time and evoked responses of the cochlear nucleus (the first relay of the auditory system) were reported to habituate.

Careful studies by Worden and associates (see Marsh et al. 1962) demonstrated that click-evoked responses at the cochlear nucleus do not show habituation. Instead, amplitudes of responses at this first relay nucleus in the auditory system are rigidly controlled by the physical properties of the sound stimulus. Because of acoustic factors, the intensity of a sound is often weaker at the floor of a test cage. If an animal gradually became bored and rested his head on the floor, the cochlear nucleus-evoked response to click would decrease because of reduced sound intensity.

In an extensive and careful study, Wickelgren (1968) found that habituation of evoked responses to auditory stimuli was not prominent below the level of the thalamus and cortex in the auditory system. This appears to be a general finding in all sensory systems—the lower ascending relay nuclei do not exhibit habituation. It would be most surprising if there were marked habituation in sensory systems. If our sensory systems habituated to sensory stimuli, we would rapidly cease to have sensation—a most unadaptive state of affairs. However, behavioral responses and phenomena that seem to be related to the probability of behavioral responding, such as EEG arousal and autonomic signs of orienting, show marked habituation. It is very adaptive to cease responding to arousing stimuli that have no particular significance.

Historically, such processes as habituation and sensitization were viewed simply as reflexes and not as phenomena of learning and memory. However, the broad definition of learning as a change in behavior as a result of experience certainly encompasses both phenomena. When this was recognized, analysis of neural mechanisms became 'legitimized' (Sharpless and Jasper 1956, Thompson and Spencer 1966). Habituation and sensitization are now seen as examples of nonassociative learning and memory. Greater progress has been made in analysis of the basic neuronal mechanisms underlying habituation than in more complex forms of behavioral plasticity such as classical and instrumental conditioning. Major reasons for this progress lie in the widespread occurrence of habituation, in its simplicity, and in the reliability with which the behavioral properties of habituation can be specified and determined. Habituation occurs both in spinal reflexes of the vertebrate (Sherrington 1906) and for responses of simple animals, as well as in complex human behavior. Consequently, neural mechanisms of habituation can be studied in simplified systems.

As noted earlier, behavioral habituation is simply response decrement as a result of repeated stimulation. For example, if a moderate electric shock is delivered to the foot of a normal but restrained animal for the

first time, the animal will show marked struggling and leg flexion movements. However, if the stimulus is repeated a few times, the responses decrease considerably and may even disappear. Habituation is commonly differentiated from response decrements due to trauma, growth, aging, etc. In the example above, if the shock to the foot is withheld for a period of time after habituation and then given again, the response will reappear with normal strength. This spontaneous recovery is a common control used to show that habituation is a *reversible* response decrement.

Response decrements due to altered sensitivity of receptors are often termed *receptor adaptation*, and response decrements due to decremental effects occurring at the neuromuscular junctions of muscles are often termed *muscular fatigue*. Neither of these processes is considered to be habituation; it is a central rather than a peripheral process. Almost all behavioral responses of the intact organism initially elicited by a given stimulus, ranging in complexity from leg flexion to exploratory and play behavior, show habituation with repeated stimulation.

Response sensitization may be defined simply as an increased response strength to a given stimulus as a result of some other, usually strong, stimulus. The galvanic skin reflex (GSR) is a good case in point. Suppose we were to measure the GSR activity of sweat glands in the palm of the hand in response to a sound of moderate intensity. The response would probably be of relatively low amplitude and somewhat irregular in occurrence. If a strong shock were delivered to the skin, the sound alone now would evoke a much larger GSR response; the response would have been sensitized. The response increase due to sensitization is distinguished from conditioning in that no pairing of shock and sound is necessary to produce the effect. Sensitization itself typically habituates; after a number of shock presentations the response to the sound decreases to the initial weak level. In fact, it seems to decay spontaneously following a sensitizing stimulus.

The study of habituation has a long history; some of its basic properties were described in classic works by Jennings (1906), Humphrey (1933), Prosser and Hunter (1936), and Harris (1943). Thompson and Spencer (1966) surveyed the by then very extensive behavioral literature on habituation and identified some nine basic parametric properties or characteristics exhibited by behavioral habituation.

1. Properties of Habituation

(a) Given that a particular stimulus elicits a response, repeated applications of the stimulus result in decreased response (habituation). The decrease is usually a negative exponential function of the number of stimulus presentations.

(b) If the stimulus is withheld, the response tends to recover over time (spontaneous recovery).

(c) If repeated series of habituation training and spontaneous recovery are given, habituation becomes successively more rapid (this might be called potentiation of habituation).

(d) Other things being equal, the more rapid the frequency of stimulation, the more rapid and/or more pronounced is habituation.

(e) The weaker the stimulus, the more rapid and/or more pronounced is habituation. Strong stimuli may yield no significant habituation.

(f) The effects of habituation training may proceed beyond zero or asymptotic response level.

(g) Habituation of response to a given stimulus exhibits stimulus generalization to other stimuli.

(h) Presentation of another (usually strong) sensitizing stimulus results in recovery of the habituated response (dishabituation or sensitization).

(i) Upon repeated application of the sensitizing stimulus, the amount of dishabituation produced habituates (this might be called habituation of sensitization).

It is important to emphasize that these characteristics of habituation are simply the empirically observed properties in a wide variety of behavioral studies on many species and as such do not imply particular theories or inferences. These properties have come to serve as the detailed definition of habituation. Thus, any particular stimulus-response system is said to exhibit habituation if the response decrement to repeated stimulation exhibits these properties.

The most potent variables determining the course of habituation are stimulus frequency and intensity. The frequency range must be behaviorally relevant. Very high frequencies can yield fatigue, potentiation, and other processes. The range of stimulus frequencies over which habituation occurs, i.e., behavioral response decrements exhibiting the properties noted above, is an empirical question dependent upon the organism, the stimulus, and the response system engaged. For the orienting or arousal response, the effective range may be from minutes to hours to days; for the spinal flexion reflex, the range is over seconds.

The fact that rate and degree of habituation are inversely related to stimulus intensity (property (e)) is true if the same intensity is used for test and habituating stimuli (the test stimulus is given sufficiently infrequently that it does not by itself cause habituation). However, if the test and habituating stimuli differ in intensity, more complex effects can occur (Davis and Wagner 1968).

2. Dishabituation vs. Sensitization

Thompson and Spencer (1966) employed habituation of the spinal flexion reflex, first used as a model system to study habituation by Prosser and Hunter (1936), to

characterize the detailed properties of habituation and analyze possible neuronal mechanisms. They discovered that the process of 'dishabituation' caused by presentation of a strong extra stimuli was not in fact dishabituation. That is, it did not disrupt the process of habituation at all. Instead, it added a superimposed, independent process of increased excitability, namely sensitization. The fact that dishabituation, *per se*, does not appear to exist as such, but rather reflects a separate process of sensitization, appears to hold for all vertebrate preparations studied, ranging from reflexes to the startle response to EEG arousal and the orienting response. One exception to this rule has been reported for a monosynaptic system in the invertebrate *Aplysia* (Marcus et al. 1988). However, the neuronal mechanisms underlying the processes of sensitization/dishabituation appear to differ in vertebrates vs. the *Aplysia* circuit. In the latter, 'dishabituation' appears to be a presynaptic process, occurring at the sensory neuron terminals on the motor neurons (see Sect. 3 and Kandel 1976); whereas, at least in spinal reflexes, sensitization can involve increased excitability in motor neurons (Thompson and Spencer 1966). Actually, recent evidence suggests that changes may occur in motor neurons in the *Aplysia* circuit as well (Murphy and Glanzman 1997).

The fact that dishabituation does not exist as such, but rather reflects a separate independent process of sensitization in vertebrate preparations led Groves and Thompson (1970) to develop the dual-process theory of habituation: the outcome of repeated stimulus presentations is due to the net interaction of the inferred processes of habituation and sensitization. This theory is able to account for an astonishingly wide range of behavioral phenomena (see, e.g., Peeke and Herz 1973).

A somewhat differently construed and widely influential theory of habituation is the stimulus-model comparator theory developed in Russia by Evgeny Sokolov (1963), based largely on data from the human orienting response. In this view, repetition of a given stimulus results in the development of a neuronal 'model' of the stimulus in forebrain structures. If a repetitive stimulus matches the model, then there is little output, only the habituated response, which could be no response, i.e., no EEG arousal. However, if there is a mismatch, then the difference is amplified via the reticular formation and a larger response results.

The dual-process theory is more successful in accounting for basic phenomena of habituation and sensitization in a wide range of species; the model-comparator theory can more easily account for more complex phenomena like the 'missing stimulus' effect (increased response to the absence of one presentation of the habituating stimulus, a phenomenon sometimes seen in EEG arousal but not in reflex systems). But note that a process of synaptic depression could underlie the formation of the neuronal model of the

stimulus in the model-comparator theory; the theories are at different levels and not incompatible. A view somewhat similar to that of Sokolov was developed by Wagner (1979) to account for long-term habituation. Thus, the central model of the habituating stimulus is considered as a relatively long-term memory trace, yielding the basic properties of a learned response. In their classic study of human EEG arousal habituation, Sharpless and Jasper (1956) stressed the distinction between short-term and long-term habituation.

As noted earlier, primary sensory afferent system neurons do not exhibit habituation. This makes perfect sense; in order to 'decide' how much to respond to a stimulus it is necessary to know what the stimulus is. Virtually all behavioral responses elicited by stimuli show habituation to repeated stimulation under at least some conditions. Central brain processes that are behavior-related like cortical EEG arousal show clear habituation, as do brain motor systems. Davis and associates completed a comprehensive analysis of the startle circuit and startle habituation in the rat (Davis 1984). The circuit includes auditory relay nuclei: the cochlear nuclei and the nuclei of the lateral lemniscus; a nucleus in the brainstem reticular system (nucleus reticularis pontis caudalis) and its projections to brainstem and spinal cord premotor and motor nuclei. Evidence suggests that habituation to the acoustic startle stimulus occurred before the motor systems, i.e., at the nucleus reticularis pontis caudalis (Davis et al. 1982). The reticular nucleus is a 'bridge' nucleus between sensory and motor systems.

3. Neural Mechanisms

Unlike associative learning, in habituation the stimulus-response neuronal circuit is hardwired. In systems where some degree of analysis of neuronal mechanisms is possible, the neural substrate of habituation occurs within the stimulus-response circuit itself and is not due to any extrinsic circuit action like external inhibition acting on the habituating circuit.

Greatest progress in analyzing neural mechanisms of habituation and sensitization has come from simplified neuronal systems. Two of these are the gill withdrawal reflex of the mollusk *Aplysia* and vertebrate spinal reflexes. Kandel (1976) and his many associates completed a detailed analysis of the processes of habituation and sensitization in the *Aplysia* monosynaptic circuit. They were able to show that habituation was a presynaptic process involving a decrease in the probability of neurotransmitter release as a result of repeated activation. Similarly, a process of synaptic depression is responsible for habituation of the tail-flip escape response in the crayfish, another very productive invertebrate model system (Krasne 1969). In the *Aplysia* model, synaptic decrement is due to a persisting decrease in calcium ion influx into the terminals. Sensitization, elicited for example by strong

shock to the tail of the animal, was due to activation of a polysynaptic circuit that acted presynaptically on the sensory neuron terminals, the same terminals that habituated. The effects on calcium channels are due in turn to complex and not fully understood biochemical cascades within the neurons. Kandel and associates have also described a long-term sensitization effect involving structural changes in the synapses that they have used as a model for long-term memory storage (Hawkins and Kandel 1984).

In the vertebrate spinal cord, the flexion reflex is polysynaptic. Thompson and Spencer (1966) were able to rule out any changes in excitability in the sensory neuron terminals and in the motor neurons themselves during habituation. However, sensitization did involve increases in motor neuron excitability, as noted above. They suggested, but could not prove, that a process of low-frequency depression in the interneurons was responsible for habituation, a process analogous to the presynaptic depression process demonstrated by Kandel and associates in *Aplysia*. Using a monosynaptic pathway in the isolated frog spinal cord, Farel and Thompson (1976) demonstrated a presynaptic process for habituation. The story concerning neuronal mechanisms of habituation is particularly satisfying because the same process of synaptic depression appears to underlie habituation in both invertebrate and vertebrate systems. This simple process of synaptic decrement can account for a wide range of behavioral phenomena of habituation.

The study of habituation continues to be a most active field of research, ranging from responses of nematode *C. elegans* (Wicks and Rankin 1996) to the startle response in mammals (Pilz and Schnitzler 1996) to behaviors in the human fetus and infant (Leader 1995, Kaplan et al. 1990).

See also: Arousal, Neural Basis of; Classical Conditioning, Neural Basis of; Conditioning and Habit Formation, Psychology of; Eyelid Classical Conditioning; Learning and Memory: Computational Models; Learning and Memory, Neural Basis of

Bibliography

- Davis M 1984 The mammalian startle response. In: Eaton R C (ed.) *Neural Mechanisms of Startle Behavior*. Plenum, New York, pp. 287-351
- Davis M, Parisi T, Gendelman D S, Tischler M D, Kehne J H 1982 Habituation and sensitization of electrically elicited 'startle' reflexes. *Science* **218**: 688-9
- Davis M, Wagner A R 1968 Startle responsiveness after habituation to different intensities of tone. *Psychonomic Science* **12**: 337-8
- Farel P B, Thompson R F 1976 Habituation of a monosynaptic response in frog and spinal cord: Evidence for a presynaptic mechanism. *Journal of Neurophysiology* **39**: 661-6
- Groves P M, Thompson R F 1970 Habituation: A dual-process theory. *Psychological Review* **77**: 419-50

- Harris J D 1943 Habituation response decrement in the intact organism. *Psychological Bulletin* **40**: 358–422
- Hawkins R D, Kandel E R 1984 Is there a cell-biological alphabet for simple forms of learning? *Psychological Review* **91**: 373–91
- Hernandez-Peon R, Jouvett M, Scherrer H 1957 Auditory potentials at cochlear nucleus during acoustic habituation. *Acta neurologica latinoamericana* **3**: 144–56
- Humphrey G 1933 *The Nature of Learning*. Harcourt, Brace, New York
- Kandel E R 1976 *Cellular Basis of Behavior: An Introduction of Behavioral Biology*. W. H. Freeman, San Francisco
- Kaplan P S, Werner J S, Rudy J W 1990 Habituation, sensitization and infant visual attention. In: Rovee-Collier C, Lipsitt L P (eds.) *Advances in Infancy Research*. Ablex Publishing, Norwood, NJ, Vol. 6, pp. 61–109
- Krasne F B 1969 Excitation and habituation of the crayfish escape reflex: The depolarization response in lateral giant fibers of the isolated abdomen. *Journal of Experimental Biology* **50**: 29–46
- Leader L R 1995 The potential value of habituation in the prenat. In: Lecanuet J-P, Fifer W P, Krasnegor N A, Smotherman W P (eds.) *Fetal Development: A Psychobiological Perspective*. Lawrence Erlbaum, Hillsdale, NJ, pp. 383–404
- Marcus E A, Nolen T G, Rankin C H, Carew T J 1988 Behavioral dissociation of dishabituation, sensitization, and inhibition of *Aplysia*. *Science* **241**: 210–13
- Marsh J T, Worden F G, Hicks L 1962 Some effects of room acoustics on evoked auditory potentials. *Science* **137**: 280–2
- Murphy G G, Glanzman D L 1997 Mediation of classical conditioning in *Aplysia Californica* by LTP of sensorimotor synapses. *Science* **278**: 467–71
- Peeke H V S, Herz M J (eds.) 1973 *Habituation, Volume I: Behavioral Studies, Volume II: Physiological Studies*. Academic Press, New York
- Pilz P D, Schnitzler H-U 1996 Habituation and sensitization of the acoustic startle response in rats: Amplitude, threshold, and latency measures. *Neurobiology of Learning and Memory* **66**: 67–79
- Prosser C L, Hunter W S 1936 The extinction of startle responses and spinal reflexes in the white rat. *American Journal of Physiology* **41**: 609–18
- Sharpless S, Jasper H H 1956 Habituation of the arousal reaction. *Brain* **79**: 655–80
- Sherrington C S 1906 *The Integrative Action of the Nervous System*. Yale University Press, New Haven, CT
- Sokolov Y N 1963 *Perception and the Conditioned Reflex*. Trans. Waydenfeld S W, Pergamon Press, Oxford, UK
- Thompson R F, Spencer W A 1966 Habituation: A model phenomenon for the study of neuronal substrates of behavior. *Psychological Review* **73**: 16–43
- Wagner A R 1979 Habituation and Memory. In: Dickinson A, Boakes R A (eds.) *Mechanisms of Learning and Motivation: A Memorial Volume for Jerry Konorski*. Lawrence Erlbaum, Hillsdale, NJ, pp. 53–82
- Wickelgren W O 1968 Effects of acoustic habituation on click-evoked responses in rats. *Journal of Neurophysiology* **31**: 777–84
- Wicks S R, Rankin C H 1996 Recovery from habituation in *Caenorhabditis elegans* is dependent on interstimulus interval and not habituation kinetics. *Behavioral Neuroscience* **110**: 840–4

R. Thompson

Halévy, Élie (1870–1937)

Élie Halévy was a French historian and philosopher who made three principal contributions to the social sciences: a study of the English Utilitarians; a historical account of English society in the eighteenth and nineteenth centuries which is noteworthy for the role it attributes to Methodism and Evangelicalism in maintaining political stability; and an analysis of the development of European socialism culminating in what he designated an ‘era of tyrannies.’

1. Biography

Élie Halévy was born at Étretat to wealthy, artistic parents on September 6, 1870, which marked a turbulent period in French history, with the origins of the Third Republic, France’s defeat by Prussia, and the collapse of the Paris Commune. His father, of Jewish descent, was a successful librettist for such composers as Bizet and Offenbach, while his mother was of Swiss Calvinist origins. Both Élie and his brother Daniel were brought up as Protestants under her strong moral influence: like Élie’s contemporary Max Weber (see *Weber, Max (1864–1920)*) there was a strong familial influence on the study of religious—and especially Christian Protestant—ideas. Halévy’s grandfather had been a disciple of Saint-Simon and succeeded Comte as his secretary for a time. Though the family had important connections in French political life, Élie remained distant from, but also keenly concerned with, political affairs.

While still a student at the *Lycée Condorcet* he planned, with his brother and other associates, the *Revue de Métaphysique et de Morale*, a journal with which he was associated for the whole of his life. At the *École Normale Supérieure* he studied philosophy, completing a thesis on Platonism. Already his political and moral principles had become established as liberal, individualist, rationalist, and secular—principles that were to characterize him for the rest of his career. In 1898 Halévy was invited to lecture at the *École Libre des Sciences Politiques*, which had been founded after France’s defeat in 1871 with the goal of diagnosing French political instability, and he remained there all his professional life. The *École Libre* was much admired in Britain by the group of Fabians and their associates who founded the London School of Economics and Political Science, and Halévy came to know them well as he moved in British intellectual circles.

The Dreyfus affair and the defense of the Third Republic were issues which united Halévy with Émile Durkheim (see *Durkheim, Emile (1858–1917)*), though on a philosophical level he rejected Durkheim’s positivism and collectivism. On the other hand he attributed his initial interest in socialism, to which he

devoted half of his scholarly life, to Durkheim's lectures on socialism at Bordeaux in 1895. The affinity that Halévy perceived between communist and fascist régimes—since he viewed 1914 as the start of an 'era of tyrannies'—was an unpopular stance in the intellectual climate of the interwar years, but he persisted in his opinion until his death at Sucy-en-Brie on August 21, 1937. The enthusiasm of the Webbs and Shaw for fascist or communist experiments in no way surprised him, and he ended his life firmly committed to the liberal, rational, democratic ideals of his youth.

2. Intellectual Contributions

In his thesis, published in 1896 as *La théorie platonicienne des sciences*, Halévy explicated a dialectical approach, which he labeled Platonic, by means of which the contradictions in a subject were stated as systematic dichotomies. By exposing the core assumptions and practical implications behind a doctrine or belief system, which usually involved logical conflicts, his own analysis of it could proceed. This fundamental principle of methodology was deployed in his subsequent studies of Bentham's thought, of Wesleyan doctrine, and finally of socialism. In his study of the English utilitarians he found just such a contradiction between Bentham's political and economic theories and argued that 'the contradiction was continually breaking out in the current formulae of Benthamism' (Halévy 1952, p. 489). In his political theory, Bentham postulated the necessity of a legislator who would reconcile conflicts of interest and thus secure the greatest good of the greatest number: this he called the artificial identity of interests. In his economic theory, however, Bentham assumed with Adam Smith (see *Schumpeter, Joseph A (1883–1950)*) and his successors that there was a spontaneous identity of interests and that individual interests were naturally harmonious. Halévy's book is a reminder that the Bentham who designed the *Panopticon*—'a new application of the principle of the artificial identification of interests, the idea of which he had found in Helvetius' (Halévy 1952, p. 82)—also propagated a belief in *laissez-faire* and the 'hidden hand.'

Soon after the completion of *The Growth of Philosophic Radicalism*, Halévy's interest turned to Methodism in eighteenth- and nineteenth-century England. Sometimes regarded as the complete antithesis of utilitarianism, he regarded Methodism as an ally of utilitarianism in shaping the social morality of modern England. In 1905 he traveled to England to work on material relating to the early history of Methodism and it was at this time that he expressed his intention of writing a history of England in the nineteenth century. The Methodism project had a more immediate outcome, however, in two articles on *La naissance du Méthodisme en Angleterre* in the *Revue de Paris* in 1906. With very few exceptions these articles

went unremarked by historians until their translation into English in 1971 (Halévy 1971).

The fundamentals of Halévy's interpretation of the role of Methodism in *La naissance* correspond with his later treatment in *A History of the English People* in 1815 (Halévy 1924)—the first volume of his five-volume *A History of the English People* which was published in French between 1912 and 1926 and in English between 1912 and 1932—but the focus and detail of the earlier work provide a particularly strong statement of the 'Halévy thesis.' He traces the earliest development of manufacturing industry in the first three decades of the eighteenth century and points to the growth of pauperism in the industrial centers. By the late 1730s this was compounded by a commercial crisis of overproduction accompanied by working class agitation, and it was this that stimulated the 'eruption,' in 1739, of Methodism. He argues that working class despair was transformed by Methodist doctrine and discipline into a religious and mystical movement of a conservative nature.

In this process Halévy gave a key role to the *bourgeoisie*, a class which he saw as an essential component of leadership in harnessing the enthusiasm of the proletariat. Had the *bourgeoisie* in 1739 been fired by revolutionary ideals there might well have been the conditions for social revolution, but the English *bourgeoisie* retained a strong element of Puritanism and a revulsion against the excesses of the seventeenth-century Civil War. In Halévy's words, 'There was the raw material for a general revolt. But popular discontent took the form the discontent of the *bourgeoisie* wanted to give it: a religious and conservative form' (Halévy 1971, p. 75).

On this foundation is constructed the broader thesis with which Halévy is more generally known. In *A History of the English People* in 1815 he stated the problem as that of explaining 'the extraordinary stability which English Society was destined to enjoy throughout a period of revolutions and crises; what we may truly term the miracle of modern England, anarchist but orderly, practical and businesslike, but religious, and even pietist' (Halévy 1924, p. 339). The explanation he gives is complex and is focused not exclusively on Methodism, nor solely on the working class, for Methodism's impact is traced first on Nonconformity, then on the Evangelical wing of the Church of England, and through them on the ruling class, and finally on secular thought. In contrast to the view of religion as an 'opiate' which Marxian critics of Halévy such as E. P. Thompson have employed, the thesis is better interpreted as a version of *embourgeoisement*. Halévy did not attribute the influence of Methodism to the depths of the working class but saw it as having an impact on upwardly mobile workers. Methodism, originating in the Church of England, occupied an interstitial position between that Church and Nonconformity, offering a means of passage between the two: 'Puritan nonconformity thus tended

to become a transitional creed, a stage in the history of an English family. The unskilled laborer becomes in turn a skilled workman, an artisan, the head of a small business, a businessman possessed of a modest capital, and as he rises out of the barbarism in which the working class was plunged, he becomes a Nonconformist. If he himself rises still higher on the social ladder, or if his children rise after his death, he, or they go over to the Church of England' (Halévy 1924, p. 371). Wesleyan Methodism, he claimed, rendered the transition imperceptible. Paradoxically, as the religious institutions facilitated such upward social mobility they also served to sharpen status distinctions, for while the middle-class Nonconformist might be despised by members of the Church of England, to which his family might one day belong, compensation could be found in contempt shown towards the working class from which the family had come. Hence, Halévy's opinion that 'the élite of the working class, the hard-working and capable bourgeois' had been instilled with values which did not threaten the existing order of society.

Wesleyan Methodism had originated in the Church of England among a group of conservative High Church clergy. In preaching an egalitarian doctrine of salvation for all it had successfully mobilized an important segment of the rising working class. Here is an excellent example of Halévy's utilization of systematic dichotomy at the core of a belief system. He goes on to demonstrate the influence which Methodism had on establishment opinion by proposing that Wesley had left a 'rear-guard' in the Church of England in the form of the Evangelical movement, which sought to realize his original goal of Church regeneration: 'If the Wesleyan sect, with its hierarchic constitution, and frank political conservatism, constituted the High Church of Nonconformity, the new Low Church or evangelical party was a species of Anglican Methodism' (Halévy 1924, p. 379). With powerful representatives at the University of Cambridge, and with wealthy and influential advocates in the 'Clapham sect' (which was comprised principally of Members of Parliament like William Wilberforce), the Evangelicals were in a position to influence not only ecclesiastical reform but also to campaign for a wider reform of morals—including those of the ruling class. In these several ways, argued Halévy, the influence of Wesleyan Methodism had permeated throughout English society and had helped establish the temper of Victorian England. By contrast, no comparable institution or creed had existed in eighteenth-century France.

Halévy's *Histoire du socialisme européen* (1948) and *L'ère des tyrannies* (1938) are the most cited of his works in France. They further extend the central theme of his work on English ideas, where the contrast between reform and revolution provided his comparative focus. Just as the contradiction between natural and artificial interests had been identified in

utilitarianism, and that between egalitarian salvation and hierarchical discipline in Wesleyanism, so the contradiction in socialism derived from the conjoint goals of human emancipation and the zeal to organize society. The decisive event in Halévy's view of socialism, which changed from a skeptical liberalism to a critical pessimism, was the First World War. The growth of the power of the state and the spread of bureaucracy were seen not only as threats to parliamentary democracy but also as evidence of the dominance of the organizational over the emancipatory principle in socialism.

Halévy's central thesis is clearly stated: 'Since its beginnings, in the early years of the nineteenth century, socialism has suffered from an internal contradiction. On the one hand, its partisans often present it as the outcome and fulfilment of the Revolution of 1789, a revolution of liberty, a liberation from the last remaining subjection after all the others have been destroyed: the subjection of labor by capital. But, on the other hand, it is also a reaction against individualism and liberalism; it proposes a new compulsory organization in place of the outworn institutions destroyed by the Revolution ...' (Halévy 1966, p. 265). It is known that the pessimistic conclusions he drew from this diagnosis surprised many in his intellectual circle at the *Société Française de Philosophie*, and arrangements were being made for a further debate on his propositions when he died in 1937.

3. Impact and Controversy

Halévy's work continues to be a focus of research and debate among social scientists. Two biographies have appeared, the first relating him to the French and European contexts of his time and the second emphasizing the major themes of his work: both characterize the 40-year span of his writing with the same term, 'unity' (Bramsen 1978, p. 348, Chase 1980, p. 3). In Chase's work there is a particularly clear exposition of the 'Halévy thesis' which gives effective treatment to aspects of his argument which have sometimes been obscured in subsequent debate, such as the broader significance of Nonconformity and his portrayal of the 'hard-working and capable bourgeois' as the 'élite of the working class.' The republication in 1995 of a French edition of his book on English utilitarianism has stimulated critique of a work that had been widely accepted as a definitive analysis of its subject. Thus, Halévy has been taken to task for allegedly interpreting the principle of utility, or seeking of pleasure, as both a positive and a normative statement—as both an 'is' and an 'ought' (Vergara 1998).

Even more extensive has been the attention paid to his claim that the Methodist and Evangelical movements were responsible for England's remarkable stability during a period of political ferment in Europe.

Shortly after the publication of the first volume of *A History of the English People in the Nineteenth Century*, the Hammonds offered their portrayal of Methodism as having a repressive and intolerant stance towards the working class. In 1957, in *History Today*, E. J. Hobsbawm questioned Halévy's claims of the extent to which Methodism penetrated the working class and offered an alternative explanation of England's political stability deriving from the ruling classes's ability to retain control. E. P. Thompson's *The Making of the English Working Class* in 1963, however, attributed even more influence to Methodism in subduing the working class than had Halévy. In a 'Postscript on the Halévy Thesis' in *Victorian Minds*, Himmelfarb (1968) commented on the curious status of the thesis in current historical literature, where it was often assumed that the thesis had been discredited or greatly modified while in reality there had been no substantial exploration of it. After reviewing the main features of the debate she concluded by paraphrasing Halévy: 'Will Halévy defeat Marx as the interpreter of this crucial period in English history?' (Himmelfarb 1968, p. 299).

Himmelfarb's postscript became the preface to two decades of research into 'the Halévy thesis.' Semmel's 1971 translation of Halévy's 1906 articles *The Birth of Methodism in England* (Halévy 1971) was published with an introduction which noted that 'the full dimensions of the thesis, as Halévy stated them, have been imperfectly understood'. This was because the 1906 articles had gone virtually unread. The thesis was further explored in *The Methodist Revolution* (Semmel 1973), which began by noting that in the major discussions of the 'Halévy thesis' the role of Methodist theology had been afforded little significance, in contrast to social scientific discussions of seventeenth-century Calvinism. Semmel devoted several chapters to the elements which contributed to the synthesis of Methodist beliefs, restating Halévy's observation of an underlying contradiction between the revolutionary and counter-revolutionary elements of Wesleyan doctrine.

The thesis continued to be the subject of discussion throughout the 1970s and attracted the attention of sociologists examining religion and social change: in this context the 'Halévy thesis' has been described as 'the political corollary, applied in a more restricted historical and geographical context, of the broader thesis put forward by Weber' (Hill 1973, p.183). As closer attention has been paid to Halévy's work, some of his specific historical observations—and their interpretation by more recent writers—have provided further stimulus for debate. One such area has been the role of foreign missions, and specifically of the Wesleyan Methodist Missionary Society. Did such activity provide a safety valve for social and religious tensions, rechanneling potentially disruptive enthusiasm by exporting it, or was there equal concern for home missions? These questions have been addressed

in a revisiting of Halévy's original account (Piggin 1980).

In a wide-ranging review of the debate, Itzkin, like Himmelfarb before her, has noted that many of the contentious issues remain unresolved. While Halévy's antagonists have succeeded in revising some of his claims, the central thesis has remained viable and continues to constitute a 'working historical hypothesis' (Itzkin 1975, p. 56). In the work of revision, attention has focused on the numerical strength of Methodism together with its geographical distribution; on the nature and political stances of the variety of bodies which came within the overall rubric of Methodism; and on the way in which moderate political deviance can make 'an important contribution to public order and political stability' (Gilbert 1979, p. 392).

With the culmination of two decades of lively debate over the 'Halévy thesis,' a collection of most of the important statements was published in Olsen (1990). Extracts from Halévy, from historians of Methodism and Evangelicalism such as Maldwyn Edwards, John Walsh, and Kiernan, and from critics like the Hammonds, Hobsbawm, and E. P. Thompson provide a striking testimony to the continuing vitality of a thesis first advanced in 1906. The parallel with the 'Weber thesis,' first published in 1904–5, is too close to ignore.

Halévy's lectures on the history of socialism, reconstructed from his students' notes and published posthumously as *Histoire du Socialisme Européen* in 1948, remains untranslated into English. *The Era of Tyrannies*, first published posthumously in French in 1938, has, however, been translated and published in English (Halévy 1966). In the latter work the author provides not only a summary account of socialism but also a treatment of the era of tyrannies—using the term in the classical Greek sense of a group of armed men who seize the state—under Stalin, Hitler, and Mussolini. At the end of World War One Halévy had written: '... I recognize that socialism holds the key to the secret of the future. But I cannot decipher this secret nor can I say whether socialism will lead us to a Swiss-type, universalized Republic or to European Caesarism' (Aron 1984, p. 407). In the last years of his life he took an increasingly pessimistic view of Europe's future in the event of another war, which he thought likely. For him, that would represent the final blow to liberal traditions; but for this judgment Aron has criticized him. While cautioning against over-confidence, Aron thought Halévy had 'sinned through over-pessimism' (Aron 1984, p. 422). With the benefit of hindsight Aron's criticism may be justified, but it should be recalled that in the era of the 1930s Halévy was not alone among social scientists who believed that the capitalist élite were becoming 'highly placed bureaucrats.'

See also: Bentham, Jeremy (1748–1832); Durkheim, Emile (1858–1917); Liberalism; Marx, Karl (1818–89);

Socialism; Utilitarianism: Contemporary Applications; Weber, Max (1864–1920)

Bibliography

- Aron R 1984 *Élie Halévy. Government and Opposition* 19(4): 407–22
- Bramsen M B 1978 *Portrait D'Élie Halévy*. Grüner, Amsterdam
- Chase M 1980 *Élie Halévy: An Intellectual Biography*. Columbia University Press, New York
- Gilbert A D 1979 Methodism, dissent and political stability in early industrial England. *Journal of Religious History* 10(4): 381–99
- Halévy É 1924 *A History of the English People in 1815* [trans. Watkin E I, Barker D A]. Fisher Unwin, London
- Halévy É 1952 *The Growth of Philosophic Radicalism* [trans. Morris MJ]. Faber and Faber, London
- Halévy É 1966 *The Era of Tyrannies* [trans. Webb R K]. New York University Press, New York
- Halévy É 1971 *The Birth of Methodism in England* (trans. Semmel B). University of Chicago Press, Chicago
- Hill M 1973 *A Sociology of Religion*. Heinemann Educational, London
- Himmelfarb G 1968 *Victorian Minds*. Weidenfeld and Nicolson, London
- Itzkin E 1975 The Halévy thesis—A working hypothesis? *Church History* 44(1): 47–56
- Olsen G W (ed.) 1990 *Religion and Revolution in Early-industrial England: The Halévy Thesis and its Critics*. University Press of America, Lanham, MD
- Piggin S 1980 Halévy revisited: The origins of the Wesleyan Methodist Missionary Society. *Journal of Imperial and Commonwealth History* 9(1): 17–37
- Semmel B 1973 *The Methodist Revolution*. Basic Books, New York
- Vergara F 1998 A Critique of Élie Halévy. *Philosophy* 73(283): 97–111

M. Hill

Hall, Granville Stanley (1844–1924)

The age of the modern research university began with the founding of the Humboldt University in Berlin in 1810. One of the 9,000 Americans who studied at this and other German universities during the nineteenth century was Granville Stanley Hall. He, like they, sought to gain more advanced knowledge and research skills than was available elsewhere. This educational opportunity aided him in becoming the first professor of psychology and pedagogics at the new Johns Hopkins University, and later the first president of research-oriented at Clark University in the USA. During his academic career, Hall launched the American discipline of psychology (Ross 1972, 1979) and the science of education (White 1992), out of the then

broader study of philosophy. He achieved these intellectual accomplishments through his long faculty and presidential career, numerous books and more than 400 other publications, creation of research journals, and extensive professional conference presentations. In this way, Hall played the initial pivotal and defining role in founding psychology and education as social and behavioral sciences in American higher education (Mitchell and Haro 1999, p. 50, Lagemann 2000).

1. Hall's Early Life, Career, and Intellectual Pursuits Until 1880

Born in Ashfield, Massachusetts on February 1, 1844 to Granville Bascom Hall and Abigail Beals Hall, Stanley was the oldest of nine siblings. His upbringing on the family farm offered his inquisitive mind opportunities to observe nature intimately, which would so powerfully shape his future intellectual outlook. Eschewing agrarian life, Hall completed college preparatory studies at Williston Seminary in Easthampton, Massachusetts in 1862 and entered Williams College in the following year. There he studied under two famous American educators and philosophers, Mark Hopkins and John Bascom. Their influences shaped his interest in philosophy and theology, respectively. Finishing his bachelor's degree in 1867, Hall next decided to study for the ministry and entered Union Theological Seminary in New York. Yet, two years of study, some pastoral experience, and a frank discussion about his intellectual interests with the imposing American theologian Henry Ward Beecher resulted in Hall's first trip to Germany to study philosophy at the University of Berlin in 1869. In Friedrich Adolf Trendelenburg's seminary, he became an avowed Hegelian and contemplated becoming a philosopher. With the outbreak of the Frank-Prussian war, Hall returned to New York and finished his bachelor's degree in divinity in 1871. Yet, his intellectual and philosophy study there led him to embrace the evolutionary ideas of Charles Darwin and Herbert Spencer (Hall 1923).

Forsaking the ministry then, Hall sought the academic life. He became a professor at Antioch College in Ohio, where he taught English, French, and German from 1872 to 1876. Becoming increasingly disenchanted with Hegel intellectually during this time, Hall openly became an enthusiast for Darwin, Spencer, and Huxley and positivistic science (Hall 1923, p. 199, Ross 1972, pp. 58–9). The turning point in his intellectual life occurred with his reading of Wilhelm Wundt's (1873) first volume of *Physiological Psychology*, after which he decided to specialize in psychology. Leaving Antioch, Hall completed a Ph.D. at Harvard University in 1878 under William James; it was the first doctorate in psychology awarded in the USA. His dissertation explored the physiological

relationships between muscles, nerves, and consciousness from a genetic perspective (Ross 1972, pp. 70–5, Wilson 1914, p. 64). Hall's research showed his fascination with experimental psychology and its links to Darwinian and Spencerian evolutionary thought. To bolster his research in the study of psychology, and because he could not obtain a suitable professorial appointment at a major university, Hall again went to Germany to study under Hermann von Helmholtz and William Wundt at the universities of Berlin and Leipzig, respectively (Goodchild 1996, pp. 72–3).

2. Hall's Faculty Appointments at Johns Hopkins University from 1882 to 1887

After returning to the USA in 1880, Hall lectured at Harvard at the invitation of President Charles W. Eliot and then at Hopkins by courtesy of President Daniel Coit Gilman. Later the Hopkins president appointed him a lecturer in 1882 and then Professor of Psychology and Pedagogics in 1884. From these positions, Hall furthered the study of education and psychology through a series of major articles exploring their relationships to human development. First, he surveyed pedagogical study in the USA and then explored the moral and religious training of both children and adolescents in one of his first discussions of child development. In analyzing the needs of education later, the professor stated boldly that 'we must study and follow the child's nature as it actually is' (Hall 1883a, p. 286). Hall further proclaimed his new concept of 'natural' methods of education where education and psychology may both be utilized to understand the relationship between intelligence and human development in his important article, *New Departures in Education* (Hall 1885a, pp. 147–8). Finally, he and one of his doctoral students outlined the entire field of education in their 1886 book, *Bibliography of Education* (Hall and Mansfield 1886).

Second, in psychology, Hall quickly created one of the first experimental psychological laboratories in America at Hopkins and authored many of its early articles in experimental psychology (Boring 1950, Hawkins 1960, pp. 202–3). One of his important contributions was the publication of 'The new psychology' (1885b) which outlined this developing field. In 1887, he created a new research journal, the *American Journal of Psychology*, to publish psychological research. It was pivotal in establishing this new discipline. Third, his pedagogical and psychological ideas had already coalesced a few years earlier in his first major empirical studies using questionnaires on child development. 'The contents of children's minds' (1883b) and *The Study of Children* (1883c) described 'the real nature' of children, used the reports of actual teachers' observations from the questionnaires, and suggested appropriate pedagogical techniques. They

played a major role in launching Hall's national reputation and his central position as the founder of the child study movement in the USA (White 1990). Some of his doctoral students followed him as his career evolved; others, particularly John Dewey and James McKeen Cattell, were not favored and moved onto their own significant accomplishments (Ross 1972, pp. 144–7). Thus Hall's work at Hopkins set the stage for the remainder of his scholarly career and publications as he pursued research in education, psychology, and their mutual implications for understanding human development.

3. Hall's Presidency of Clark University from 1888 to 1920

In Worcester, Massachusetts, wealthy industrialist Jonas Clark founded an institution of higher learning with a \$1 million gift in 1887. He intended it to have a research mission for older male students and be a 'useful' college for local boys (Koelsch 1987, pp. 9–12). Hall was named its first president in April 1888. He was determined to create a new graduate research university comparable to Hopkins, following the German ideal. The first three years saw considerable fulfillment of his vision. However, conflict with the founder over the college and loss of many faculty members to the new University of Chicago severely constrained the president. The struggling and much smaller institution still embraced the primacy of research as the reform ideal for the mission of all American universities (Goodchild 1996, Goodchild and Miller 1997). Eventually, Hall allowed an undergraduate college to be created when funds from the founder's estate became available. The president remained highly productive in his research and continued to oversee the institution until his retirement in 1920.

4. Hall's Prolific Scholarly Achievements During his Presidency

The president's accomplishments in psychology, education, and their implications for a new genetic interpretation of human development established the American prominence of these new scientific fields of study. In psychology, Hall founded the American Psychological Association in 1892, became its first president, and anchored his *American Journal of Psychology* to the organization. It remained the pre-eminent national association for this growing academic discipline. In 1904, Hall's most famous psychological study, *Adolescence: Its Psychology and Its Relations to Physiology, Anthropology, Sociology, Sex, Crime, Religion and Education*, appeared. This definitive two-volume work on adolescence comprised many of his earlier ideas on the child's next stage of development, established this stage as a scientific

concept, related this developmental period to various human behaviors, and discussed its meaning for other emerging academic fields, especially education. Hall's interest in the sexual development of youth in this work was ground breaking. It led him to invite Sigmund Freud, Carl Jung, and many other leading psychologists to the famous 1909 twentieth Anniversary Conference of the founding of Clark University (Koelsch 1987, pp. 246–7). As if to secure his role in this developing discipline, Hall published the valuable book, *Founders of Modern Psychology*, in 1912. After World War I, the scholarly president turned to his lifelong interest in religion and its psychological meaning. He authored a major reinterpretation of Christianity's founder in his *Jesus the Christ, in the Light of Psychology* (1917) and its application in the advocacy of a new life ethic in his *Morale: The Supreme Standard of Life and Conduct* (1920a) in which Hall espoused a holistic understanding of health and well-being as the rule for human conduct. With retirement from Clark's presidency in 1920, Hall reflected on his accomplishments in his *Recreations of a Psychologist* (1920b), and *Life and Confessions of a Psychologist* (1923). Hall achieved another scholarly triumph by completing his work on the last stage of human development with the publication of *Senescence: The Last Half of Life* (1922). In all this pioneering work, he made some significant missteps, especially claiming child development recapitulates human evolution (White 1990, p. 31) and a too pervasive Darwinian and Spencerian 'evolutionism' (Ross 1972, p. 92). On the other hand, he awarded the first doctorate in psychology to an African-American and encouraged Asian students to study psychology at Clark (Goodchild 1996).

By his death on April 24, 1924, Hall had established psychology as an American discipline and study, almost entirely graduated the next generation of psychologists (Goodchild and Miller 1997), and secured the role of behavioral and genetic psychology as being an integral part. To insure the continuation of his work, Hall's estate established at Clark an endowment for a genetic psychology professorship and for his educational journal that was redirected and renamed *The Journal of Genetic Psychology*.

Similarly, his work in education commanded national attention. Hall believed that education was applied psychology. He followed identical approaches in establishing the importance of education as a field of study, as he had done in psychology. Hall offered a series of university lectures on kindergarten, elementary, secondary, and higher education (Goodchild 1991). He next founded a research journal, *The Pedagogical Seminary: An International Record of Educational Literature, Institutions, and Progress*, in 1891. It provided a new research outlet from studies intended to reform education from American and international perspectives. Further, he extracted his pedagogical ideas from his books on adolescence and

released them as *Youth: Its Education, Regimen, and Hygiene* (1906) to serve educators and teachers who were implementing his ideas in the classroom. Five years later, Hall presented his definitive study on education with another extensive two-volume work, *Educational Problems* (1911). He claimed to have fomented an American educational revolution where 'now the school, including its buildings, all its matter and method, revolve about the child' (p. 1: vi). This 'natural' education was to be individualized to the student's developmental stage, sex, and race. Thus pedagogy was to assist in the intellectual development of every person, although higher education was intended for the most able. At Clark, Hall had been instrumental in creating the study of education—particularly in conducting empirical research on children to develop better pedagogy, offering the first courses in higher education, and espousing genetic psychology as the basis for educational practice. His extensive publications in higher education focused on the mission of the new research university and its faculty (Goodchild 1996). Similarly, missteps occurred in education as Hall advocated separate education for women, decried 'identical' coeducation, and espoused elitism in higher education (Ross 1972, pp. 311–12). Nevertheless, the president greatly assisted in the arrival of education as an academic field of study (Cremin 1961).

Hall's most significant contribution may be the creation of developmental psychology and its relationship to education (White 1992). His research on child study, adolescence, and old age enabled him to argue for five stages of human development: (a) childhood; (b) adolescence; (c) middle life, from 25 to 45; (d) senescence, after 40; and (e) senectitude, approximately near the end of life (Hall 1922, p. vii). His psychological and educational ideas and practices were focused on what he believed, would further the human race. His Darwinian and Spencerian concepts of evolution pervaded his understanding of human genetic development. He consistently maintained these ideals, even in the face of demands for coeducation in schools, growing behaviorism in psychology, and expansive progressivism in education. Hall was a man of his times, his scientific worldview centered on the mid- to late nineteenth century ideas and their implications for professional practice.

Finally, Hall's most important achievement when assessing the rise of the social and behavioral sciences was to engender the development of psychology and education from their philosophical progenitor. His patterns consisted of a critical assessment of the condition of the field first, next an assertion of a new position for the field, then a new research journal to publish these findings, and finally definitive works exploring the field of study. Hall thus played an essential role in the establishment of psychology and education as new academic fields of study in the USA. In 1992, the centenary meeting of the American

Psychological Association pointed to his pivotal role in these founding moments (Goodchild 1992, White 1992).

See also: Adolescent Development, Theories of; Aging, Theories of; Education: Phenomena, Concepts, and Theories; Helmholtz, Hermann Ludwig Ferdinand von (1821–94); James, William (1842–1910); Life Course: Sociological Aspects; Lifespan Development, Theory of; Psychology: Historical and Cultural Perspectives; Rousseau, Jean-Jacques (1712–78); Spencer, Herbert (1820–1903); Wisdom, Psychology of; Wundt, Wilhelm Maximilian (1832–1920); Youth Culture, Sociology of

Bibliography

- Boring E C 1950 *A History of Experimental Psychology*, 2nd edn. Appleton-Century-Crofts, New York
- Cremin L A 1961 *The Transformation of the School: Progressivism in American Education, 1876–1957*. Vintage Books, New York
- Goodchild L F 1991 Higher education as a field of study: Its origins, programs, and purposes, 1893–1960. In: Fife J D, Goodchild L F (eds.) *Administration as a Profession. New Directions for Higher Education*, no. 76. Jossey-Bass, San Francisco
- Goodchild L F 1992 G. Stanley Hall's Psychological Avocation: Founding the Study of Higher Education. Research paper given at the 100th American Psychological Association annual meeting in Washington, DC
- Goodchild L F 1996 G. Stanley Hall and the study of higher education. *The Review of Higher Education* **20**: 69–99
- Goodchild L F, Miller M M 1997 The American doctorate and dissertation: Six developmental stages. In: Goodchild L F, Green K E, Katz E L, Kluever R C (eds.) *Rethinking the Dissertation Process: Tackling Personal and Institutional Obstacles. New Directions for Higher Education*, no. 99. Jossey-Bass, San Francisco
- Hall G S 1883a Educational needs. *North American Review* **236**: 284–90
- Hall G S 1883b The content of children's minds. *Princeton Review* **11**: 249–72
- Hall G S 1883c *The Study of Children*. Printed privately
- Hall G S 1885a New departures in education. *North American Review* **140**: 144–52
- Hall G S 1885b The new psychology. *Andover Review* **3**: 120–36, 239–48
- Hall G S, Mansfield J M 1886 *Hints Toward a Select and Descriptive Bibliography of Education*. Heath, Boston
- Hall G S 1904 *Adolescence: Its Psychology and Its Relations to Physiology, Anthropology, Sociology, Sex, Crime, Religion and Education*, 2 Vols. Appleton, New York
- Hall G S 1906 *Youth: Its Education, Regimen, and Hygiene*. Appleton, New York
- Hall G S 1911 *Educational Problems*, 2 Vols. Appleton, New York
- Hall G S 1912 *Founders of Modern Psychology*. Appleton, New York
- Hall G S 1917 *Jesus, The Christ, in the Light of Psychology*. Doubleday Page, New York

- Hall G D 1920a *Morale: The Supreme Standard of Life and Conduct*. Appleton, New York
- Hall G S 1920b *Recreations of a Psychologist*. Appleton, New York
- Hall G S 1922 *Senescence: The Last Half of Life*. Appleton, New York
- Hall G S 1923 *Life and Confessions of a Psychologist*. Appleton, New York
- Hawkins H 1960 *Pioneer: A History of Johns Hopkins University, 1874–1889*. Cornell University Press, Ithaca, NY
- Koelsch W A 1987 *Clark University, 1887–1987: A Narrative History*. Clark University Press, Worcester, MA
- Lagemann E C 2000 *An Elusive Science: The Troubling History of Educational Research*. University of Chicago Press, Chicago
- Mitchell T R, Haro A 1999 Poles apart: Reconciling the dichotomies in educational research. In: Lagemann E C, Shulman L S (eds.) *Issues in Educational Research: Problems and Possibilities*. Jossey-Bass, San Francisco
- Ross D 1972 *G. Stanley Hall: The Psychologist as Prophet*. University of Chicago Press, Chicago
- Ross D 1979 The development of the social sciences. In: Oleson A, Voss J (eds.) *The Organization of Knowledge in Modern America, 1860–1920*. Johns Hopkins University Press, Baltimore, MD
- White S H 1990 Child study at Clark University: 1894–1904. *Journal of the History of the Behavioral Sciences* **26**: 31–50
- White S H 1992 G. Stanley Hall: From philosophy to developmental psychology. *Developmental Psychology* **28**: 25–34
- Wilson L N 1914 *G. Stanley Hall: A Sketch*. Stechert, New York
- Wundt W 1873 *Principles of Physiological Psychology*. Engelmann, Leipzig, Germany

L. F. Goodchild

Hallucinations, Neural Basis of

Hallucinations, by definition, are percepts experienced in the absence of sensory input. They can be expressed in any of the sensory modalities and arise from a wide range of pathological conditions. Hallucinations are also a normal component of conscious experience. For instance, they arise routinely as ordinary dreaming during sleep. Sometimes daydreams and relaxed waking consciousness produce imagery so vivid that it is momentarily experienced as real (Foulkes and Fleisher 1975). The main types of hallucinations and their distinctive phenomenological characteristics have been the subject of considerable prior study (Brown 1985, Benson and Gorman 1993). A brief summary of the major types follows:

Visual hallucinations: Peduncular hallucinations consist of animals, face, or human figures and are traditionally associated with vascular lesions in the upper brainstem (involving, especially, the cerebral peduncles). They can also occur in cases of encephalitis or during the transition from waking to sleep (Lhermitte 1922, from Brown 1985). *Charles Bonnet syn-*

drome is similar and occurs in the elderly, generally with significant visual impairment. However, human figures are characteristically small and doll-like, and can actually amuse the subject (Benson and Gorman 1993). *Palinopsia* is a pathological persistence of a visual percept originating from external stimulation that arises from lesions of the occipital cortex, seizures, or drug-induced states (Benson and Gorman 1993, Brown 1985). So-called *elementary* hallucinations consist of bands of light or phosphenes that can arise from pathology of the retina or optic nerve, and certain drug-induced states (Brown 1985). *Hypnagogic* hallucinations occur normally during the transition from sleep to wakefulness and are experienced as the continuation of a typical dreamstate. Hallucinations of faces or human figures are reported by approximately 20 percent of patients with schizophrenia. Visual hallucinations also often occur during seizures in patients with epilepsy. These hallucinations can have a geometric design or reflect a recollection of visual images based on past experience. Severe alcohol withdrawal often is accompanied by visual hallucinations of bugs or spiders. Lysergic acid (LSD) and related compounds often produce visual hallucinations consisting of light flashes, colors, geometric figures, and generally are accompanied by distortions of the geometric form of percepts (Crowley 1995). Higher doses produce fantastic forms, distorted facial images, etc. So-called *autoscopic* hallucinations reproduce an actual visual image of the self and are generally associated with epilepsy though these experiences also can occur in patients with schizophrenia or depression (Denning and Berrios 1994).

Auditory hallucinations: Most commonly, auditory hallucinations consist of spoken speech. These hallucinations are reported by 70–80 percent of patients with schizophrenia and less commonly by patients with other psychiatric disorders, such as depression or mania. These patients may also experience hallucinations of non-speech sounds, such as rattling, footsteps, etc. Other common causes of auditory hallucinations of speech include chronic alcoholism and phencyclidine use, or chronic amphetamine abuse. Hallucinations of speech consist of one or more ‘voices’ which comment on the patient’s speech, thought, or behavior, command them to perform certain actions, or converse with each other. A nonpathological variant of this experience is the momentary experience of hearing one’s name being spoken in the absence of an actual speaker. Both pathological and non-pathological variants of hallucinated speech are often triggered by indistinct sounds such as background conversation, street sounds, and running water. Often, if visual hallucinations are also experienced by patients with schizophrenia, they are phenomenologically linked with the auditory hallucination, e.g., a face will appear to speak the hallucinated words. Auditory hallucinations of music can occur in psychiatric patients with an affective disorder such as depression

(Wengel et al. 1989) as well as in patients with deafness and steroid-induced psychotic states. *Palinacousis* is the auditory analogue of *palinopsia* and consists of the persistence of previously heard sounds, music, or speech.

Tactile or haptic hallucinations: These experiences commonly occur during the *delirium tremens* stage of severe alcohol withdrawal and consist of *formication*, i.e., the sensation of insects crawling on the body. Amphetamine or cocaine-induced psychosis can also lead to formication experiences. *Phantom limb* hallucinations follow arm or leg amputation and consist of the persistent sensation of the lost limb that can include the experience of movement and/or pain. Sensation of altered or duplicated body shapes can

occur in epilepsy, toxic encephalopathy, or with hallucinogenic drugs. When such hallucinations are accompanied by delusions that are fantastic, they are likely to be related to a schizophrenic process (Benson and Gorman 1993).

Hallucinations of bodily sense and movement have also been reported following direct electrical stimulation of the temporal lobe in humans (Ishibashi et al. 1964).

Olfactory hallucinations: These hallucinations most typically occur during seizure auras, migraine attacks, or as a result of damage to the olfactory bulb at the base of frontal lobe (e.g., by tumors). Olfactory hallucinations generally have a noxious content, consisting of a burning smell or some other foul odor. They are also reported occasionally by patients with schizophrenia and depression.

Mechanisms of hallucinations: Benson and Gorman (1993) outline a simple, but intuitively appealing, approach dividing hallucinations into two categories: the first ‘release’ hallucinations and the second ‘ictal’ hallucinations. The former arise when ‘normal sensory processing channels are blocked, allowing stored images to be experienced’ (pp. 266–267). Typical examples would be musical hallucinations following deafness or the Charles Bonnet syndrome arising from visual impairment in the elderly. These hallucinations tend to be longer in duration, can often be well-formed and quite elaborate, and occur only in the defective receptive field. ‘Ictal’ hallucinations arise from ‘abnormal neuronal discharge’ and tend to be briefer, less well-formed, and associated with alterations in consciousness. Benson and Gorman also highlight the importance of alterations in attention/arousal in producing certain types of hallucinatory phenomena (1993, p. 278). Examples include dream-like hallucinations arising from the transition from sleep to arousal, the disturbed consciousness of drug-induced states, sleep deprivation, narcolepsy, or delirium. The postulated mechanism is an alteration in the gating of material into conscious awareness. In addition, limbic disorders and frontal dysfunction were postulated to cause hallucinations. We will discuss below some evidence suggesting a role of the former.

More recent work on the details of hallucinatory phenomena, considered in the remainder of the article, suggests some revision may be necessary in the useful framework introduced by Benson and Gorman.

First, consider again phantom limb phenomena, a classic example of 'release' hallucination. These experiences are almost universal, occurring in approximately 85 percent of amputees (Benson and Gorman 1993) and are experienced as if the cortical representation of the limb is retained somehow as a conscious experience. Phantom limb experiences, especially early on, often are finely articulated with detailed sensations of the digits of the hands or feet. After loss of a hand, the person may still, for instance, re-experience the sensation of a long-worn ring. One puzzling observation is that patients who also lose sensation and use of a limb due to spinal cord injuries do *not* generally have phantom limb hallucinations or, if they do, these are very indistinct and fleeting. Why does loss of limb produce these experiences, while paraplegia does not? In both cases, deafferentation has occurred and 'release' phenomena should, in theory, arise.

Recent work by Spitzer et al. (1995) has provided an updated model suggesting that phantom limb hallucinations arise from a progressive reorganization of the sensory area of the cerebral cortex responsible for limb representation. The critical observation underlying Spitzer's model is that sensory ganglia in the spinal cord are known to fire randomly following denervation. These random impulses are then projected in a sustained fashion to the somatosensory cortex. Random firing from these sensory ganglia is not generated, in contrast, in the case of spinal cord injury. Using computer simulations, Spitzer demonstrated that chronic random inputs cause a deafferentized cortical sensory system to reorganize so that activation patterns become self-sustaining. Moreover, a cortical region whose input is limited to noise was shown to assimilate receptive field properties of neighboring cortical areas gradually. Consequently, representations of the original limb representation are gradually usurped. Progressive reorganization of limb cortical maps in this fashion could produce the characteristic subjective 'shrinking' of these hallucinations over time—the phantom limb grows smaller over a period of weeks to months so that, in later stages, it is only experienced as, say, the size of a postage stamp superimposed on the stump.

The limits of the Benson/Gorman approach are also suggested by the hallucinations experienced during amphetamine psychosis. These hallucinations typically consist of spoken speech of whole phrases or sentences or the perception of faces or bodily forms and are often accompanied by paranoid fears. Their phenomenology often resembles that of schizophrenia. Consequently, amphetamine psychosis is studied as an experimental model of this disorder.

These hallucinations do not readily fit a reduced attention/arousal model of hallucinations since

amphetamine generally heightens attention and arousal. There are other data to suggest that perhaps amphetamine-induced hallucinations reflect an 'ictal' or 'neuronal activation' model of hallucinations. It is well-known, for instance, that a single dose of amphetamine produces widespread, heightened activation of the cerebral cortex (Porrino et al. 1984). However, amphetamine-induced hallucinations almost invariably do not emerge unless the person is a chronic abuser of the drug (Bell 1973).

Induction of apparent hallucinatory behavior in monkeys also requires repeated exposure to this drug (Castner and Goldman-Rakic 1999). Thus, it seems clear that the hallucinations induced by amphetamine do not arise simply from heightened brain activation—which is readily detected following a single exposure to the drug—but instead reflect complex changes in brain function generated over time by repeated drug exposure. The exact nature of these brain 'sensitization' changes has yet to be determined. However, recent studies suggest that alterations of neural circuitry in brain regions which receive projections from dopamine neurons and ordinarily respond to amphetamine are critical (Pierce and Kalivas 1997). These alterations probably involve a cascade of molecular events that enhance cellular aspects of learning/conditioning mediated by receptors for the excitatory amino acid neurotransmitter, glutamate, as well as elaboration of new neuroanatomic connections (Robinson and Kolb 1999, Wolf 1998). These changes produce widely distributed neurocircuits that consequently become 'primed' to respond to the activating effects of amphetamine.

Reports of direct electrical stimulation in the human brain also suggest that neural activation alone is not sufficient to cause hallucinations. These studies have been conducted in patients with intractable epilepsy who were evaluated for neurosurgical resection of seizure loci. Penfield, a pioneer in this area, reported that complex hallucinations whose content consisted of 'engrams' of previous memories could be elicited in unanesthetized human subjects by electrically stimulating temporal lobe neocortical gray matter (cf. Penfield and Perot 1963). Penfield proposed that such stimulation directly accessed neural networks that actually coded for memories. Later workers found it difficult to obtain similar results, however. Halgren et al. (1978), Gloor et al. (1982), and Horowitz and Adams (1970), for instance, were able to elicit experiential responses but by stimulating deeper limbic structures (i.e., hippocampus and amygdala) rather than neocortical areas.

These experiences tended not to be memories but did include a range of hallucination-like experiences including unformed sensory experiences (lights, streaks, flashes) and complex visual experiences of persons, scenes, or objects or autoscopic experiences of the self. The location of sites that elicited hallucinations is curious since it is generally assumed that only

neocortical structures (not limbic structures) code for actual sensory experiences. This raises the question of how hippocampal and amygdala stimulation, whose known functions pertain to registration of emotional salience, memory storage, and memory reactivation, could produce hallucinations.

One clue was that hallucinations were more likely to emerge when limbic stimulation produced not simply activation, but sustained afterdischarges, i.e., rhythmic activity detected via EEG recordings (Halgren et al. 1978, Horowitz and Adams 1970). After-discharges of this sort are more likely to lead to a temporary functional lesion of neural systems rather than activation of coded information (Gloor et al. 1982). Analogously, rhythmic electrical stimulation of neocortical tissue produces temporary aphasia or speech arrest if directed to language processing areas of the brain (Ojemann and Mateer 1979). Halgren et al. (1978) and Horowitz and Adams (1970) proposed that sustained disruption of limbic structures induced by after-discharges lead to the 'release' of sensory material from neocortical regions into consciousness. Consistent with this view are EEG studies in animals showing that the REM stage of sleep and LSD administration, which each produce vivid hallucinations in humans, cause dramatic *reductions* in hippocampal activation (Halgren et al. 1978). These observations suggest that the actual production of hallucinations by intracerebral stimulation arises from a dynamic interaction between neocortical and limbic systems, rather than simple activation of a group of neurons.

As with other forms of hallucinations, a new picture of auditory hallucinations in patients with schizophrenia is now emerging. Functional neuro-imaging studies indicate that these hallucinations are accompanied by activation from speech perception and auditory neurocircuitry (Silbersweig et al. 1995, Dierks et al. 1999); one of these studies also strongly implicated activation of hippocampal structures as well (Silbersweig et al. 1995). However, it is not clear if this distributed pattern of activation produces auditory hallucinations or arises as a consequence of some other pathophysiological process. For instance, shifts in attention due to the experience of hallucinations (from outward focus to internal imagery focus) may produce regional brain activation effects. If so, brain activation of this type will accompany hallucinations but will not in itself signal a causal mechanism.

Transcranial magnetic stimulation (TMS) is a method whereby an electromagnet placed on a scalp transmits magnetic pulses or waves to a small portion of the brain. Repetitive transcranial magnetic stimulation (rTMS) delivered at a low frequency (once per second) has been shown to reduce the reactivity or excitability of the part of the brain stimulated and other brain regions functionally connected to the region stimulated (Chen et al. 1997, Wassermann et al. 1998). A pilot study using a double-blind, crossover

design compared effects of active versus sham low frequency rTMS in schizophrenic patients with auditory hallucinations (Hoffman et al. 1999, 2000). Based on the aforementioned neuro-imaging studies, the left temporoparietal cortex, a brain region critical for speech perception, was stimulated. A total of forty minutes of rTMS was administered over a four-day period. Statistically robust improvements in auditory hallucinations following active rTMS relative to sham rTMS were observed. These rTMS data support the hypothesis that activation of speech processing neurocircuitry contributes to the production of these hallucinations rather than emerging as a 'downstream' consequence of some other pathophysiological process.

These findings still do not address why speech processing areas of the brain, such as temporoparietal cortex, are pathologically activated in patients with schizophrenia. To understand this process better, computer simulations of neural networks have been developed that reproduce some aspects of speech processing of continuous narrative speech (Hoffman and McGlashan 1997). These simulations suggest that the ultimate 'cause' of auditory hallucinations is a loss of neuronal connections within these networks. Reductions in cortico-cortical connectivity are a normal component of 'late' development in human association cortex that continues during human adolescence (Huttenlocher and Dabholkar 1997). There are empirical data suggesting that reductions in synaptic density accompany some forms of learning (Scheich et al. 1991). Along these lines, computer simulations demonstrate that selective pruning of connections within recurrent neural networks produces significant information processing advantages (Hoffman and McGlashan 1997). In this model, pruning connections caused the system to fill in some of the gaps due to ambiguous phonetic input. It is well known that ordinary continuous speech, when produced at normal rates, has significant phonetic ambiguity that we do not 'hear' because our brains are also very skillful in 'filling in the gaps.' When pruning was excessive in our simulated networks, however, the system began to produce outputs in the absence of any phonetic input at all, thereby simulating speech hallucinations. This model therefore suggests that the hallucinations of schizophrenic psychosis arise as a result of an otherwise normal neuro-developmental trend that is excessive or unregulated. An appeal of this model is that it provides an account for the characteristic age of onset of schizophrenia, namely late adolescence and young adulthood.

Overall, the 'release' and 'ictal' models of hallucination described by Benson and Gorman appear to be good starting points. A comprehensive understanding of their nature, however, needs to consider interactions between different brain regions as well as between different 'levels' of neural processes involving molecular and neuroplastic alterations on the one

hand, and large-scale neural networks dynamics on the other. In spite of our current limited understanding of hallucinations, this fascinating area is likely to provide important insights into the pathophysiology of brain disturbances, as well as the normal physiology of perception and consciousness.

See also: Phantom Limbs; Visual Imagery, Neural Basis of

Bibliography

- Bell D S 1973 The experimental reproduction of amphetamine psychosis. *Archives of General Psychiatry* **29**: 35–40
- Benson D F, Gorman D G 1993 The neural basis of hallucinations. *Current Neurology*, Vol. 13, Mosby-Year Books, pp. 265–87
- Brown J W 1985 Hallucinations: Imagery and the microstructure of perception. In: Fredriks J A M (ed.) *Handbook of Clinical Neurology*. Elsevier, New York, Vol. 1, pp. 351–372
- Castner S A, Goldman-Rakic P S 1999 Long-lasting psychotomimetic consequences of repeated low-dose amphetamine exposure in rhesus monkeys. *Neuropsychopharmacology* **20**: 10–28
- Chen R, Classen J, Gerloff C, Celnik P, Wassermann E M, Hallett M 1997 Depression of motor cortex excitability by low-frequency transcranial magnetic stimulation. *Neurology* **48**: 1398–403
- Crowley T J 1995 Hallucinogen-related disorders. In: Kaplan H I, Sadock B J (eds.) *Comprehensive Textbook of Psychiatry*. Williams and Wilkins, Baltimore, MD, Vol. 1, pp. 831–8
- Denning T R, Berrios G E 1994 Autoscopic phenomena. *British Journal of Psychiatry* **165**: 808–17
- Dierks T, Linden D E J, Jandl M, Formisano E, Goebel R, Lanfermann H, Singer W 1999 Activation of Heschl's gyrus during auditory hallucinations. *Neuron* **22**: 615–21
- Foulkes D, Fleisher S 1975 Mental activity in relaxed wakefulness. *Journal of Abnormal Psychology* **84**: 66–75
- Gloor P, Olivier A, Quesney L, Andermann F, Horowitz S 1982 The role of the limbic system in experiential phenomena of temporal lobe epilepsy. *Annals of Neurology* **12**: 129–44
- Halgren E, Walter R D, Cherlow D G, Crandall P H 1978 Mental phenomena evoked by electrical stimulation of the human hippocampal formation and amygdala. *Brain* **101**: 83–117
- Hoffman R E, Boutros N N, Berman R M, Roessler E, Krystal J H, Charney D S 1999 Transcranial magnetic stimulation of left temporoparietal cortex in three patients reporting hallucinated 'voices.' *Biological Psychiatry* **46**: 130–2
- Hoffman R E, Boutros N N, Hu S, Berman R M, Krystal J H, Charney D S 2000 Transcranial magnetic stimulation of left temporoparietal cortex in schizophrenic patients reporting auditory hallucinations. *Lancet* **355**: 1073–5
- Hoffman R E, McGlashan T H 1997 Synaptic elimination, neurodevelopment and the mechanism of hallucinated 'voices' in schizophrenia. *American Journal of Psychiatry* **154**: 1683–9
- Horowitz M, Adams J 1970 Hallucination on brain stimulation: Evidence for revision of the Penfield Hypothesis. In: Kemp W (ed.) *Origin and Mechanism of Hallucinations*. Plenum Press, New York
- Huttenlocher P R, Dabholkar A S 1997 Regional differences in synaptogenesis in human cerebral cortex. *Journal of Comparative Neurology* **387**: 167–78
- Ishibashi T, Hori H, Endo K, Sato T 1964 Hallucinations produced by electrical stimulation of the temporal lobes in schizophrenic patients. *Tohoku Journal of Experimental Medicine* **82**: 124–39
- Lhermitte J 1922 Syndrome de la calotte du pédoncule les troubles psychosensoriels dans les lésions du mésencéphale. *Revue Neurologique* **38**: 1359–65 November 9
- Ojemann G, Mateer C. 1979 Human language cortex: Localization of memory, syntax, and sequential motor-phoneme identification systems. *Science* **205**: 1401–3
- Penfield W P, Perot P 1963 The brain's record of auditory and visual experience. A final summary and discussion. *Brain* **86**: 595–696
- Pierce R C, Kalivas P W. 1997 A circuitry model of the expression of behavioral sensitization to amphetamine-like psychostimulants. *Brain Res. Brain Res. Rev.* **25**: 192–216
- Porrino L J, Lucignani G, Dow-Edwards D, Sokoloff L 1984 Correlation of dose-dependent effects of acute amphetamine administration on behavior and local cerebral metabolism in rats. *Brain Research* **307**: 311–20
- Robinson T E, Kolb B 1999 Alterations in the morphology of dendrites and dendritic spines in the nucleus accumbens and prefrontal cortex following repeated treatment with amphetamine or cocaine. *European Journal of Neuroscience* **11**: 1598–604
- Scheich H, Walhauser-Franke E, Braun K 1991 Does synaptic selection explain auditory imprinting? In: Squire L R, Weinberger N M, Lynch G, McGaugh J L (eds.) *Memory: Organization and Locus of Change*. Oxford University Press, London, pp. 114–59
- Silbersweig D A, Stern E, Frith C, Cahill C, Holmes A, Grootenok S, Seaward J, McKenna P, Chua S E, Schnorr L, Jones T, Frackowiak R S J 1995 A functional neuroanatomy of hallucinations in schizophrenia. *Nature* **378**: 176–9
- Spitzer M, Bohler P, Weisbrod M, Kischka U 1995 A neural network model of phantom limbs. *Biological Cybernetics* **72**: 197–206
- Wassermann E M, Wedegaertner F R, Ziemann U I, George M S, Chen R 1998 Crossed reduction of human motor cortex excitability by 1-Hz transcranial magnetic stimulation. *Neuroscience Letters* **250**: 141–4
- Wengel S P, Burke W J, Holemon D 1989 The sounds of silence? *Journal of the American Geriatrics Society* **37**: 163–6
- Wolf M E 1998 The role of excitatory amino acids in behavioral sensitization to psychomotor stimulants. *Progress in Neurobiology* **54**: 679–70

R. E. Hoffman

Harlow, Harry Frederick (1905–81)

Open any Introductory Psychology textbook published since 1960, start thumbing through the pages, and most likely you will come across a photograph of a rhesus monkey infant clinging to a cloth-covered dummy adjacent to another dummy, this one with a wire surface and a protruding milk-filled bottle. Along with the photo will be a description of Harry Harlow's classic 'surrogate mother' experiments, in which he

dramatically demonstrated that tactile contact was more important than feeding in establishing a monkey infant's bond to its surrogate. Harlow's conclusions from these studies, that 'there is more to mother love than mere milk' (Harlow 1958), had important and lasting consequences. They effectively demolished the prevailing theoretical consensus that an infant's primary social relationship is formed as a direct consequence of the feeding process, be it through oral gratification, as advocated by orthodox psychoanalysts, or by its reinforcement history, the contemporary behaviorist view. More importantly, Harlow's findings also provided compelling empirical support for specific biological foundations underlying Bowlby's attachment theory, which went on to become the predominant view of social and emotional development over the latter third of the twentieth century. But Harlow's contributions to, and influences on, both psychology and a remarkably diverse range of other disciplines went far beyond his legacy as the father of the surrogate mother. His numerous ground breaking discoveries over a lifetime of research with monkeys have not only stood the test of time and scientific debate, but indeed provided the very basis for many new areas of scientific inquiry that still remain active in the year 2001.

Harry F. Harlow was born Harry F. Israel on October 31, 1905 in Fairfield, Iowa, where he spent his childhood years. He left this small Midwestern town to go to college, initially matriculating at Reed College for a year (as an English major) before transferring to Stanford University, where he changed his major and then earned both his BA (1927) and Ph.D. (1930) in psychology. His graduate advisors at Stanford, concerned that his family name might restrict his academic employment opportunities during a time of economic depression and latent anti-Semitism, suggested a name change. Harlow obliged, and later in life he often quipped that they had given him not only his degree but his very name as well. In the fall of 1930 he was hired as an Assistant Professor in the Department of Psychology at the University of Wisconsin in Madison, where he remained for the rest of his professional career. Following his retirement in 1974 Harlow moved to Tucson, Arizona, where he served as a Emeritus Professor of Psychology at the University of Arizona until his death on December 3, 1981.

Although Harlow's graduate training in comparative psychology had been largely limited to research with rodents, he began working with non-human primates shortly after his arrival at Wisconsin, carrying out his initial studies with a variety of monkeys and apes then residing in Madison's public zoo. He quickly discovered that the standard types of conditioning and memory tests then being used in studies with rats or pigeons were simply not adequate to tap the considerable cognitive capabilities of his non-human primate subjects. He and his students therefore began creating new tasks and test batteries

that were substantially more challenging. Shortly after moving into his own laboratory building on the University campus, Harlow designed and built the first prototype of a apparatus that would profoundly alter the course of research in primate cognition and that continues to be used in primate laboratories around the world in the year 2001. This testing device was named the Wisconsin General Test Apparatus (WGTA); it enabled an experimenter to present a large number of discrete learning and memory tasks in a rapid and highly standardized fashion to subject after subject (Harlow and Bromer 1938). At the same time, he began directing most of his efforts away from broad cross-species comparisons to more focused studies using a single species—the rhesus monkey (*Macaca mulatta*), an unusually hardy and adaptive species that thrived in captivity.

Armed with a laboratory specifically designed to house and test primates, in which individual monkeys could be maintained in good physical health for year after year, and a testing apparatus that enabled the same monkey to be tested repeatedly under standardized conditions, Harlow began changing the field of animal learning. In the 1940s he and his students created a battery of learning and memory tasks that provided a standardized 'intelligence test' for his monkeys. He then proceeded to study cortical localization of learning capabilities by lesioning different brain areas and recording subsequent differential patterns of deficits across the various components of the test battery. This research established his credentials as a physiological psychologist and it opened up areas of inquiry and an experimental methodology that continue to be pursued in neuroscience research.

In 1949 Harlow made a major conceptual and methodological breakthrough with his discovery of 'learning sets' in rhesus monkeys. He found that monkeys presented with a long series of six-trial two-choice discrimination problems soon learned to achieve near-perfect performance on the second and all subsequent trials of each problem. This demonstrated unequivocally that the monkeys had acquired a general strategy for problem solution, i.e., an abstract concept (in his words, 'learning to learn'), rather than a product of simple associative learning (Harlow 1949). This finding rendered untenable the neo-Decartes characterization of animals as 'reflex machines,' and it ultimately has led to increasingly more sophisticated studies of primate cognitive capabilities that foster direct comparisons with those of young human subjects. Largely because of this discovery, Harlow was elected to the National Academy of Sciences in 1951.

Harlow became interested in the ontogeny of rhesus monkey cognitive capabilities in the early 1950s when he developed a program of research designed to precisely establish when specific cognitive capabilities first emerged during development and when they subsequently achieved adult levels. He demonstrated

that infant monkeys could be both classically and operantly conditioned and performed associative learning tests at adult levels as early in life as such tests could be administered. In contrast, monkeys did not achieve adult levels of performance on more abstract problems such as learning set and oddity until their juvenile and adolescent years, respectively (Harlow 1959). In the process of carrying out these studies he also observed that young monkeys would solve problems and otherwise engage their physical environment in the absence of appetite rewards. This led him to conclude that in many circumstances curiosity was a more powerful motivator than a reduction of any primary drive (Harlow 1953). These observations provided the conceptual foundation for the 'surrogate mother' studies cited above.

In the 1960s Harlow shifted the major focus of his research from primate cognition to the development of social behavior, both normal and abnormal. He developed the concept of 'affectional systems,' in which he characterized emerging relationships formed by monkey infants with their mothers, peers, and others in the social group as differing in qualitative fashion and serving complementary roles in the socialization process (Harlow and Harlow 1965). He and his students also studied the consequences of blocking the formation of different affectional systems via various degrees and duration of social deprivation initiated at birth (Harlow and Harlow 1962, 1969). These studies established convincingly the importance of early social experiences for the development of species-normative adult social activities, including reproduction and maternal behavior, and they provided the conceptual and methodological basis for current studies of early experience effects on brain as well as behavior. In recognition for this research Harlow was awarded the National Medal of Science, America's highest official scientific award, in 1967.

Harlow continued his investigations of abnormal behavior during his final years at Wisconsin, generating two more areas of research in the process. First, he and his students demonstrated that the devastating behavioral consequences of total social isolation for the first 6 months or year of life, previously thought to be permanent, could be largely reversed by exposing the monkeys upon emergence from isolation to conspecific 'therapists' deliberately chosen to not only be socially normal but chronologically younger than the isolates themselves (Suomi and Harlow 1972, Novak and Harlow 1975). These findings directly refuted a strict 'critical period' interpretation of social deprivation effects, providing compelling evidence of developmental plasticity and once again opening up a new area of study that remains at the forefront of developmental neuroscience research today. At the same time, Harlow set out to create a rhesus monkey model of human affective disorders, relying on earlier observations by Rene Spitz and Bowlby that certain forms of social

deprivation were associated with infant and childhood depression. He and his students used a variety of social separation procedures to precipitate analogous behavioral states in young monkeys and then employed both social and pharmacological interventions to reverse to behavioral disorders (Harlow and Suomi 1974, Suomi and Harlow 1977). In recognition of these efforts, Harlow received the Kittay Scientific Award in 1975, the most prestigious award in the field of psychiatry at that time.

Harlow's influence on the field of psychology extended far beyond his empirical contributions, which resulted in over 323 original scientific publications over his 40-year career (a complete bibliography of Harlow's works may be found in Suomi and LeRoy 1982). He served as Editor for the *Journal of Comparative and Physiological Psychology* for over a quarter century. He also took an active role in many scientific societies, including the National Academy of Sciences, the Society for Experimental Psychology, and the American Psychological Association, for which he was elected President in 1958. Harlow served as the Chief Psychologist for the US Army during the Korean War, and he subsequently was appointed to numerous US government planning and review committees. He played a major role in the establishment of the US Regional Primate Research Center System in the early 1960s and was instrumental in establishing the University of Wisconsin as the site for one of the 7 original Centers, which he directed from 1964 from 1971.

Despite all of his research activities and administrative responsibilities, Harlow maintained an active teaching role throughout his career. Not only did he train numerous graduate students and postdoctoral fellows (a total of 35 students completed their Ph.D. dissertations under his tutelage), but he was also legendary as an undergraduate instructor, especially for Introductory Psychology, which he taught nearly every year during his tenure at Wisconsin. In addition, Harlow worked hand-in-hand with veterinary and technical personnel to establish numerous basic advances in primate husbandry, including the development of the first self-sufficient primate colony, the first large-scale primate neonatal nursery, and various quarantine and screening procedures that are still used.

Harlow was a colorful character who enjoyed dealing with controversy. He delighted in generating empirical findings that directly challenged existing theoretical dogma. In addition to his considerable skills as a writer and editor, he was a gifted public speaker, who for years was one of the most sought-after scientific lecturers in the country. As a result, he was able to present his findings not only to psychologists around the world but also to representatives of numerous other scientific disciplines, as well as to lay audiences. Indeed, 20 years after his death Harlow's scientific contributions were still cited not only by

comparative, physiological, developmental, cognitive, and clinical psychologists but also by others in fields ranging from anthropology to zoology. Clearly, he was much more than the ‘father of the cloth mother.’

See also: Animal Cognition; Learning to Learn; Psychologism, History of

Bibliography

- Harlow H F 1949 The formation of learning sets. *Psychological Review* **56**: 51–65
- Harlow H F 1953 Mice, monkeys, men, and motives. *Psychological Review* **60**: 23–32
- Harlow H F 1958 The nature of love. *American Psychologist* **13**: 673–85
- Harlow H F 1959 The development of learning in the rhesus monkey. *American Scientist* **47**: 459–79
- Harlow H F, Bromer J 1938 A test-apparatus for monkeys. *Psychological Record* **2**: 434–6
- Harlow H F, Harlow M K 1962 Social deprivation in monkeys. *Scientific American* **207**: 136–46
- Harlow H F, Harlow M K 1965 The affectional systems. In: Schrier A J, Harlow H F, Stollnitz F (eds.) *Behavior of Nonhuman Primates*. Academic Press, New York, Vol. II pp. 287–334
- Harlow H F, Harlow M K 1969 Effects of various mother-infant relationships on rhesus monkeys behaviors. In: Foss B M (eds.) *Determinants of Infant Behaviour*. Methuen, London, Vol. IV pp. 15–36
- Harlow H F, Suomi S J 1974 Induced depression in monkeys. *Behavioral Biology* **12**: 273–96
- Novak M A, Harlow H F 1975 Social recovery of monkeys isolated for the first year of life: I. Rehabilitation and therapy. *Developmental Psychology* **11**: 453–65
- Suomi S J, Harlow H F 1972 Social rehabilitation of isolate-reared monkeys. *Developmental Psychology* **6**: 487–96
- Suomi S J, Harlow H F 1977 Production and alleviation of depressive behaviors in monkeys. In: Maser J, Seligman M E P (eds.) *Psychopathology: Experimental Models*. W H Freeman, San Francisco, CA pp. 131–73
- Suomi S J, LeRoy H A 1982 In memoriam: Harry F. Harlow (1905–1981). *American Journal of Primatology* **2**: 319–42

S. J. Suomi

practice at the Chancery Bar throughout the 1930s. In 1940, he became a civil servant working in military intelligence; in 1941 he married Jenifer Fischer Williams, a civil servant in the Home Office, who was later herself to become an academic.

During the war, Hart’s keen interest in philosophy was stimulated by regular conversations with Stuart Hampshire and Gilbert Ryle, who were working in a related Department, and who introduced him to the new movements in linguistic philosophy. In 1945, after some hesitation, but encouraged by his close friend Isaiah Berlin, Hart accepted a fellowship in philosophy at New College. Despite his 16 years outside the academy, he established himself quickly as a leading member of the influential group of analytic philosophers, including Ryle, J. L. Austin, G. A. Paul, and Friedrich Waissman, working in Oxford at the time. Such was his reputation that, notwithstanding having published only one article and a book review, he was in 1952 elected Professor of Jurisprudence at University College, Oxford. During his tenure of the Chair, he revived what had become an intellectually sterile subject, bringing to jurisprudence the rigor of a philosophical method of which it had been deprived since the days of Bentham and John Austin, and introducing the fresh insights of Wittgenstein’s and, primarily, J. L. Austin’s philosophy of language.

Outside the UK, his reputation was especially strong in the US, where he spent several periods, notably at Harvard in 1956–7, and in Israel, which he first visited in 1964. He was elected President of the Aristotelian Society in 1959 and Fellow of the British Academy in 1962. In 1968 he resigned from the Chair, taking up a Senior Research Fellowship at University College, Oxford, in order to devote himself to the editing and analysis of the works of Bentham. From 1967–73 he served as a member of the Monopolies Commission; he also chaired a committee that made important recommendations for the reform of Oxford’s governance in the light of student unrest in the late 1960s. From 1973–8 he was Principal of Brasenose College; following his retirement he returned to University College as an honorary fellow. He died in Oxford on December 19, 1992.

Hart, Herbert Lionel Adolphus (1907–92)

Herbert Lionel Adolphus Hart was born on July 18, 1907 in the spa town of Harrogate where his family, which was Jewish, ran a furrier’s business. Hart’s was the first generation of the family to receive a university education: he studied at New College, Oxford, where he took an outstanding First Class degree in Greats (ancient history and philosophy) in 1929. He then qualified as a barrister, and enjoyed a successful

1. Hart’s Contribution to Legal, Moral, and Political Philosophy

Hart’s intellectual contribution lies in two distinct but related fields. First, he revived the tradition of analytical jurisprudence, exploring both the concept of law in general and a number of specific concepts—causation, rights, obligation, responsibility, rules—which structure legal reasoning and doctrines. In this work, Hart emphasized J. L. Austin’s assertion that one may use a ‘sharpened awareness of words to sharpen our perception of phenomena.’ He insisted,

however, that the task was not one of simple definition: rather, the aim was to build a theory or elucidation of the relevant concepts by examining their operation within the practical contexts in which they are used, and paying heed to the ‘open texture’ of language, with its core meaning surrounded by an unsettled penumbra. Perhaps the most outstanding example of these methodological commitments lies in Hart’s and Honoré’s monumental *Causation in the Law*, first published in 1959, which sets its conceptual analysis of causation within the context of hundreds of legal cases, and which draws criteria for the identification of causes out of the common sense understandings which underlie everyday usage in the relevant contexts.

Hart’s second contribution was in the fields of normative legal, moral, and political philosophy. Building out from his analytic conceptions, he made highly influential interventions in normative and policy debates about matters such as capital punishment, abortion and the regulation of homosexuality. His treatment of these issues was located within a tradition stretching back to Hobbes, Hume, Locke, and, most obviously, J. S. Mill, to which Hart contributed, among other things, an elaboration of principles limiting the proper scope of state action, a novel account of the justification of punishment, and an analysis of the relationship between utility and rights.

To understand why Hart’s work in analytical jurisprudence had such an extraordinary impact, one has to understand something about the development of jurisprudence in the century preceding his election to the Oxford Chair. From the late eighteenth to the mid-nineteenth century, the utilitarian thinkers Jeremy Bentham and his disciple John Austin had developed an influential theory of law known as ‘legal positivism.’ Rejecting the idea that law derives its authority from God, or from some metaphysical conception of nature or reason—so-called ‘natural law’—Bentham and Austin argued that law is essentially man-made: it is a command, backed by sanctions, issued by a political superior or sovereign, to whom the populace is in a habit of obedience. This positivist theory had, however, been little developed since Austin’s death. The one exception was the brilliant Austrian jurist Hans Kelsen, whose theory of law as a system of norms was, however, mainly influential on the continent of Europe. In Britain, the slow development of legal education in universities contributed to the impoverishment of the jurisprudential tradition, which in the first half of the twentieth century concerned itself primarily with the decontextualized analysis of legal terms such as ownership, possession, or the corporation. This analysis was not enriched by any systematic exchange with philosophy, nor was there any attempt to link it to any broader idea of the nature of law, or to consider how technical legal concepts assisted law to serve its various social functions. In the USA, by contrast, jurisprudence had moved in a sociological direction which arguably

emphasized the empirical at the expense of a sustained examination of the conceptual dimensions necessary to an analysis of law’s distinctive normativity. Prescriptive questions about what purposes law ought to pursue were left to the attention of moral and political philosophy—the latter itself a field that was relatively stagnant at this time.

Hart’s first single-authored book, *The Concept of Law* (1961), which stands as his most important work, claimed to provide a general, descriptive theory of law which was at once a contribution to ‘analytical jurisprudence’ and to ‘descriptive sociology.’ In other words, he sought to elucidate a concept of law that would be relevant to an understanding of all forms of law, wherever or whenever they arose. In pursuing this project, Hart returned to the insights of Austin and Bentham, but—in a crucial innovation—combined their methods with those of the new linguistic philosophy represented by J. L. Austin and Ludwig Wittgenstein.

The nub of Hart’s theory is the startlingly simple idea that law is a system of social rules structurally similar to the rules of games such as chess or cricket. The rules are of different kinds, with complementary functions. Some—‘primary rules’ such as criminal laws—directly govern behavior; others—‘secondary rules’—provide for the identification, interpretation, and alteration of the former. Distinctive among these ‘secondary’ rules is the ‘rule of recognition.’ This rule provides criteria by which officials identify valid rules belonging to the legal system. It, therefore, also serves to identify the content of one particular legal system. Yet it is not, Hart argued, itself a valid rule of the system. Rather, it exists in the constellation of official attitudes and practices. The existence of a legal system is hence, fundamentally, a complex matter of social fact.

Like his nineteenth-century counterparts, Hart therefore insisted that law is a social, human invention: though legal rules generate genuine obligations, they are not moral rules. Their authority derives not from their content but from their source, which lies in a distinctively institutionalized system of social recognition. For example, the rule that we should drive on the left is authoritative not because there is any intrinsic value to driving on the left. Rather, it is because the rule can be identified in accordance with an agreed set of criteria for recognition, such as parliamentary enactment or judicial precedent. Precisely the same is true, moreover, of legal rules that overlap with moral standards: the legal prohibition on murder is not the same as, and derives its validity in a different way from, the moral injunction against killing. Hart also teased out, however, what he saw as important grains of truth in the natural law tradition. While he insisted that the connection between law and morality was a contingent one, he also pointed out that a rule-based legal order brings with it a distinctive conception of formal justice. He also argued that,

given certain features of human being—our physical vulnerability and limited altruism, and the limited resources of the natural world, for example—legal systems, given their purpose of facilitating human survival, would have to include a core set of rules governing—justly or unjustly—matters such as physical integrity and property. This he termed, perhaps misleadingly, ‘the minimum content of natural law.’

Hart’s account of how legal rules are recognized as valid served, on the other hand, to distinguish law from a mere system of force, or ‘orders backed by threats.’ For—crucially—according to Hart, legal rules have not only an external but also an internal aspect: we know that a rule is in existence not only because it is regularly observed, but also because those subject to it use it as a reason or standard for behavior, criticizing themselves or others for breaches of the rules. Here the linguistic philosopher’s focus on the understandings embedded in ordinary usage becomes important. For example, Hart explores the distinction between habitual behavior (going to a café on Sunday morning) and rule-governed behavior (going to Church on Sunday morning); between being obliged to do something (handing over money because someone threatens to kill you if you refuse) and having an obligation to do it (paying your taxes). Of crucial importance to his status as a positivist is his insistence that the decision to adopt an ‘internal attitude’ to legal rules may be based on a variety of reasons, ranging from moral commitment to mere pragmatism: indeed, for a legal order to exist, it is theoretically only necessary for the officials of the system to adopt an internal attitude to its rules. In his elucidation of the internal attitude, Hart claimed to have detached the concept of law from that of sanction: the normativity of law derives from the internal aspect of rules and not from the threat of punishment for their breach.

By moving from the early positivist notion of law as a sovereign command to the notion of law as a system of rules, Hart produced a theory that resonated with the impersonal idea of authority embedded in modern democracies, providing a remarkably widely applicable rationalization of the nature of law in a secular and pluralistic world. He also offered an account of legal validity that purported to explain the (limited) sense in which citizens have an obligation to obey the law. Unlike the ‘command’ theory, it accounts not only for laws which impose duties but also the wide variety of laws which confer powers upon or create normative facilities for citizens: the law of property, contract, marriage, or succession, for example.

Hart also succeeded in reviving the idea that the project of jurisprudence should be not only to provide a clear, descriptive concept of law; but also—though as a separate project—to elaborate a framework for systematic criticism of law and for law reform. Hart’s work became increasingly concerned with these prescriptive questions: from the late 1950s on he published highly influential papers dealing with issues such as the

punishment of those who had committed what would now be regarded as serious crimes in Nazi Germany; capital punishment, the decriminalization of homosexuality, and the legalization of abortion. In these works, and in his widely publicized debate with the conservative judge Patrick Devlin, he regenerated a liberalism that finds its roots in J. S. Mills’s *On Liberty* (1859). In *Law, Liberty and Morality*, Hart (1963) argued, as against Devlin, that democratic states are not entitled to enforce moral standards for their own sake. With the exception of special cases where paternalistic legislation can be justified, or where grave public offence is at issue, the state should respect individual freedom, intervening only to punish the commission of tangible harms. This book stands, over 30 years after its publication, as the resounding late twentieth-century statement of the most influential genre of liberal social policy. The philosophical ideas on which it rests are articulated closely with those expressed in Hart’s conception of rights as founded in respect for individual human liberty and choice (‘Are there any natural rights?’), and in his argument that consequentialist aims of punishment should be tempered by considerations of individual fairness in the distribution and quantum of punishment (Hart 1968).

2. Hart’s Importance for the Social Sciences

Since the mid-1960s, Hart’s work has been subject to intense scrutiny and criticism. The most significant criticism has come from two different directions. First, from the direction of ‘natural law’ theory, it has been argued that, ultimately, the reason why those subject to legal rules adopt an ‘internal attitude’ to them has to do with moral factors, and that consequently Hart’s restricted sense of ‘legal obligation,’ and with it his distinction between law and morality, collapse. Second, from the direction of the social sciences, it has been argued that although the techniques of analytic philosophy may establish that law is simply one form of social rule, the further question of just what is distinctive about legal as opposed to other social rules can only be understood in terms of historical and social facts in which Hart shows surprisingly little interest. Notwithstanding these criticisms, there is strong reason to think that Hart’s work has real significance for the social sciences.

First, in reviving the tradition of descriptive legal theory—the attempt to provide a systematic elucidation of law—Hart illuminated an institution whose distinctive importance in the organization of modern societies has become a fundamental preoccupation of the social sciences. In spite of his universalistic philosophical ambitions, and of the undoubted tension in his work between his ‘analytic’ and his ‘descriptively sociological’ methods, his theory can be argued to provide a persuasive account of the distinctive nature of legal authority—law’s normativity—while main-

taining that law's existence is, ultimately, a matter of social fact and hence susceptible of empirical inquiry.

Second, therefore, in elucidating the nature of law as a social institution, Hart arguably provides precisely the sort of legal theory which social scientists need in incorporating law into broader social analyses. In his insistence on the structural continuity between legal and other social rules yet on law's distinctive mode of institutionalization, Hart provides a vision that is at once analytically well delineated yet relatively modest in terms of its substantive assumptions. Hence, the feature of his theory which might at first sight seem least sympathetic from a social science point of view—its pretension to provide a universal theory of law—may in fact be argued to be one of its strengths. For it provides a clear model for the purposes of testing hypotheses about, for example, the relationship between legal and other forms of social authority. This idea of his concept of law as, at root, a social science model is backed up by his own insistence, in terms reminiscent of Weber (though, in a significant ambiguity, also of Aristotle), that he is delineating a 'central case' with reference to which other, 'penumbral' cases can be identified and analyzed rather than 'banished to another discipline.'

Third, in his insistence on the internal aspect of rules, and hence on the idea that the social phenomenon of law can only be understood by reference to the attitudes and understandings of the agents who are its subjects, Hart's thinking meshed with a cluster of theoretical developments in disciplines such as sociology and anthropology. In this respect, his theory is a vast step forward from the cruder empiricism of the nineteenth-century positivists. Although it is not something with which Hart himself was centrally concerned, this feature of his method undoubtedly opens up the possibility of an analysis of the symbolic aspects of legal power and authority such as would be central to any adequate social science account of legal practices. In his emphasis on the importance of linguistic usage, moreover, his ideas relate, albeit obliquely, to the so-called 'linguistic turn' in much social theory of the late twentieth century.

Finally, in his revival of the utilitarian tradition of systematic critique of law and argument for its reform, Hart provided an example of the way in which conceptual analysis can be marshaled in aid of social policy. Clear thinking and reason—Hart's deepest intellectual commitments—are used, in an apparently modest and sometimes painstaking way, to clear the confusion in social and political debate, opening up in often unexpected ways paths forward in the development of policy and the construction of social consensus. Though Hart's own work is surprisingly little concerned with empirical—social, psychological, economic—evidence, his arguments often serve to clarify the relevance of just this sort of social science data. Though the fulcrum of his work lay firmly at the analytical-philosophic rather than the descriptive-

sociological end of the spectrum, he can, therefore, justly claim a place among those whose thinking has made a significant contribution to the social and behavioral sciences.

Bibliography

- Bayles M D 1992 *Hart's Legal Philosophy*. Kluwer, Dordrecht, The Netherlands
- Gavison R (ed.) 1987 *Issues in Contemporary Legal Philosophy: The Influence of H. L. A. Hart*. Clarendon Press, Oxford, UK
- Hacker P M S, Raz J (eds.) 1997 *Law, Morality and Society: Essays in Honour of H. L. A. Hart*. Clarendon Press, Oxford, UK
- Hart H L A 1955 Are there any natural rights? *The Philosophical Review* 64: 175–91
- Hart H L A 1961/1994 *The Concept of Law*, 2nd edn. Clarendon Press, Oxford, UK
- Hart H L A 1963 *Law, Liberty and Morality*. Stanford University Press, Stanford, CA
- Hart H L A 1965 *The Morality of the Criminal Law*. Oxford University Press, Oxford, UK
- Hart H L A 1968 *Punishment and Responsibility*. Clarendon Press, Oxford, UK
- Hart H L A 1982 *Essays on Bentham*. Clarendon Press, Oxford, UK
- Hart H L A 1983 *Essays in Jurisprudence and Philosophy*. Clarendon Press, Oxford, UK
- Hart H L A, Honoré A M 1959/1985 *Causation in the Law*, 2nd edn. Clarendon Press, Oxford, UK
- MacCormick N 1981 *H. L. A. Hart*. Edward Arnold, London
- Martin M 1987 *The Legal Philosophy of H. L. A. Hart*. Temple University Press, Philadelphia, PA
- Mill J S 1859/1974 *On Liberty*. Penguin, Harmondsworth, UK

N. Lacey

Hate Crimes

1. What is a Hate Crime?

There is no universally accepted definition of what precisely constitutes a hate crime, nor agreement as to the precise conceptual boundaries that delimit hate crimes from mundane crimes. The sole factor stressed uniformly is the motivation of the perpetrator. In hate crimes the principal motivation is one of negative affect (*animus*) toward the victim because of his or her skin color, ethnicity, religion, sex, national origin, or sexual orientation, for example. It is this *animus* that drives the criminal behavior rather than more common motives of pecuniary gain, retribution, jealousy, or status seeking. Thus, any crime may become a hate crime if the perpetrator's motivation is *animus* toward the victim based on some socially constructed category.

Some care must be taken to differentiate two seemingly related circumstances: (a) situations where the fundamental cause of the behavior is hatred; and (b) situations where hatred is coincident, with the primary motivation being something other than hatred. Presumably, it is only a true hate crime if the former condition exists. In other words, having extreme negative affect founded on skin color, ethnicity, religion, sex, national origin, or sexual orientation is both a necessary and sufficient condition for a hate crime.

In terms of behavior, hate crimes range from spray-paint vandalism and verbal harassment to physical assault and murder. The perpetrators may be single individuals, organized or unorganized groups, or entire nation-states. Similarly, the victims may be solitary individuals, small groups, or large aggregates of human population. In fact, a hate-crime universe can be conceptualized in terms of a three dimensional space: along one dimension is the degree of aggregation or organization of the perpetrator (ranging from single persons to nation-states), along another dimension is the degree of aggregation of victims (from single persons to nation-states), and the third dimension is the degree of severity (extending from vandalism to lethal sanctioning). Within this heuristic model, behavior can be located as diverse as a sole actor desecrating a synagogue with swastikas, to nation-states with genocidal policies aimed at despised populations. Yet underlying each is the assumption that the primary motive is negative affect based on a socially constructed status such as race or religion.

While the term hate crime, and the public notoriety it has received, has been particularly discernible during the 1980s and 1990s, the phenomenon is far from recent and has long been a frequent factor in human societies across the globe. This century alone has witnessed the genocide of Christian Armenians at the hands of Muslim Turks (1915–18), the Nazi attempt to exterminate European Jewry and the Roma, the mass murder of Vietnamese and Cham Muslims by the Khmer Rouge in Cambodia in the 1970s, the 1994 Hutus inspired genocide of Tutsis in Rwanda, and the Serbian ‘ethnic cleansing’ of Muslims from Bosnia. These are but a few examples of mass-hate crimes involving large numbers of perpetrators and with their victims experiencing the most severe levels of sanctioning. Often, however, public attention has been drawn to smaller-scale events involving fewer victims: for example, in 1991 skinheads in Hoyerswerda, Germany attacked several Asians and fire-bombed a hostel used by asylum-seekers (similar attacks were made on immigrants and refugees in many, if not most, Western European countries during the 1980s and 1990s)—the burning of some African-American churches in the American South during the early 1990s—or the 1998 lynching of James Byrd Jr, a 49-year old African-American in Texas by three white supremacists.

In sum, hate crimes can range from killing of genocidal proportions perpetrated by a nation state to verbal harassment by a single actor. They are frequent in human societies globally.

2. Why Hate Crime?

Most theories of hate crime involve, implicitly or explicitly, the notion of ‘threat’: that hate crime perpetrators perceive members of the target group as representing danger or peril to their life chances, lifestyles, social status, or belief systems. This threat does not have to be ‘real’ in any demonstratively physical sense—it only needs to be thought to be real to motivate some persons to commit crimes of hate.

Macropolitico-economic theories emphasize the importance of large-scale changes in the economic, political, or demographic arenas in generating apparent threats to the status quo. Typically these theories postulate that social change brings uncertainty, disruption, and competition between dominant and emergent groups for economic standing or political power. That contested terrain breeds fear and motivates some to strike out against members of the perceived threatening group. The effect of intergroup competition (threat) seems to be especially pronounced when the target group is culturally dissimilar from the dominant one. This is a point emphasized by cultural explanations of hate crime. Cultural theories stress threats to traditional or dominant belief-systems. Differences between the dominant culture and the target group subculture are seen as incompatible, if not totally hostile—the culture and belief systems of the victim group may be viewed as contaminating the dominant culture and in that sense, representing a threat to the cultural status quo. It is quite likely, for example, that the hate crime directed at guest workers, asylum seekers, and refugees in many counties was rooted in both the fear that new immigrants would take jobs away or lower wages of the native labor force and in the fear that a ‘foreign’ culture would fundamentally challenge traditional values and beliefs.

3. Hate Crime and Hate Speech Laws

Hate crime laws often take the form of sentence enhancements to penalties for crimes motivated by animus. For illustration, in the United Kingdom the 1998 Crime and Disorder Act provides that the maximum sentence for common assault increase from 6 months to two years if the assault was racially motivated. In the USA, the vast majority of states have enacted legislation providing for sentence enhancements for some hate crimes (but not others). At the federal level, the Hate Crimes Sentencing Enhancement Act of 1994 increased sentencing by about one-third when a crime is proven to have been hate motivated and occurred on federal property.

In many countries, hate speech, as well as hate crime, is criminalized. In the UK, speech or writings that may provoke racial animosity are prohibited by the Public Order Act. Similarly, Section 319 of the Criminal Code of Canada prohibits deliberately inciting or promoting 'hatred against any identifiable group.' There are comparable provisions in the criminal codes of Austria, Poland, Sweden, France, and Ireland. The Indian penal code not only criminalizes racial and religious-hate speech but caste-hate speech as well. In Germany it is illegal to incite violence or racial hatred against groups or minorities, and it is a crime to deny the Holocaust. Likewise, 1995 laws in both Switzerland and Belgium prohibit denying or defending genocide. There are no comparable laws against hate speech in the USA, however.

4. Trends in Hate Crimes

There is considerable rhetoric by advocacy groups that many countries are undergoing an unprecedented wave of hate crime, yet there is little undisputed empirical evidence supporting that claim. All authorities acknowledge that existing data on hate crime are fraught with errors, so any apparent rise in frequency of hate crimes may be due to: (a) a more sensitized public willing to report incidents of hate that had gone previously unreported; (b) the media and advocacy groups more frequently publicizing hate crimes; (c) an actual increase in the incidence of hate crimes; or (d) some combination of these. Where 'official' statistics are offered, it is suspected that they too may be seriously compromised because of under reporting of hate incidents by victims, the reliance upon local authorities to recognize and define an incident as a hate crime, and changes and ambiguities in the definition of what constitutes a hate crime. Given the shortcomings of official hate crime data and the unknown reliability and validity of anecdotal data from advocacy groups, it would seem impossible to state definitively whether hate crimes are actually increasing in frequency, decreasing, or remaining the same.

Given that caveat, what meager empirical data exists suggests that in the USA the frequency of hate crime incidents increased from the early 1990s until 1996, and has been declining since. In most of Western Europe, the trend appears that hate crimes may have become somewhat less frequent toward the end of the 1990s than they were earlier in the decade, with the possible exception of the UK.

5. Victims of Hate Crimes

While time trend data on hate crimes are especially problematic, a cross-sectional inspection of official reports of hate crime can illuminate who were the

victims. In many European countries during the time 1970–2000, reported hate crime has targeted immigrants, refugees, asylum-seekers, and guest workers. Additionally, there have been many attacks on Jewish cemeteries and synagogues. In the USA, the situation is somewhat different.

Based on hate crimes reported in 1998 by 10,730 local law-enforcement agencies covering a population of 216,235 million, the US federal government cites 7,755 incidents of hate crime involving 9,722 victims. Of these incidents, 56 percent involved race hatred, and of those over two-thirds were anti-black. While race-hatred incidents were the most common, incidents of religious hatred were frequent, and 78 percent of these incidents involved anti-Jewish activity. Reports of sexual orientation bias were also common, with 67 percent of these incidents targeted at male homosexuals. Of all the hate crimes reported in 1998, 62 percent (4,832 incidents) were incidents that were anti-black, anti-Jewish, or anti-gay male. Of those offenses where the suspected offender's race was known, whites committed 76 percent of the hate crime offenses.

6. Consequences of Hate Crimes

Being a victim of virtually any crime extracts a definite psychological toll, but there is some evidence that the emotional mark of hate crimes may be more severe and longer lasting than the impact of normal (nonhate motivated) crime. Furthermore, some argue that when one member of the targeted group is victimized by hate crime, then all members of the group are adversely affected—that is, the negative effect of hate crime permeates the victim community more thoroughly than other forms of victimization. Because of the very nature of hate-motivated crime, a poignant and unmistakable message is delivered to both the victim and his or her community.

See also: Disgust, Psychology of; Genocide: Anthropological Aspects; Genocide: Historical Aspects; Heterosexism and Homophobia; Mass Killings and Genocide, Psychology of; Personality and Crime; Racism, History of; Racism, Sociology of; Religion: Morality and Social Control; Religion: Peace, War, and Violence; Religious Fundamentalism: Cultural Concerns; Social Trauma; Violence and Media; Violence: Public; Xenophobia

Bibliography

- Berk R A 1994 What is a hate crime? In: Hamm M S (ed.) *Hate Crime: International Perspectives on Causes and Control*. Anderson, Cincinnati, OH, pp. v–ix

- Grimshaw A D 1999 Genocide and democide. In: Kurtz L, Turpin J (eds.) *Encyclopedia of Violence, Peace, and Conflict*. Academic Press, San Diego, Vol. 2, pp. 53–74
- Hamm M S 1994 Conceptualizing hate crime in a global context. In: Hamm M S (ed.) *Hate Crime: International Perspectives on Causes and Control*. Anderson, Cincinnati, OH, pp. 173–94
- Jones T D 1998 *Human Rights: Group Defamation, Freedom of Expression and the Law of Nations*. Martinus Nijhoff, The Netherlands
- Myra H I (ed.) 2000 *Ethnic Violence*. Greenhaven, San Diego
- Jacobs J B, Potter K 1998 *Hate Crimes: Criminal Law and Identity Politics*. Oxford University Press, New York
- Kelly R J, Maghan J 1998 Introduction. In: J Maghan R J, Maghan J (eds.) *Hate Crime: The Global Politics of Polarization*. Southern Illinois University Press, Carbondale, IL, pp. 1–21
- Lawrence F M 1999 *Punishing Hate: Bias Crimes under American Law*. Harvard University Press, Cambridge, MA
- Witte R 1995 Racist violence in Western Europe. *New Community* 21: 489–500

E. M. Beck

Hayek, Friedrich A von (1899–1992)

Friedrich von Hayek is regarded generally as the most influential intellectual figure in the twentieth-century revival of classical liberalism (Kukathas 1990) and as the most articulate critic of socialist doctrines (Caldwell 1997). He also deserves recognition as one of the preeminent social philosophers of that century. In an age of increasing specialization within the social sciences and the humanities, he developed an approach to social analysis and social reform that integrates insights from economics, law, politics, philosophy, psychology, and evolutionary theory into an impressively coherent outlook at the foundations and the evolution of social order. Hayek's scholarly writings span more than six decades and include numerous books and articles, republished in a 19-volume edition of his collected works (Hayek 1988).

1. Biography

Hayek was born in Vienna on 8 May, 1899. He earned doctorates in law (1921) and in political economy (1923) at the University of Vienna, where he became *Privatdozent* for Political Economy in 1929. Interrupted by a year of study in New York (1923–4) he worked from 1921–6 as a legal consultant in a government office in Vienna, directed by his mentor Ludwig von Mises, in whose famous *Privatseminar* he participated and with whom he founded, in 1927, the Austrian Institute for Business Cycle Research, of which he was director until 1931. A series of invited lectures that Hayek gave at the London School of

Economics (LSE) in 1931 led to his appointment to the Tooke Chair of Economic Science and Statistics at the LSE, a position Hayek held from 1932 to 1950, the year he was appointed Professor of Social and Moral Sciences at the University of Chicago's Committee on Social Thought. After retiring from Chicago in 1962 he taught as Professor of Economic Policy at the University of Freiburg, Germany, until 1967. An honorary professorship at the University of Salzburg from 1969 to 1974 brought Hayek back to his native Austria. In 1977 he returned as professor emeritus to Freiburg where he died on 23 March, 1992. In 1974 Hayek was awarded the Nobel Prize in Economics.

2. Hayek's Social Philosophy

About his personal development Hayek (1967, p. 91) has said that he was led 'from technical economics into all kinds of questions usually regarded as philosophical.' It was his early work on monetary theory, capital theory, and business cycle theory that earned him his appointment at the LSE, and it was about these themes that he engaged in controversies with John Maynard Keynes, Piero Sraffa and Frank H. Knight in the early 1930s. An important turning point in his scholarly interests is marked by his work, in the mid-1930s, on an English edition of selected contributions to the 'socialist calculation debate,' including Mises' early (1920) article on the subject (Hayek 1935). As Hayek has noted in retrospect, in preparing this edition he began to develop his growing interest in questions of the nature and unavoidable limitations of our knowledge, and in the issue of what these limitations imply for our theoretical efforts at understanding the social world as well as for our practical efforts at organizing our social life. The concern with these questions, often referred to as the 'knowledge problem,' was to become the core theme of Hayek's life work.

In Hayek's own assessment, the critical step in his move towards social philosophy was his 1937 article on 'Economics and Knowledge' (1948, Chap. 2), an article in which Hayek addressed what he describes as the 'central question of all social sciences,' the question of how the fragments of imperfect knowledge that exist dispersed in individual minds are coordinated in the social process to allow for productive cooperation among multitudes of persons. With his 1945 article on 'The Use of Knowledge in Society' (1948, Chap. 4), his most often cited publication, Hayek restated and extended his earlier argument, stressing that the 'problem of what is the best way of utilizing knowledge initially dispersed among all the people is at least one of the main problems of economic policy—or of designing an efficient economic system' (1948, p. 78). In these two articles, which he considered his most

original contributions to economics, Hayek developed a fundamental critique of two bodies of thought prevalent at the time, of formal equilibrium economics on the one hand and of socialist concepts of central planning on the other, censuring both of them for failing to account for the theoretical and practical difficulties that are posed by the imperfect, fragmentary, and dispersed nature of human knowledge.

In Hayek's view, the models of mathematical equilibrium economics provide only a pseudo-answer to the question that they are meant to illuminate, namely, how markets work. As he put it, with its 'assumption of a perfect market where every event becomes known instantaneously to every member' (1948, p. 45), formal equilibrium analysis begs the very question that an empirically contentful theory of market processes would have to answer, namely how 'a solution is produced by the interaction of people each of whom possesses only partial knowledge' (1948, p. 91). Ironically, the failure committed in concepts of central planning is of a very similar kind. By assuming perfect knowledge on part of the planners, advocates of central planning provide only a pseudo-answer to the practical problem of how effective coordination of economic activities can be achieved in the absence of markets. Assuming that all relevant knowledge is available to, and can be administered by, a planning agency, raises the question of how such an agency is supposed to come to terms with the fact that much of the relevant knowledge 'never exists in concentrated or integrated form, but solely as the dispersed bits of incomplete... knowledge which all the separate individuals possess' (1948, p. 77), as knowledge that can be activated only by these very individuals.

What is needed, Hayek argued, is a theory that accounts for 'the unavoidable imperfections of man's knowledge and the consequent need for a process by which knowledge is constantly communicated and acquired' (1948, p. 91). He called on his colleagues in the economics profession to turn their attention from the formal statements of equilibrium analysis to empirically contentful 'propositions about what happens in the real world' (1948, p. 46), propositions about how knowledge is actually acquired and communicated. While his call did not find much response in the profession, in his own work Hayek very much pursued the research agenda that he suggested.

Hayek has especially addressed the issue of how we acquire knowledge in his *The Sensory Order* (1952a), a most unlikely book for an economist and one of his least known, but nevertheless one that is of greatest systematic significance for his entire work. In this book (based on an early manuscript from Hayek's student days), subtitled 'An Inquiry into the Foundations of Theoretical Psychology,' Hayek develops a theory of the human mind, the essence of which is in amazing accord with modern cognitive science. Seeking to explain how the human mind establishes its connections 'between the input of (external and

internal) stimuli and the output of action' (1988, p. 288), Hayek interprets the working of the mind as an adaptive process of constant classification and re-classification on many levels (sensory perceptions, emotions, concepts), as a process in which conjectural internal models of the outside world are adaptively formed and guide actions.

The issue of how knowledge is communicated and advanced in society is the theme of Hayek's theory of the market as a communication system (1976, Chap. 10) and of competition as a discovery procedure (1978, Chap. 12), a theory that forms the centerpiece of his work. The prices formed in the competitive market process, Hayek argues, signal to all market participants the relevant general information about changes in relative scarcities that they need to know in order to adjust their behavior adequately, utilizing their unique local knowledge of the particular circumstances of time and place, knowledge that could never be communicated to, and utilized by, a central planning agency. Beyond serving in such fashion as a device for utilizing dispersed knowledge, the market serves also as an arena for the continuous competitive exploration of new and potentially better solutions to economic problems, thus inducing the discovery and creation of new knowledge.

The relevance of Hayek's work, not only for economics but for the social and behavioral sciences more generally, is most clearly visible in his theories of spontaneous social order and of cultural evolution that he has developed as generalizations and extensions of his arguments on market processes and competition. The thrust of these theories is that human societies are highly complex phenomena, and that the processes of social coordination and societal evolution are inappropriately simplified and misconceived if they are approached in the spirit of the mechanistic models of physics (1952b). He argues that acknowledging the complexity of social phenomena means to acknowledge that all we can reasonably hope to achieve in the social realm, as in all areas of complex phenomena, are explanations of the principle and pattern predictions, by contrast to the explanation and prediction of specific events (1967, Chap. 2).

Hayek places his own social theoretical outlook in the tradition of the Scottish moral philosophers (David Hume, Adam Smith, Adam Ferguson and others), whom he credits with the discovery of 'the twin concepts of the formation of spontaneous orders and of selective evolution' (1988, p. 146). Central to this paradigmatic tradition is a methodological individualism and subjectivism, that is, on the one hand, the notion that social aggregate phenomena are, ultimately, to be explained in terms of the actions and interactions of individual human beings and, on the other hand, the insight that 'the analysis of what people do can start only from what is known to them' (1948, p. 44), that is, from their subjective view of the world. Equally central to it is the notion that the

productive coordination and division of labor that we observe in society is to a large measure not due to deliberate planning and conscious design, but to forces of spontaneous mutual adjustment and evolution, as are the rules and institutions on which societal cooperation is based. From these premises important conclusions are drawn with regard to the limits of our ability to improve our social condition by deliberate construction.

According to Hayek, there are essentially two basic ways in which the actions of individuals can be socially coordinated, namely either through spontaneous mutual adjustment within a framework of general rules of conduct, or through the commands of a central coordinator (1973, Chap. 2). As he argues, the respective merits and limits of the two kinds of order—spontaneous order and organization—have to do with the nature of their principal means of coordination, namely general rules versus commands. Command-based organization is an effective method of social coordination where specific, limited purposes are to be achieved. Because of its reliance on central decisions it is, however, severely limited in its capacity to utilize the knowledge dispersed among its individual participants. The advantage of rule-based spontaneous social orders is that, by allowing individuals to pursue their own ends within the general confines defined by the rules, they can utilize and activate to a much higher degree localized, dispersed knowledge. The price that has to be paid for their greater capacity for utilizing knowledge is that the particular outcomes that they generate must inevitably remain largely unpredictable. In Hayek's view, considering the respective merits and limits of the two kinds of order should lead us to conclude that, while organization can be a useful method for solving a multitude of specific, limited tasks, it cannot possibly be appropriate for the extended order of an economy or a society at large. There is ample room for organizations within the spontaneous order of society, but efforts to turn society itself into an organization can only end in disaster.

An important element in Hayek's theory of spontaneous social order is the insight that, while the particular outcomes of a rule-based spontaneous order remain necessarily largely unpredictable, the general patterns of its outcomes are determined systematically by the nature of its rules. There exists, Hayek emphasizes, a systematic interplay between the 'order of rules' and the 'order of actions' (1967, Chap. 4), that is, between the nature of the 'rules of the game' on which the interactions of the participating individuals are based, and the nature of the action-patterns that emerge from the individuals' behavioral choices within the rules, just as the rules of a sport allow us to predict—if not particular outcomes—the general patterns that we observe on the playing field. In fact, Hayek argues, our knowledge of the systematic relation between the order of rules and the order of

actions is the principal instrument that we can employ to improve our condition, by choosing rules of the game that can be predicted to generate more desirable overall patterns of outcomes.

Hayek's emphasis on the basic facts of 'the dispersion and imperfection of all knowledge' leads him to be highly critical of a mind set that he labels 'rational constructivism' or 'constructivist rationalism' (1978, Chap. 1). Under this rubric he subsumes approaches to issues of social policy and social reform whose recommendations are based on a 'pretence of knowledge' (1978, Chap. 2), that is, on unjustified presumptions concerning the extent of our knowledge and our intellectual powers. The primary targets of Hayek's critique are approaches that, like socialist concepts of central planning, amount to the 'claim that man can achieve a desirable order of society by concretely arranging all parts in full knowledge of all the relevant facts' (1967, p. 88). Against such claims Hayek points out that, because it is simply impossible for us to take 'conscious account of all the particular facts which enter into the order of society' (1973, p. 13), we need to rely on general rules 'as the means to create order in social affairs' (1948, p. 19), and that our efforts at building a better society must largely rely on the indirect method of improving the rules and institutions under which we live, instead of on attempts to directly produce desired outcomes through specific interventions.

To be sure, in our efforts to improve the framework of rules and institutions under which we live, we need, Hayek stresses, to be no less careful to avoid the pitfalls of a pretense of knowledge. Since the desirability of the outcome patterns that particular rules or institutions tend to generate can only be judged appropriately in terms of their working properties over extended periods of time, it would be presumptuous to claim that we can easily design, at the drawing board, rules and institutions that serve us better than those that have evolved over generations. Just as, in the ordinary market, the discovery process of competition helps us to find out how our 'economic' problems can be solved better than before, competition in the realm of rules and institutions can serve the function of allowing us to find out what kinds of rule arrangements are better suited to allow for mutually beneficial cooperation among multitudes of persons. A process of competitive selection can here, just as in ordinary markets, help gradually to evolve solutions that are much more effective than what any rational design could have ever invented at the outset. In fact, so Hayek argues, the institutions on which our civilization rests, including the rules of the market, must be viewed as the outcome of a process of cultural evolution that extends over thousands of generations and that reflects the experience of many more trials and errors than the most wise and intelligent rule designer could ever consider. Our efforts at institutional reform and rule design should, therefore, always

be informed by due respect for the potential wisdom implicit in inherited institutions, and we should expect further improvement ‘more from slow experimental piecemeal evolution than from redesign of the whole’ (1964, p. 8).

The theory of cultural evolution is the theme to which Hayek devoted much of his later work (1979, Epilogue, 1988), and it is this part of his theory that has become a primary focus of discussions on his ideas. A central issue in these discussions is how the notion of competition among institutions—a notion that is at the heart of Hayek’s ideas on cultural evolution—is to be specified, and under what conditions such institutional competition can be expected to generate ‘good’ rules, that is, rules with working properties that are desirable to those who are affected by them.

3. Impact and Significance of Hayek’s Work

With his arrival at the LSE in the early 1930s Hayek became almost instantaneously a well-recognized figure in the economics profession, engaged in prominent controversies on economic theory with such opponents as John Maynard Keynes, and on the issue of rational economic calculation under socialism with Oskar Lange and others. As his involvement in the latter debate induced the shift in his interest towards general issues of social and political philosophy, he increasingly removed himself from the standard professional discourse. He turned his attention to a fundamental critique of the philosophical foundations and the political implications of socialist doctrines, reflected most prominently in the publication of his most famous book *The Road to Serfdom* in 1944, a book that quickly made him a widely known public figure, in particular in the USA, where a condensed version appeared in the *Reader’s Digest* in 1945. While the book earned him public fame, it was met with considerable hostility among his colleagues in the economics profession, many of whom were, at the time, much more inclined to interventionist views than to Hayek’s warnings against socialist concepts of social and economic planning. With the rise of Keynesianism in the 1950s and 1960s, Hayek’s work fell almost into oblivion, in academia as well as in public debate, and when in 1960 his major restatement of the principles of classical liberalism, *The Constitution of Liberty*, appeared, it was widely regarded, even by sympathetic observers, as an outdated treatise, out of step with modern economic and political thought. It came as a surprise to many, including Hayek himself, when in 1974 he was awarded the Nobel Prize in Economics. Even if the fact that he had to share the prize with Gunnar Myrdal, an economist of a quite different orientation, reflected significant ambivalence on part of the Nobel committee, the

award was an early indicator of a reawakening interest in classical liberal thought in general, and Hayek’s work in particular, an interest that grew significantly in the later half of the 1970s, as a response to an increasing disillusionment with Keynesian economic policies and to accumulating evidence of the failure of ‘real socialism.’ In fact, Hayek became very much the symbol of what has been described as the ‘modern rebirth of classical liberalism.’ His work regained increasing attention in the academic community, his ideas influenced politicians like Margaret Thatcher and Ronald Reagan, and his publications found growing interest among dissidents and reformers in the disintegrating communist world. Hayek lived long enough to witness the final collapse of the communist regimes in Europe as a confirmation of his lifelong critique of socialist constructivism.

In modern theoretical economics Hayekian ideas play an important role in a number of approaches that seek to advance an alternative theoretical outlook to the orthodox equilibrium paradigm. His ideas on the nature of human knowledge and learning are finding increasing attention among scholars who seek a corrective for deficiencies of the traditional rational choice model. Recent developments in evolutionary economics, in the new institutional economics, and in the economics of complex adaptive systems, either explicitly draw on, or give support to core ideas of his work, to his theory of the market as a spontaneous order, his outlook on competition as discovery process, and his evolutionary approach to institutions.

The reception of Hayek’s work in social sciences other than economics is still in its infancy, yet his approach to social theory may well come to play a significant role as a paradigmatic contribution towards a theoretically integrated social science. It was one of Hayek’s major concerns that the disciplinary fragmentation of modern social science left what should be its central issue, namely ‘the problem of an appropriate social order,’ without adequate attention. As he put it, even though this issue ‘is studied today from the different angles of economics, jurisprudence, political science, sociology, and ethics, the problem is one which can be approached successfully only as a whole’ (1973, p. 4). More than any of his contemporaries in the social sciences Hayek has made an effort in his own work to address the fundamental issue of an ‘appropriate social order’ from a general social theoretical perspective that cuts across traditional disciplinary boundaries. With his ideas on the two kinds of social order, on the function of rules and institutions, and on cultural evolution he has made important steps towards a theoretically integrated social science.

See also: Economics, History of; Knowledge, Sociology of; Liberalism; Liberalism: Historical Aspects; Liberalism: Impact on Social Science; Socialism: Historical Aspects

Bibliography

- Caldwell B 1997 Hayek and socialism. *Journal of Economic Literature* 35: 1856–90
- Hayek F A (ed.) 1935 *Collectivist Economic Planning*. Routledge, London
- Hayek F A 1944 *The Road to Serfdom*. Routledge, London
- Hayek F A 1948 *Individualism and Economic Order*. University of Chicago Press, Chicago
- Hayek F A 1952a *The Sensory Order*. University of Chicago Press, Chicago
- Hayek F A 1952b *The Counter Revolution of Science*. Free Press, Glencoe, IL
- Hayek F A 1960 *The Constitution of Liberty*. University of Chicago Press, Chicago
- Hayek F A 1964 Kinds of order in society. *New Individualist Review* 3: 3–13
- Hayek F A 1967 *Studies in Philosophy, Politics and Economics*. University of Chicago Press, Chicago
- Hayek F A 1973, 1976, 1979 *Law, Legislation and Liberty*, 3 vols. University of Chicago Press, Chicago
- Hayek F A 1978 *New Studies in Philosophy, Politics, Economics and the History of Ideas*. University of Chicago Press, Chicago
- Hayek F A 1988 *The Collected Works of F. A. Hayek*. University of Chicago Press, Chicago
- Kukathas C 1990 *Hayek and Modern Liberalism*. Clarendon Press, Oxford, UK
- Mises L von 1920 Die Wirtschaftsrechnung im sozialistischen Gemeinwesen. *Archiv fuer Sozialwissenschaft* 47: 86–121
- Vanberg V J 1994 *Rules and Choice in Economics*. Routledge, London

V. Vanberg

Healing

In the context in which healing is studied here, it applies to any human or socially organized intervention having as its purpose the improvement of a physical or mental condition. In sociological analysis healing is judged by its intention, not by its effect. However, the differential success rates of opponent forms of healing are the basis of conflict between their practitioners and of the legitimization of the organizations within which they are practiced. The evidence is that while religious organizations are in other spheres of human life losing ground to secular agencies, there has been a recent increase of the resort to forms of healing that are alternative to secular medicine.

1. Types of Religious Healing

The widely used concept of shaman is a precise one, having been crystallized in a classic work by Mircea Eliade in the 1960s. A shaman is a person who enters a state of abnormal consciousness, being induced either by rhythmic music or by psychedelic drugs in the form of snuff or mushrooms. As a functionary in the

business of healing, the shaman may be distinguished from looser and more popular concepts such as witch doctor and medicine man that suggest skills in the use of locally available medications without the element of trance or ecstasy. Again, the role of the shaman in healing is not that of magician: whereas the magician uses ritual or other mysterious means to control the course of natural events, the shaman is merely a channel or borrower of the powers of the universe through whom some benign outcome is effected. The word shaman is taken from a tribe in north Siberia but shamanism within this narrow conception may be found also in, for example, the upper Amazon, the Orinoco basin, Australia, and in south-east Asia.

Mental healing, which is based on the diagnosis of illness as a condition of the mind, may be traced within modern Christian practice to the theology of Emanuel Swedenborg (1688–1772), a Swedish scientist and initiator of the religious sect which bears his name. He perceived healing as a harmony of physical and spiritual realities, achievable by the adjustment of the body to the mind, the mind to the soul, and the soul to the angelic order. In mid-nineteenth century America the Swedenborgian idea of mental adjustment was current in the fashionable mesmerism of Charles Poyen, in the teachings of Phineas Quinby, and subsequently in the writings of his most famous patient, Mary Baker Eddy, the founder of Christian Science. In her seminal *Science and Health with Key to the Scriptures* (1875) she turned to Matthew 9, 2–8 to demonstrate the central tenet that sickness and pain were delusions curable by the correction of mental error.

The notion of faith healing implies a degree of prior belief, even confident expectation, either in the seeker or in the healer. Healing which is reckoned to depend on the commitment of the patient lends itself to explanations in terms of the will power of the individual and other psychological factors. But the term faith healing is sometimes applied also to practices in which the healer claims authority over evil spirits or practices the laying on of hands in confidence of a miraculous outcome.

Divine healing, by contrast, is believed to be an intervention that does not rely on the faith of those involved. These conceptions of healing are prevalent in various of the Christian sects though not exclusive to them. Most notable among these are the ‘Four-square’ pentecostal churches of north America, one of the four squares being ‘Christ the Healer.’

In recent years there has been a burgeoning interest in healing in the mainstream churches in which the term healing is used in a more general way to include acts of reconciliation of individuals or groups, either as a result of conflicts between themselves or in changing circumstances such as tragedy and bereavement. The ministry of healing which is particularly associated with the Church of England does not connote the sense of the miraculous which distinguishes

divine healing and faith healing. It recognizes the skills of the minister who will accordingly remain conscious during the consultation.

The conciliation of mind and spirit that takes a variety of forms and is pursued by a still greater variety of methods is the essence of spiritual healing. It is not to be confused with spiritualist healing in which the practitioner is said to be in contact with spirits in another domain, such as that of the dead.

2. *The Social Context of Faith and Healing*

The pentecostal churches have flourished lately among the dispossessed of central and south America (Martin 1990, Chesnut 1997). The routine Marxist explanation is that the hope of a life hereafter enables the faithful to endure the discomforts of the present world. In Brazil, however, Chesnut found relatively less conviction in remote consolation than hope of immediate relief from the illnesses that attend poverty. The gravitation to the Pentecostal churches may to that extent be compared with the pilgrimages to Lourdes, to the sacred waters of the Ganges, and to the healing shrines and centers that abound in India and in the Catholic world.

What is more distinctive of recent years is the interest in ritual healing among the American middle class. Maguire (1988) recognizes in 'the metaphysical movement' the confluence of a number of nominally Christian traditions from New Thought, mesmerism, the mind-cure groups of the 1880s, Spiritualism and Christian Science, and an experimental interest in currents of spirituality that are or were native American, Latin American, Caribbean, and Polynesian. The turning to metaphysical healing in the suburbs has an intellectual basis involving an alternative pathology, a redefinition of illness and disharmony, and an open-minded disposition to eastern thought emanating from student subcultural movements in the late 1960s.

The host of life and death concerns which are addressed by hands-on therapy rituals in the religious life of middle America provide the agenda too for the ministry of healing in the Church of England. This neither trades on the expectation of the spectacular nor publishes its results against those of the medical services. The ministry operates, for example, in the rehabilitation of counseling and psychotherapy within the sacred domain, in the hospice movement pioneered in Britain by Dr Cicely Saunders, and in the provision of healing services within the regular ministrations of the mainstream churches.

Church-type healing is distinguishable from that of the sects in a number of ways: (a) the sectarian cult of the personality tends to elude church-type practitioners; (b) the expectation is not of the miraculous, nor are successes celebrated; (c) ministry to the dying will have the purpose of enabling them to die well, not

to recover to full health; (d) needs and benefits are largely reserved within the private domain; (e) the ministry is integrated with the Church's sacraments; (f) practitioners tend not to claim divine gifts so much as seek to enhance their effectiveness through conventional training of knowledge and skills; (g) the Church's ministry of healing is reckoned to complement conventional medicine and not to supplant it. While the miraculous healing tradition is rooted in the Pauline formulation of spiritual gifts—notably in Corinthians 1, 12—and denotes the healer as a direct agent of divine power, the Churches' ministry of healing looks to the Gospel for a motivation but to contemporary science for a method.

This movement by the mainstream churches may be seen as an attempt to evolve a new role and purpose in a secularizing world. In Mexican folk religion, too, *curanderismo* has adapted to address current issues and aspirations such as love, marriage, and happiness.

Some religious communities in which the practice of healing is central may stand or fall by their results. Others, however, lay hands on the sick because they are directed to do so (Mark 16, 18) and not with the primary purpose of a validating outcome. In an age of accountability, of the perception of an audience as a market, and of the importance of reputation and the fear of litigation, there has been a systematic movement to professionalize the alternative practice of healing. Practitioners have formed themselves into associations, established courses of study and qualifications, written ethical codes, and introduced a corporate discipline. These developments have been much more evident within complementary medicine outside the churches than within them.

3. *The Charisma of the Healer*

There is a tendency for the personalities of those involved in healing in sects and new religious movements to be rather higher in public profile than those who practice therapy in the suburbs and within the ministrations of the churches. Their names and faces are better known and are used in drawing crowds to religious meetings. The consequences of divine workings through them are more spectacular. Their disposition to secular medicine is more rejectionist and there is a triumphalism about their healing stories. When Melvin Banks testifies to the miracles God has worked through him, he shows the picture of a woman holding aloft her walking stick in defiance of the doctors who said she would always need it. Similarly, Her Holiness Mataji Nirmala Devi Srivastava, the 'Divine Mother' and teacher of Sahaja Yoga passes on to her disciples healing skills, including the use of vibration therapy in the cure of cancer. In the rhetoric, if not always in the understanding of the faithful, it is made explicit that the power is divine and the healer a mere agent. The claim to be a channel of divine power

is of course a greater enhancement of charisma than the claim to possess skills. The shaman possesses knowledge about healing which is not available to the community and which signifies an exclusive relationship with the spiritual domain. However, no man or woman would describe himself of herself as a shaman: the sense is that the power is invested in the shaman and an ethic of humility restrains self-promotion (Doore 1988). Of course, humility can itself be an element of charisma. Many *curanderos*, the folk-healers of Mexico, believe their healing powers to be a gift from God, invested in them and giving them certain responsibilities to alleviate misery. They claim no credit for themselves and, far from rejoicing in their powers, they may even regret the burden of their responsibilities: albeit they are full-time and charge fees.

But charisma invites suspicion from the secular world and from mainstream churches. It is in view of the capacity to exploit the charisma in the role of healer, both Church and the medical establishment distance themselves from alternative healing practices of new religious movements.

4. *Conflicts of Faith and Secular Practice*

It is reported (2 Chronicles 16, 12) that King Asa 'in the thirty and ninth year of his reign was diseased in his feet, until his disease was exceeding great: yet sought he not unto the Lord but the physicians: ... and in the one and fortieth year of his reign Asa slept with his fathers.' The Chronicler implies an error of judgement on Asa's part with a fatal consequence. Asa lacked faith. In modern times the resort to secular medicine frequently is perceived to compromise religious conviction.

The *curanderos* are condemned with suspicion by the fundamentalist churches which attribute to the devil the mysterious source of their power. For its part the medical establishment has invoked both research and the law to counter the rising tide of religious healing. The British Medical Association (1986) issued a report warning against the use of healing in new religious movements to confirm allegiance; it contended that they had caused harm to recruits and their families. The report was not well received, either by politicians or by the BMA's own members, but it survives as a significant expression at the high point of a conflict between the medical profession and alternative practitioners.

In the tradition of King Asa, Christian Scientists and the Peculiar People in England and others have at various times appeared in courts of law for refusing medical attention to the children. One of the more publicized conflicts between secular medicine and religious principle relates to the belief of Jehovah's Witnesses that blood transfusions are unscriptural and dangerous; in the United States, Canada, and

Australia medical decisions to transfuse children without parental consent—for example, by making children wards of court—have been more frequent than in the United Kingdom.

The rejection of secular medicine is, for some groups, axiomatic in the rejection of the world and its wisdom. For others, medicine does not need to be rejected with the world: the boundaries that exclude electricity and the motor car from Amish territory do not resist conventional medicine.

The conflict of religious and secular may be overcome by a strategy of incorporation: the religious organization may reckon to control not only its own ministrations but also those which are conducted outside it. So Christians in the church tradition may pray for those in hospital 'that by thy blessing upon them and upon those who minister to them they may be restored ...' while those in the Pentecostal church may ask, 'We pray thee, use that surgeon Lord to thy great glory.' And again, 'We prayed about the medicine and it began to work.'

5. *Conclusion*

The process by which religious organizations lose control of aspects of human life is familiar to students of religion as secularization. In recent years conventional medicine has suffered a similar attrition and the revival of religious healing has traded upon it.

There have emerged in the western world two movements that function as alternatives to conventional medicine and compare themselves with it. The religious tradition typified by but not exclusive to the practice of the Pentecostal churches has compelling parallels with shamanism from which it does not directly draw. The second tradition is that of complementary or alternative medicine which has burgeoned since the 1960s. The alternative is distinguished by the claim of its therapists to understand and control the process of cure while the religious involves the engagement of rituals in the expectation of a miraculous effect. There are several features common to both movements, including a suspicion if not rejection of conventional medicine and the role of physical contact between the healer and the healed, be it in the form of massage or manipulation or of the ritual laying on of hands. However, there are significant differences between the two movements. Complementary medicine has highly organized professional structures and the authority to practice comes from formal study and qualification. Clients of alternative medicine seek practitioners because of their methods, not for their charisma or personal reputation.

Such a typification highlights the peculiar case of the Church's Ministry of Healing which conforms much more to the secular alternative model than to the religious: its practitioners cultivate skills that have been developed in secular practice: it draws on the

Gospel for the spirit of its ministry but not for the letter of its technique.

See also: Religion and Health

Bibliography

- British Medical Association 1986 *Alternative Therapy*. BMA, London
- Chesnut R A 1997 *Born Again in Brazil*. Rutgers University Press, New Brunswick, NJ
- Doore G (ed.) 1988 *Shaman's Path: Healing, Personal Growth and Empowerment*. Shambhala, Boston
- Eliade M 1964 *Shamanism: Archaic Techniques of Ecstasy* (trans. Trask W R, Bollingen Series no. 76). Bollingen Foundation, New York
- Maddocks M 1981 *The Christian Healing Ministry*. SPCK, London
- Maguire M B 1988 *Ritual Healing in Suburban America*. Rutgers University Press, New Brunswick, NJ
- Martin D 1990 *Tongues of Fire*. Blackwell, Oxford, UK
- Pullar P 1988 *Spiritual and Lay Healing*. Penguin, London
- Trotter R T II, Chavira J A 1997 *Curanderismo: Mexican American Folk Healing*, 2nd edn. University of Georgia Press, Athens, GA

R. Homan

Health and Illness: Mental Representations in Different Cultures

Mental representations of health and illness are the cognitive structures that guide people's response to symptoms and illness as well as to various kinds of health threats. These mental representations have variously been described as 'cognitive schemata,' 'mental prototypes,' 'structured vocabularies,' and 'social representations.' However one describes them, they provide the foundation for responding to states of health and illness. These mental representations also vary from culture to culture and provide a basis for the wide variety of ways in which people in different cultures respond to health and illness related events. This article explores the nature of these mental representations, the ways in which they vary from culture to culture, and some of the implications of this variation.

1. Understanding Mental Representations of Health and Illness

Although the biological nature of disease remains constant from one culture to another, the evidence is

very clear that how diseases, and health and illness in general, are understood varies dramatically. This variation occurs from culture to culture as well as between individuals and subgroups within the same culture. As such, illness, which refers to the experiential component of disease and ill health, is best understood as a social and cultural construction. Whereas disease refers to objective pathology, illness, as experience, is constructed on the basis of shared understandings in addition to individual experience (Angel and Thoits 1987, Landrine and Klonoff 1992).

This construction of health and illness has been described in various ways. At one level it can be described in terms of structured vocabularies used by people to define and express feelings of distress. The basic idea is that culture and experience provide people with categories for defining experience as well as a language for expressing feelings of distress. For example, the common concept of influenza provides many people with a category for understanding the experience of fever, chills, headache, nasal congestion, sore throat, and general achiness, whereas the concept of *susto pasado* is a Mexican-American category used for understanding the experience of stomach trouble, diarrhea, lack of appetite, lethargy, irritability, and loss of weight after experiencing a strong fright.

A useful approach to understanding the shared nature of health and illness concepts is via the theory of social representations. The key idea in this approach is that in the process of attempting to understand their bodily experience, people create socially shared representations of health and illness that then become part of the heritage of the society as a whole or of a subgroup within the society. These representations influence the way that individuals think about health and illness but are a property of the larger social collective. This approach emphasizes the essential social nature of concepts of health and illness and the way in which they are shared between individuals, a key prerequisite for communicating private bodily experience to others. A good example of this approach is Herzlich's (1973) examination of illness beliefs in France.

A contrasting but complementary perspective is the information processing approach that examines the cognitions which individuals have about health and illness. This approach has its roots in cognitive psychology and, in particular, work on cognitive schemata and prototypes. Research using this approach has shown that people have implicit theories of disease that exert a strong influence on how they cope with health problems as well as how they deal with treatment. For example, Leventhal and Diefenbach (1991) have shown that the mental representations held by hypertension and cancer patients strongly influence their behavioral response to treatment. Research using this approach has also pointed out that such mental representations operate at different levels (sensory, affective, and conceptual) and have a number

of components. These components include the identity of the condition (the label it is given as well as the symptoms involved), its cause, time line (expected length and course), consequences, cure, and perceptions of the type of person who gets the condition. Taken together these components make up the prototype that a person has of a particular disease. For example, a common prototype of 'heart disease' among British young people includes symptoms of lack of stamina, breathlessness, and chest pains as well as the beliefs that it is caused by lifestyle and high blood pressure, is serious but can be cured by surgery, and mainly affects older people (Lalljee et al. 1993). These prototypes, in turn, affect the recognition of disease conditions (see *Symptom Awareness and Interpretation*) as well as the person's illness behavior (see *Illness Behavior and Care Seeking*). Research from the information processing perspective has also explored the cognitive organization of disease concepts, noting that these concepts are organized cognitively in characteristic ways that have implications for how people interpret illness threats and how they respond to people with particular diseases (Bishop 1991).

2. Cultural Differences in Mental Representations

One of the more striking features of health and illness concepts is the way in which they vary from culture to culture. This variation is obvious even to the casual observer, as concepts common in one culture or society may not even exist in another. Although there are any number of ways in which health and illness concepts may vary, three major ways stand out. First, cultures show substantial variation in conceptions of the basic nature of health and illness. For example, in Western societies placing a strong emphasis on the scientific understanding of health and illness, these concepts are understood in natural terms with disease considered as pathology deriving from natural causes, such as germs, injury, or genetic aberrations. By contrast, in other societies, as well as among some ethnic-cultural minorities within Western societies, illness is understood more in interpersonal as well as supernatural terms. For example, illness may be seen as resulting from violations of standards of interpersonal behavior or other social norms, particular emotions (such as jealousy or envy), transgressions of moral and/or religious taboos, or such quasinatural agents as hot-cold foods, drinks, weather, 'balance,' or particular characteristics of one's blood (e.g., thin, weak, or bad) (Landrine and Klonoff 1992). These differences in the general nature of health and illness are then reflected in the beliefs people hold about specific conditions.

Second, cultures often differ in their disease vocabularies. In some cases the differences may result from using different terminology for what appears to be the

same condition. For example, Kleinman (1980) notes that symptoms that are often diagnosed as depression in the West are interpreted as neurasthenia in China. In other cases, however, a condition defined in one culture may have no referent in another. For example, conditions such as *susto*, *empacho*, *caída de mollero*, and *mal de ojo*, experienced by some Mexican-Americans, have no direct referent among conditions known to Anglo-Americans or to medical professionals.

Third, the underlying structure of illness representations often differs between cultures. Work using techniques such as multidimensional scaling and factor analysis has shown that the mental representations people have of different diseases show characteristic organization in memory and that this organization differs between different cultures. For example, research with Anglo-Americans has found that disease representations appear to be organized along with dimensions of contagiousness and severity, whereas those of Mexican-Americans were organized according to concepts of 'hot' vs. 'cold' and whether a disease was associated with children or old people (D'andrade et al. 1972). As one might expect, the organization of disease concepts is closely related to the general models of health and illness that people are exposed to. Thus among Anglo-Americans whose folk model of health derives primarily from the biomedical model, contagiousness, and severity, central concepts in the biomedical model, have been found to be prime organizing dimensions. However, among Chinese in Singapore where Chinese traditional medicine exists side-by-side with Western scientific medicine, disease concepts are organized according to dimensions deriving from both Chinese and Western medicine (Bishop 1998).

3. Implications

The mental representations that people have of health and illness, and cultural differences in these representations, have a number of implications for health and illness related behavior. First, the disease concepts that people have can be expected to have a strong influence on their interpretation of symptoms and bodily experience in general. In brief, having specific disease concepts makes it possible, and perhaps likely, that a person will interpret bodily experience in line with that concept. Thus, having available the concept of *susto* makes it possible for Mexican-Americans to experience this condition, an experience that is not available to others not possessing this concept. Along these same lines, the relative currency of the concept of neurasthenia in China and of the concept of depression in North America increases the likelihood that the similar symptoms associated with these two conditions

will be interpreted, and perhaps experienced, differently in these two cultures.

Second, and following from this, health and illness representations have implications for what people do when they experience distress. For one thing, these representations provide a definition for when bodily distress represents illness and requires assistance from a medical professional or other healer. For example, research with heart attack patients points out that one reason for delay in seeking treatment is that the person misinterpreted the symptoms and failed to recognize the seriousness of the situation. In addition, disease representations serve to direct the person's help-seeking behavior. For example, research in Singapore has noted that degree of adherence to beliefs associated with Chinese traditional medicine is a key determinant of whether the person seeks help from a practitioner of Western or Chinese traditional medicine (Bishop 1998).

Third, once help is sought, disease representations play an important role in determining whether the person stays in treatment and follows the treatment advice. A good example of this is research with hypertension patients showing that the cognitive 'model' they had of hypertension and its relationship to specific symptoms was a major determinant of whether they stayed in treatment (Leventhal and Diefenbach 1991).

Finally, illness representations play a key role in how people respond to potential health threats. Specifically, the beliefs that people have about particular diseases, how they are contracted, and their consequences have a strong influence on how they respond to people having the condition as well as what precautions, if any, they take to prevent the disease. For instance, beliefs among some Africans that anyone who has ever had sex with a person who has died of AIDS will automatically die as well appear to lead to refusal on the part of those holding such beliefs to seek medical advice, be tested for HIV, or change their sexual behavior so as to reduce their risk of infection (Ankrah 1991).

4. Remaining Questions

In many respects work to date has only scratched the surface of this fascinating topic. As noted, progress has been made in the theoretical understanding of mental representations of health and illness but much more can be done to flesh out this understanding and, in particular, explicitly account for cultural differences. Also much remains to be done in further exploring the content of health and illness concepts in different cultures and relating these to other aspects of the cultures in question as well as to dimensions, such as individualism-collectivism and power distance, used to describe cultures.

See also: African Studies: Health; Cultural Psychology; Culture as a Determinant of Mental Health; Illness Behavior and Care Seeking; Latin American Studies: Health; Psychosomatic Medicine; South Asian Studies: Health; Southeast Asian Studies: Health; Western European Studies: Health

Bibliography

- Angel R, Thoits P 1987 The impact of culture on the cognitive structure of illness. *Culture Medicine and Psychiatry* **11**: 465–94
- Ankrah E M 1991 AIDS and the social side of health. *Social Science Medicine* **32**: 967–80
- Bishop G D 1991 Understanding the understanding of illness: Lay disease representations. In: Skelton J A, Croyle R T (eds.) *Mental Representation in Health and Illness*. Springer-Verlag, New York
- Bishop G D 1998 East meets West: Illness cognition and behaviour in Singapore. *Applied Psychology* **47**: 519–34
- D'andrade R G, Quinn N R, Nerlove S B, Romney A K 1972 Categories of disease in American-English and Mexican-American Spanish. In: Shepard R N, Romney A K, Nerlove S B (eds.) *Multidimensional Scaling*. Seminar Press, New York
- Herzlich C 1973 *Health and Illness: A Social Psychological Analysis*. Academic Press, London
- Kleinman A 1980 *Patients and Healers in the Context of Culture*. University of California Press, Berkeley, CA
- Lalljee M, Lamb R, Carnibella G 1993 Lay prototypes of illness: Their content and use. *Psychology and Health* **8**: 33–49
- Landrine H, Klonoff E A 1992 Culture and health-related schemas: A review and proposal for interdisciplinary integration. *Health Psychology* **11**: 267–76
- Leventhal H, Diefenbach M 1991 The active side of illness cognition. In: Skelton J A, Croyle R T (eds.) *Mental Representation in Health and Illness*. Springer-Verlag, New York

G. D. Bishop

Health and Mortality, Body Stature as a Measure of

In recent years a rapidly expanding research on the determinants and consequences of human stature has become central to the study of health and mortality processes. Human stature—as measured by height, weight, body mass (which combines height and weight), and a host of other anthropometric measures—has been shown to reflect health status and mortality risk across the life course. Stature is an indicator of health *in utero*, in infancy, and in childhood, and is an important predictor of chronic disease and mortality among adult populations. Moreover, anthropometric measures can be used to com-

pare health and nutritional status across populations, social classes, and historical time periods.

1. Stature as an Indicator of Nutrition, Health, and Mortality

1.1 The Importance of Height

Although auxology (the study of human growth) began as early as the seventeenth century, it was not until the nineteenth century that scholars began realizing that environment can systematically affect human growth. For instance, Villerme argued, based on a study of French soldiers, that poverty was much more important than climate in influencing growth. A few decades later, Charles Roberts estimated fitness for factory employment using frequency distributions of key anthropometric measures such as weight-for-height and chest circumference. In the early 1890s, Franz Boas conducted a national growth study, which was used to develop national standards in the USA. In later years, Boas was a pioneer in the use of statistical methods to analyze growth and furthered the work of linking environment and heredity to growth (Steckel 1995).

In the twentieth century, it became clear that a child's growth rate is likely the best single indicator of his or her health and nutritional status. Furthermore, mean values of childhood heights and weights provide powerful summary statistics of the public health of the population, and childhood anthropometric measures can be used to make meaningful comparisons of health and nutritional status across populations (Eveleth and Tanner 1990).

Adult height can also be used as a measure of public health, since adult height is a proxy for the childhood nutritional status of a population. However, adult height reflects a complex interaction of genetics, environment and the timing of nutritional shortfalls. If nutritional deprivation occurs during the 'sensitive periods' of development, differences in adult height may be quite striking. The sensitivity of growth to nutritional deprivation depends upon the age at which it occurs. Catching up is possible, but prolonged and severe deprivation results in stunting or a reduction in adult size.

Height is a significant predictor of older age mortality. Overall, mortality risk declines as height increases. One of the most comprehensive modern studies was completed by Waaler (1984) on the entire Norwegian population. He found that men measuring 165cm in height face a mortality risk that is 71 percent greater than those measuring 182.5cm. Obviously height is also strongly influenced by genetic factors, but genetics cannot explain the significant differentials in height that exist today according to socioeconomic status, nor the narrowing of differentials in industrialized nations (Floud et al. 1990).

1.2 Weight and BMI

In both children and adults, weight is an important anthropometrical measure. Whereas height is typically used as a proxy for past nutritional development, weight is seen as a measure of current health and nutritional status. In the developing world, low weight remains a significant indicator of nutrition and health. In the developed nations, high weights are the principal concern. Obesity rates, especially in the USA, have risen over recent decades and obesity is frequently found to be a risk factor for serious chronic illnesses, including cardiovascular disease, a leading cause of adult mortality in developed nations. Obesity has also shown to be correlated with race and social class.

Obviously weight can be more effectively exploited in epidemiological analysis if it is standardized for height. A common standardization is to use body mass index (BMI), which is defined as height (in meters) divided by the square of weight (in kilograms). The mortality risk is relatively constant in the range of 21–28. However, as BMI falls below 21, adult mortality rises sharply. Mortality also rises for BMI higher than 28. BMI is associated with higher mortality from all causes, though the relative importance of BMI may be confounded by a history of smoking and by previous diseases, and the impact of BMI declines with age. Some studies suggest that much of the increased mortality risk from low BMIs in the USA can be attributed to smoking. Interestingly, excess mortality among African-American women in the USA is not attributed to BMI, even though they have high rates of obesity.

2. Determinants of Stature

Over the last century, scholars have made considerable strides in characterizing and understanding human growth and how various genetic, nutritional, and environmental factors interact with the growth process. Clearly the most important time is infancy, in which the growth rate is greater than at any point thereafter. Not coincidentally, infancy is also the time when humans face the greatest risk of mortality. After infancy, the growth rate declines irregularly until adolescence, during which it rises, at its peak, to about one-half the rate during infancy. The most important study on human growth is by Eveleth and Tanner (1990), and the following sections draw heavily from this work.

2.1 Genetics

Within a community facing similar environmental influences, the distribution of heights is dominated by genetic influences. Systematic studies of heights within families confirm this everyday observation. Studies of twins and other sibling relationships provide the most

convincing evidence. Wilson (1976) reports that monozygotic (MZ) twins, who share the same genes and, in most cases, a common family environment, have, on average, only 1.1 cm difference in height at age four, while dizygotic (DZ) or fraternal twins have a 3.2 cm difference at the same age. Furthermore, differences in height in MZ twins decline as the twins get older, but the difference increases for DZ pairs.

Numerous studies of twins and other siblings provide strong evidence that the closer the genetic relationship, the closer the growth patterns. MZ twins are more closely intercorrelated than DZ twins, but siblings in general are much more similar to each other than they are to nonrelated individuals. Furthermore, the heights of parents and children are highly correlated, though no evidence exists that one parent is more important than the other.

2.2 Nutrition

Even though genetic influences are an important determinant of *individual* variation in height, differences in average height across *populations* are almost entirely the product of environmental influences, not genetics. Although many environmental factors affect body stature, chief among them is nutrition. Of particular importance is nutrition in early life. Numerous scientific studies have shown that severely malnourished children have a greatly increased mortality risk relative to normal children (Behrman and Deolalikar 1988).

Infancy is the period of life in which both the human growth rate and the risk of mortality is at its highest. It is not surprising, therefore, that it is also the time during which nutritional intake is the most critical. Indeed, the concern for nutrition may begin *in utero*, and several studies have shown that supplementing the diet of pregnant women in developing nations significantly increases birth weights—the most important predictor of infant mortality. Even in developing nations, infant mortality is higher than any other period of life. By the age of two, much of the deficit characteristic of adults has already accumulated. Nutrition during the adolescent growth spurt may be important, though the evidence is mixed. In any case, the impact on morbidity and mortality of adolescent malnutrition is nowhere near as dramatic as it is in early childhood.

Many sub-populations in developing nations have mean birth weights that are similar to those in developed nations. In these developing nations, however, weight gain begins to diminish at about six months, the time when mothers' lactating ability declines and children are often weaned. Although food supplements can help, the prevention and treatment of infectious diseases is also critical in this period in developing nations.

2.3 Infectious Disease

Infectious disease interacts closely with nutrition in the growth process. Even in nutrition-rich environments, the undeveloped immune systems of children make them more susceptible to disease. When diet is poor, most of the body's resources are devoted to basic maintenance; malnourished children, therefore, are even more susceptible to disease. Furthermore, infection reduces the body's ability to absorb nutrients, which can make children even more malnourished. Infectious disease in childhood, therefore, can have long-term consequences on growth. Some controversy does exist on this topic, however. An alternative 'small but healthy' hypothesis has been promoted by a few researchers, though it has not received widespread support.

2.4 Socioeconomic Correlates

Economic growth (which results in better nutrition, public sanitation, improved hygiene, improved medical treatment, etc.) has been the driving force behind the secular increases in height around the globe. Variation in national income can also explain much of the variation in mean heights across nations. Steckel (1995) calculates correlation coefficients between mean height and the log of per capita income that are between 0.8 and 0.9. National income remains a powerful correlate of height, even after controlling for other demographic characteristics of the nations such as urbanization and age. And in nearly all nations, both developed and developing, children with moderate to high socioeconomic status are larger than their counterparts with low socioeconomic status, where socioeconomic status is determined by education, occupation, and income.

Of course the effect of socioeconomic status is interwoven with nutritional intake and infectious disease environments. Consequently, socioeconomic differentials are much more pronounced in developing nations. Socioeconomic differentials in height are most pronounced in childhood, but exist for adults as well. Several studies show that the social class effects are greater for boys than for girls.

3. Stature and the Study of Health and Mortality

3.1 Historical Populations

In the last decades of the twentieth century scholars began using body stature, and height in particular, to determine the living standards and health status of populations for several historical time periods and in various locations around the globe. Health and mortality details, particularly at the individual level, are generally unavailable in historical data, but height data are available (typically from military service records) in several American and European popu-

lations since the seventeenth century. Height, therefore, is the most important indicator of health status we have when investigating the historical record.

These studies have confirmed traditional notions of public health, such as the abysmal living conditions in eighteenth- and nineteenth-century urban areas. There have also been some unexpected findings. For example, the USA, in contrast to European nations, achieved near-modern average height as early as the mid-eighteenth century, though heights declined somewhat over the nineteenth century (from 173 cm in 1800 to 171 in 1900 for adult men), mostly during the middle part of the century. Mortality evidence from genealogies and from plantation records indicates that life expectancy deteriorated while heights declined during the ante-bellum period. The cause of the nineteenth-century decline in height is still a puzzle. Hypotheses include increased urbanization, nutritional decline, an increase in infectious diseases created by trade, migration, and public schooling, and the disruption of the US Civil War.

Before the twentieth century, European men were smaller than Americans, probably due to far poorer diets, but there was significant diversity in height across European nations. In the early nineteenth century, residents of Ireland and Scotland were the tallest (168 cm) followed by Norway, Sweden, England, France and Austria-Hungary (163 cm). In the twentieth century, geographic distributions of height shifted. Based on data collated by Eveleth and Tanner (1990), the Dutch and Norwegians were now the tallest, followed by Americans, British, French and Austrians. Socioeconomic differentials were pronounced in Europe. Floud et al. (1990) report, for instance, that in nineteenth-century England, 14-year-old Sandhurst boys (mostly the children of aristocrats, high-ranking officers, and professionals) exceeded the stature of boys from the Marine Society (mostly from the slums of London) by 10 to 15 cm.

3.2 The Modern World

The period from the late-nineteenth to the mid-twentieth century was characterized by dramatic increases in stature and equally dramatic reductions in mortality. Life expectancy was about 40 years in 1860 but had risen to 68 years by 1950 (see *Mortality: The Great Historical Decline*). There is no single cause for this dramatic increase, but several contributing factors have been identified. Widespread public health programs were implemented, such as vaccinations, purifications of water supplies, sewage disposal, and milk pasteurization. Medical advances, such as an understanding of the germ theory of disease and the development of antibiotics, as well as increasing access to medical care, also contributed. Increases in income also facilitated better diets and housing.

Stature remains an important variable in modern studies of health and mortality, though its use varies

by the economic development of the population under study. In developing nations, stunting and wasting are still critical indicators of nutritional status and the overall health of the population. In developed nations, most of the focus is on increased mortality of the obese, as indicated by high levels of BMI. Significant differentials in all health indicators, including anthropometric measures, still exist across cultures and social classes.

Central to the improvements in the twentieth century was the fight against infectious diseases. In a wide-ranging study of numerous populations from 1900 to the early 1960s, Preston (1976) attributed 25 percent of the mortality decline to influenza, pneumonia, and bronchitis; 15 percent to 'other infectious and parasitic diseases'; 10 percent to respiratory tuberculosis; and 10 percent to diarrheal diseases (the remaining portion of the decline is most likely attributable to cardiovascular disease). As noted earlier, the impact of infectious disease, particularly among children, is impacted significantly by nutrition.

Many factors, such as the development of antibiotics, contributed in the twentieth century to the decline in infectious disease prevalence and in the associated reduction in mortality. During the period between 1950 and 1970, life expectancy in developed nations increased only slightly (see *Life Expectancy and Adult Mortality in Industrialized Countries*), which many attributed to the successful fight against infectious diseases. Some argued that because so much early mortality was due to infectious disease, modern populations were facing an era of increased life but worsening health. People with frailer constitutions were living to older ages and, it was thought, would face higher rates of chronic illness than previous cohorts of elderly (Verbrugge 1984). However, most recent evidence suggests that age-specific morbidity and disability rates have been declining in the elderly population. Also surprising to most demographers has been the sharp increase in life expectancy after the period of stagnation between 1950 and 1970. The sources behind this increase remain largely a puzzle, and little consensus exists on future projections of life expectancy.

Towards the end of the twentieth century, birth weight, as well as other infant anthropometric measures, was linked not only to infant and child mortality, but also to disease and mortality in later life. The work of Barker (1994) suggests that development *in utero* may be critical to long-term health. This and other research suggests that anthropometric measures will remain an important component in an understanding of the complex and changing relationship between nutrition, health and mortality.

See also: Genetics and Development; Longevity, Genetics of; Mortality, Biodemography of; Mortality Differentials: Selection and Causation; Mortality, Epidemiological, and Health Transitions

Bibliography

- Barker D J P 1994 *Mothers, Babies and Disease in Later Life*. BMT Publishers, London
- Behrman J R, Deolalikar A B 1988 Health and nutrition. In: Chenery H, Srinivasan T N (eds.) *Handbook of Developmental Economics*. North Holland, Amsterdam
- Eveleth P B, Tanner J M 1990 *Worldwide Variation in Human Growth*, 2nd edn. Cambridge University Press, Cambridge, UK
- Floud R, Wachter K, Gregory A 1990 *Height, Health and History: Nutritional Status in the United Kingdom, 1750–1890*. Cambridge University Press, Cambridge, UK
- Fogel R W 1986 Nutrition and the decline in mortality since 1700: Some preliminary findings. In: Engerman S L, Gallman R E (eds.) *Long-term Factors in American Economic Growth*. University of Chicago Press, Chicago
- Preston S H 1976 *Mortality Patterns in National Populations*. Academic Press, New York
- Rogers R G, Hummer R A, Nam C B 2000 *Living and Dying in the USA: Behavior, Health and Social Differentials of Adult Mortality*. Academic Press, San Diego, CA
- Steckel R H 1995 Stature and the standard of living. *Journal of Economic Literature* 33: 1903–40
- Verbrugge L M 1984 Longer life but worsening health?: Trends in health and mortality of middle-aged and older persons. *Milbank Quarterly* 62: 475–519
- Wilson R S 1976 Concordance in human growth for monozygotic and dizygotic twins. *Annals of Human Biology* 3: 1–10
- Waalder H T 1984 Height, weight and mortality: The Norwegian experience. *Acta Medica Scandinavica Suppl.* 679: 1–51

S. E. Wilson

Health: Anthropological Aspects

Health and health practices are part of the inmost complexities of social existence, permeating the domains of politics, economics, and religion and almost always connected with dimensions that go beyond the body, such as interpersonal, family, and community relationships. Health is at once an end and a means. Explanations of the causes and patterns of health and disease often convey value judgments, senses of right and wrong, and of accountability and blame, as well as revealing what is morally at stake in definitions of health and its failures. Medical anthropological studies have built upon historical and cultural analyses over a long period of time that indicate an array of difference with respect to the metaphors and meanings that signify health. These studies have emphasized cultural fabrics that give coherence and depth to these meanings such that cosmology and ethical traditions come to define the body in states of sickness and well-being.

Anthropological studies of health focus on local contexts where health and illness are recognized and responded to. Such studies also trace the effects of global flows of commodities, information, finance,

images, people, and pathogens on such worlds. Anthropologists have been as interested in the social roots and consequences of health (and illness) as in their cultural representations. But what most particularly characterizes the anthropological perspective is the use of ethnography to understand health, illness, and health care. Anthropological perspectives on health bring together individual and collective realities in the way they are organized, narrated, contested, and in every sense lived as social trajectories (Kleinman 1995).

What anthropologists have learned through research in different societies is a story of diversity in experiences of health and health care giving, as well as of the ways social relationships and emotional processes interfuse in the mediation of health-relevant experiences. The meanings of health range from holistic and organic interpretations to hyperindividualistic ones. While biomedicine has treated health as a separate domain, popular meanings and interpretations of healthy states are often inseparable from other moral, political, and economic domains. Anthropologists have been particularly concerned with the transformations in health as those transformations express differences in power, social position, and social inequality, particularly as experienced by marginal groups and individuals. But in recent years, equal interest has been devoted to studies of health professionals, bureaucrats, and scientists who contribute to the public definition of health and health problems. Culture affects health as much through the culture of biomedicine and bureaucratic institutions as through the popular culture; indeed in the age of the Internet the two are closely related.

Applied medical anthropology addresses issues such as the ones raised and addresses them at the policy level. In heterogeneous societies such as North America and Western Europe, cultural differences continue to matter in the health field in large part because of differences in health conceptions which affect the panoply of health-relevant practices, from popular assessments of risks and responses to them to the uses and misuses of health care resources. Those differences contribute to different health outcomes for people, but they also contribute to different health systems, which are themselves undergoing transformation owing to both local and global economic and social changes.

1. Images of Health and Diverse Experiences of Health

The contemporary world contains impressive variations of images of health, modes of health maintenance, efforts at prevention, and illness experiences. Much of this variation is subsumed under what demographers and epidemiologists have termed the

'health transition.' This phrase refers to an increase in life expectancy, decline in mortality, and overall decrease in infectious diseases in the twentieth century by means of interventions such as vaccinations and dietary improvements. Health status measures have been introduced and are continually revised to develop policy, planning, and evaluation; these measures shape ideas of health and health improvement according to specific indices that promote interventions and socio-behavioral change. The health transition has also been accompanied by major changes in patterns of morbidity, increases in chronic degenerative diseases, old and new epidemics and behavioral and mental pathologies and is thus more complex than its original conceptualization (Chen et al. 1994). The very economic, technical, and industrial development that has benefited so many coexists with ongoing massive societal disruptions. Developing countries, for example, are experiencing the social and health costs of rapid urbanization and economic restructuring, costs that are confirmed in statistical increases of preventable disease and illness, and borne in terms of health problems such as infant mortality, life-threatening dehydration, urban violence, and hunger (Scheper-Hughes 1992).

Such acts as war, political persecution, and homicide contribute to a deliberate production of suffering and death. The era of infectious diseases is by no means past as the spread of HIV/AIDS, multidrug resistant tuberculosis, and other infectious diseases clarify. With the collapse of authoritarian rule came the collapse of public health measures and treatment regimes, demographic changes, travel, and other human actions that intensify the flows and resistance of pathogens. It is through such mechanisms that Paul Farmer has shown the ways social forces alter disease distribution and contribute to the persistence of new microbial conditions; indeed, the way 'social forces and processes come to be embodied as biological events' (1999, p. 5). The appearance of these disorders, closely correlated with poverty and social inequality, challenges the notion of a single biological etiology for disease and illustrates that these infectious diseases are mediated by multiple interacting causes that require anthropological inquiry and policy intervention.

The burden of chronic diseases—heart disease, cancer, stroke, diabetes, asthma—continue to be distributed unevenly across minority groups. Yet statistics also disclose impressive variations. In the United States, for example, there is the so-called Latino Epidemiological Paradox: Latinos, who generally experience a low socioeconomic status, nonetheless, have health outcomes that are either equal or better than non-Hispanic Whites. African Americans, by comparison, experience significantly higher rates of infant mortality, death, diabetes, HIV/AIDS, heart attacks, strokes, and homicides. Differences in levels of institutionalized racism, dietary patterns, and the influence of the Spanish language as a protective

cultural barrier have been pointed to as possible explanations for this paradox. Notably, health indices of Latino immigrants worsen from generation to generation in the United States, owing perhaps to a loss of balanced diet or tobacco or substance abuse as well as limited accessibility of health care services (Hernandez and Charney 1998).

The latest cluster of disorders to receive attention in the health transition is the rising rates of substance abuse, violence, psychological trauma, depression, suicide, and related mental health problems (Desjarlais et al. 1995). This group of health problems has increased to epidemic proportions both in developed and developing societies. Anthropological studies indicate that these problems often cluster together in settings of deep poverty, and societal breakdown. They are compounded by the fact that services tend to be limited, and health care financing and resources are inadequate. Many psychiatric hospitals in China, Russia, and Latin America continue to rely on a model of care involving prolonged stays in locked wards with antipsychotic medication and physical treatment (Phillips 1998).

Anthropological interventions have gone beyond explication of the 'soft' variables underpinning mental health problems, and focus on the ways social, economic, and political contexts shape the formation, distribution, and manifestation of disorders. Thus in the anthropological perspective, health problems and social problems are inseparable; health policies and social policies are viewed as necessary complements in defining criteria and techniques of successful rehabilitation.

2. Local Variations in Illness and Health

In response to emerging sets of health issues, anthropologists have identified the need for cross-cultural studies of health, healing, and patterns of morbidity in local contexts. One thread linking these studies is the importance of understanding patterns of everyday life experience, local knowledge, and social networks that influence personal agency and access to health care, as well as modes of transmission of diseases. From this perspective, health is seen as not simply originating in individuals themselves or deriving strictly from social forces, but as being engendered in complex interactions between individuals and their social contexts. Such lived responses may occur in the contexts of medical pluralism where several therapeutic systems coexist in one cultural setting, or in circumstances in which symptoms are subject to several distinct classifications, diagnoses, and hence, therapeutic courses of action or, once again, where different moral and religious values shape health promoting and disease preventing activities related to diet and lifestyle practices and support (Leslie and Young 1992). By stressing contingencies and detailing local histories of

particular places and peoples undergoing many kinds of social change, these ethnographies highlight the way particular worlds define health and conceive of health equity, while also illuminating the ways in which risks to disease and illness inter-relate.

It is not surprising that ethnographers have focused on illness whose event structure is relatively defined. Emphasized is the experience of illness as an interpersonal phenomenon as it is narrated through stories of illness. Ethnographies disclose how suffering is transformed through language and cultural meanings into distinctive illness narratives whose shared metaphors, voices, and genres represent distinctive experiences and political processes. This transformation, in turn, shapes the lived experience of illness through embodiment of cultural meanings and social relations into body-self processes. Ethnographers have focused on different mediating processes—ritual, devotional practices, music, poetics—that structure a sociosomatic continuum between moral and emotional processes. This approach relates to the concept of ‘local biologies’ by which is meant the different means of relating social worlds and biological processes such that those processes become particularized. Examples include Margaret Lock’s demonstration of fundamental differences in the experiences of menopausal symptoms and the aging process among women in Japan and North America (1993); Lawrence Cohen’s contrast of distinctive interpersonal experiences of dementia in India and the West (1998); and Rayna Rapp’s description of how experiences of class in the United States and technologies of the female reproductive life cycle come to alter experiences of pregnancy, infertility, abortion, and delivery (1999).

Ethnographers have also analyzed the trajectory and outcome of illness experiences on the social course of illness. Thus, Benedicte Ingstad and Susan Reynolds Whyte (1995) have shown how disability is as much a product of social definitions and restrictions and opportunities of access as it is a physical or functional limitation or disease, thus turning disability into an experience of potential wholeness and health. Other anthropologists have investigated societal responses to highly stigmatized disorders that respond to feelings of threat with practices aimed at marginalizing sufferers. Still others have studied experiences of chronic conditions like chronic fatigue syndrome, chronic pain syndrome, and fibromyalgia—disorders that are in a limbo of partial or minimal authorization where patients feel discredited or discreditable because of the lack of social legitimacy. Research here shows how patients often resist biomedical responses that are perceived as undermining and delegitimizing.

More recently, this area of research has expanded well beyond biomedically defined conditions to include a variety of forms of violence—structural, political, domestic—that represent significant threats to health. Here researchers of violence such as Veena Das (1995) have shown how different societal idioms

of those institutions authorized to deal with social problems—legal, welfare, religious, and of course medical—construct distinctive objects of programs and policies that not only differ greatly but that coerce different courses of health and disease and outcomes of these conditions.

Perhaps the largest field of anthropological inquiry has examined the varieties of community healing systems that parallel state institutions. Studies of herbalists, shamans, and other folk practitioners (and practices) illustrate how experience is manipulated, how meaning is negotiated, and how illness comes to serve some societal ends such as social control, cultural remembering, and political resistance. The point being that the collective significance of illnesses goes beyond individual and even familial concerns and plays a role in how societies remember, how social reality is legitimated, and how social worlds change.

The anthropology of health and medicine largely began as an applied field. There continues to be important segments of the field that seek application in public health and biomedical settings and in the policy arena. Such applied work classically demonstrated that health beliefs could either facilitate or become a substantial barrier to implementing policies and progress (Paul 1955). More recently, anthropologists have engaged the nongovernmental, administrative, and political dimensions of international health agencies and the policy development process, indicating how community organization can positively inform health and social programs in arenas such as HIV/AIDS prevention, cancer or infectious disease control, and family planning (Hahn 1999). Hence, anthropology has moved from a stance of critique of application to a stance of generating novel applications combining anthropological knowledge and public health technologies to enhance health.

3. Culture of Biomedicine

Distinctions between health and disease or illness are patterned according to norms. Norms of health and disease often emerge through a combination of statistics patterns, public consensus around desirable health outcomes, and social definitions and responses to deviance (Mechanic 1997). The extent to which biomedical norms have achieved a position of preeminence in health care systems and definitions of disease is undeniable. Biomedicine’s success in eradicating certain diseases and prolonging life shows that norms, however value-laden and socially constructed they may be, can indeed be health-enhancing (DeVecchio Good 1995). Yet anthropological studies have questioned biomedicine’s reductionism, primary value of efficiency, and authority in mediating interpretations of health, illness, and suffering (Good 1994, Kleinman 1995); they point to the fact that physiological measures alone fail to capture the

subjective dimensions of health. The growth of biomedicine, its characteristic quality of redefining social problems as health problems needing biomedical intervention, and its global cultural reach have also been analyzed. That biomedicine has been closely allied with the growth of industrial capitalism has led many ethnographers to examine biomedicine as a conduit of social regulations and discipline influencing all sectors of life, from definitions of disability to norms of occupational health, to definitions of mental illness and fitness (Turner 1995). This process has led to strong critiques regarding the necessity and indeed at times harmfulness of biomedical interventions, particularly with respect to the unpredictable side effects of technical procedures and drugs used to diagnose and treat patients. Concerns for iatrogenesis are reflected in current US debates regarding the consumption of psychiatric drugs by preschool children for example.

Attention to the international political economic context in which biomedicine operates is central to understanding how ideas of health modify, as well as to understanding the modes through which they are integrated as a set of values and practices in local contexts. Cross-cultural studies reveal the extent to which this integration with local experience is contested or resisted by highlighting: (a) the many discrepancies between biomedical knowledge and subjective accounts of illness and health, particularly in non-Western countries; (b) the moral and religious tensions underlying choices of medical systems; and (c) the degree to which distinctions between health and disease are rooted (or uprooted) by global ideologies and beliefs.

4. Biotechnology and Health

In response to the large influence that public health has achieved in our times due to the advances of biomedicine and advances in technology, the power of state action in the public domain, and market forces, health has increasingly become a specialized domain. The history of public health has been marked by major advances in population-level health in the nineteenth and early twentieth centuries through sanitary and workplace improvements, housing, and nutrition. Public health has shifted priority from social medicine to individually targeted health promotion. The launching and completion of the largest biology project ever conceived—the mapping and sequencing of the human genome—has already accelerated this process of individualizing health vis-à-vis the genetic profiling of diseases and increased prognostic information. For example, silicon chips embedded with hybrid bits of DNA that can test for various genetic conditions in individuals are currently in development. Such diagnostic tools are refocusing medical practice away from intervention to risk factor analyses and disease prevention.

The diseases individuals might be predisposed to, how they might react to particular medicines, and what health-promoting behaviors they might take to prevent future illness are all encompassed in this revolution (Rabinow 1996, Eisenburg 1999). Genomic information and the multitude of genomic sequences represent novel targets for therapeutic intervention. In short, the ‘life sciences industry’—knowledge of the genetic code and how to manipulate it and therapeutic advances—is transforming public health and biological knowledge. The prospect of widespread genetic testing has prompted concerns about how such information is being used to discriminate against those labeled as having genetic risks in terms of restricting or denying insurance coverage, as is already the case with women who are BRCA1 or BRCA2 positive.

These developments also highlight the depth of demand for ‘health.’ They attest to the power of biomedicine as a rationality in everyday life (in the organization of health care and delivery, legal procedures, policy analysis) that comes up against other forms of knowledge such as religion, and that coexists with older power structures related to family, gender, ethnicity, and class. Health as commodity, to be bought and sold in the marketplace, has led to new freedoms to pursue health but also raise dilemmas and create new risks: who is caretaker of genomic information? Who will have access to genetic therapies? These questions highlight some of the dangers regarding the social uses of genetic knowledge in the genomic and postgenomic era; they make it difficult to say what counts as success in this drive toward health and profits, a drive which has both sacred and secular sources and consequences.

The tradeoff between benefits to individuals and risks to populations of any given new technology must also be carefully assessed. Xenotransplantation, which promises to end the so-called shortage in donor organs by engineering transplantable animal organs, offers one prominent example. While such breakthrough technologies may save the life of a few, they pose dangers to populations in terms of the resistant viral strains and antibodies that might develop as a result of such transplant procedures. Another example of the way scientific advancement produces new social risks is found in the case of the traffic in human organs. Such traffic has been afforded by the availability of cyclosporine, among other things. This availability, combined with demand for human organs, has turned vulnerable populations in Brazil, South Africa, and India, into potential organ suppliers. Inconsistencies across the globe regarding the rights of donors as well as the lack of follow-up care for many of them have led to prolonged suffering or premature deaths, or both. Processes related to the globalization of the life sciences industry demand ethnographic attention to technical advances in terms of their development and diffusion; patterns of accessibility and differences in improved health outcomes between the poor and the

nonpoor; and the new forms of health and vulnerability these advances engender.

Images of health, illness, and healing have become key media, a 'fragile salvation,' through which societies and groups struggle with central values and beliefs, and negotiate social and political progress. The invention and implementation of techniques of health improvement have taken on many forms, including epidemic control, disease eradication, and combinations of environmental and individual behavioral change and medical advance. Yet definitions of health depend on normative and oftentimes conflictual values. Health as an individual goal coexists with new threats to health worldwide. The spread of advertising for cigarettes or the dumping of expired pharmaceuticals in low-income countries are all part of the globalization of disease. Under such conditions, what was once health now becomes disease. Health maintenance and promotion, once so influential a concept, can be undermined owing to new political and economic realities.

Contexts of health are themselves undergoing changes and are influenced by transformations in economy and polity, in science and technology, and in culture. These contexts shape the conditions of health and are in turn affected by it. These processes demonstrate that health as an experience is irreducible to a set of norms of physiological and mental activity, or strictly to a set of cultural differences. Only through concrete understandings of particular worlds of suffering and the way they are shaped by local histories and political economies can we possibly come to terms with the complex human dimensions that protect or undermine health. Health, seen in this way, is a construction as well as a contested way of being in the world.

See also: Community Health; Cultural Variations in Interpersonal Relationships; Culture and Emotion; Development: Sustainable Agriculture; Ecology and Health; Emotions and Health; Environment and Anthropology; Food in Anthropology; Globalization and Health; Healing; Health Care Systems: Comparative; Health Psychology; Medical Geography; Medical Sociology; Postmodernism: Philosophical Aspects

Bibliography

- Chen L, Kleinman A, Ware N 1994 *Health and Social Change in International Perspective*. Harvard University Press, Cambridge, MA
- Cohen L 1998 *No Aging in India: Alzheimer's, the Bad Family, and Other Modern Things*. University of California Press, Berkeley, CA
- Das V 1995 *Critical Events: An Anthropological Perspective on Contemporary India*. Oxford University Press, Delhi, India and New York

- DelVecchio Good M-J 1995 *American Medicine: The Quest for Competence*. University of California Press, Berkeley, CA
- Desjarlais R, Eisenberg L, Good B, Kleinman A 1995 *World Mental Health: Problems, and Priorities in Low-Income Countries*. Oxford University Press, New York
- Eisenberg L 1999 *The Implications of Genetics for Health Professional Education*. Josiah Macy, Jr Foundation, New York
- Farmer P 1999 *Infections and Inequalities: The Modern Plagues*. University of California Press, Berkeley, CA
- Good B 1994 *Medicine, Rationality, and Experience: An Anthropological Perspective*. Cambridge University Press, Cambridge, UK and New York
- Hahn R 1999 *Anthropology in Public Health: Building Differences in Culture and Society*. Oxford University Press, New York
- Hernandez D, Charney E 1998 *From Generation to Generation: The Health and Well-Being of Children in Immigrant Families*. National Academy Press, Washington, DC
- Ingstad B, Reynolds Whyte S 1995 *Disability and Culture*. University of California Press, Berkeley, CA
- Kleinman A 1995 *Writing at the Margin: Discourse Between Anthropology and Medicine*. University of California Press, Berkeley, CA
- Leslie C, Young A 1992 *Paths to Asian Medical Knowledge*. University of California Press, Berkeley, CA
- Lock M 1993 *Encounters with Aging: Mythologies of Menopause in Japan and North America*. University of California Press, Berkeley, CA
- Mechanic D 1997 The social context of health and disease and choices among health interventions. In: Brandt A, Rozin P (eds.) *Morality and Health*. Routledge, New York
- Paul B 1955 *Health, Culture, and Community: Case Studies of Public Reactions to Health Programs*. Russell Sage Foundation, New York
- Phillips M R 1998 The transformation of China's mental health services. *China Journal* 39: 1-36
- Rabinow P 1996 *Essays on the Anthropology of Reason*. Princeton University Press, Princeton, NJ
- Rapp R 1999 *Testing Women, Testing the Fetus: The Social Impact of Amniocentesis in America*. Routledge, New York
- Scheper-Hughes N 1992 *Death Without Weeping: The Violence of Everyday Life in Brazil*. University of California Press, Berkeley, CA
- Turner B 1995 *Medical Power and Social Knowledge*, 2nd edn. Sage, London

A. Kleinman and A. Petryna

Health Behavior: Psychosocial Theories

The term 'health behavior' is used very broadly in this article to refer to any behavior that influences, or is believed to influence, physical health outcomes, either by increasing or decreasing their risk or severity. A number of psychosocial theories (or models—the two terms are used interchangeably in this article; see also *Health Behaviors*) has been developed to predict, explain, and change such health behaviors. These theories can be divided into two main groups which

are commonly referred to as social cognition models and stage models, respectively. These are discussed in turn.

1. Social Cognition Models

The term 'social cognition models' is used to refer to a group of similar theories, each of which specifies a small number of cognitive and affective factors ('beliefs and attitudes') as the proximal determinants of behavior. These models do not deny that behavior is influenced by many other factors (e.g., social structural, cultural, and personality factors), but they assume that the effects of such distal factors are largely or completely mediated by the proximal factors specified by the model. Unlike the distal factors, the proximal factors are assumed to be amenable to change, for example by provision of relevant information. Thus social cognition models can be used as the basis for health behavior interventions.

The five models that have been used most widely by health behavior researchers in recent years are shown in Fig. 1 in the form of causal diagrams. The diagrams depict the core constructs in each model; numerous variations appear in the literature.

1.1 The Health Belief Model

The health belief model (HBM; Becker 1974) was developed in the 1950s by a group of social psychologists working in the field of public health who were seeking to explain why some people do not use health services such as immunization and screening. The model is still in common use. There are four core constructs: the first two refer to a particular disease whereas the second two refer to a possible course of action that may reduce the risk or severity of that disease. Perceived susceptibility (or perceived vulnerability) is the individual's perceived risk of contracting the disease if he or she were to continue with the current course of action. Perceived severity refers to the seriousness of the disease and its consequences as perceived by the individual. Perceived benefits refer to the perceived advantages of the alternative course of action including the extent to which it reduces the risk of the disease or the severity of its consequences. Perceived barriers (or perceived costs) refers to the perceived disadvantages of adopting the recommended action as well as perceived obstacles that may prevent or hinder its successful performance. These factors are commonly assumed to combine additively to influence the likelihood of performing the behavior. Thus, high susceptibility, high severity, high benefits and low barriers are assumed to lead to a high probability of adopting the recommended action. Another factor that is frequently mentioned in connection with the HBM is cues to action (events that

trigger behavior), but little empirical work has been conducted on this construct.

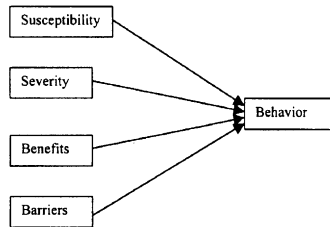
There have been two meta-analyses (quantitative reviews) of research using the HBM. Janz and Becker (1984) calculated significance ratios showing how often each HBM construct was statistically significant in the predicted direction across 46 studies. These ratios were 81 percent for susceptibility, 65 percent for severity, 78 percent for benefits, and 89 percent for barriers; the pattern of findings was similar when only the prospective studies were examined. Thus, barriers is the most consistent predictor of behavior and severity is the least consistent. Harrison et al. (1992) used extremely strict inclusion criteria; they included only 16 of the 234 studies they originally identified. Across these 16 studies, the mean correlations between HBM components and behavior were 0.15, 0.08, 0.13, and -0.21 for susceptibility, severity, benefits, and barriers, respectively. While statistically significant, these correlations are small in substantive terms. Harrison et al. found that benefits and barriers had significantly larger effect sizes in prospective compared with retrospective studies, whereas severity had a significantly larger effect size in retrospective studies.

1.2 Protection Motivation Theory

Protection motivation theory (PMT; Rogers 1983) was originally developed to explain how people respond to fear-arousing health threat communications or 'fear appeals.' It can be regarded as an adaptation of the HBM. Protection motivation refers to the motivation to protect oneself against a health threat; it is usually defined operationally as the intention to adopt the recommended action. Of the determinants of intention specified by the model, the four that have received the most empirical attention are vulnerability and severity (equivalent to perceived susceptibility and severity in the HBM), response efficacy (the belief that the recommended action is effective in reducing the threat), and perceived self-efficacy (the belief that one can successfully perform the recommended action; Bandura 1997). Thus, a person will be more motivated to protect himself or herself (i.e., have a stronger intention to adopt the recommended action) to the extent that he or she believes that the threat is likely if the current course of action is continued, that the consequences will be serious if the threat occurs, that the recommended action is effective in reducing the likelihood or the severity of the threat, and that he or she is able to carry out the recommended action.

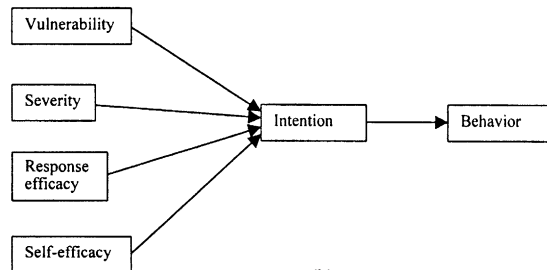
In many studies using this model (e.g., Wurtele and Maddux 1987), specific PMT variables are experimentally manipulated in a factorial design and their effects on intention (and sometimes behavior) are measured. In fact, PMT is unique among social cognition models with respect to the relatively large

Health Belief Model (HBM)



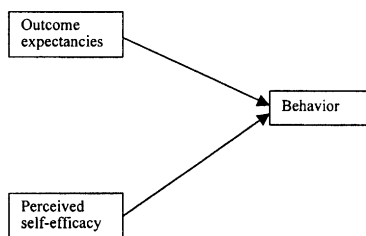
(a)

Protection Motivation Theory (PMT)



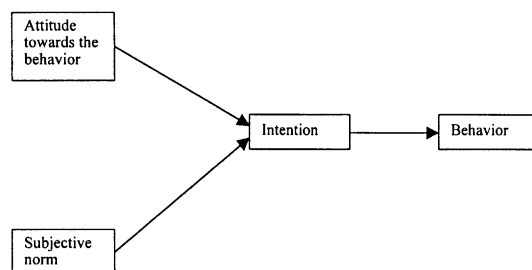
(b)

Self-efficacy Theory (SET)



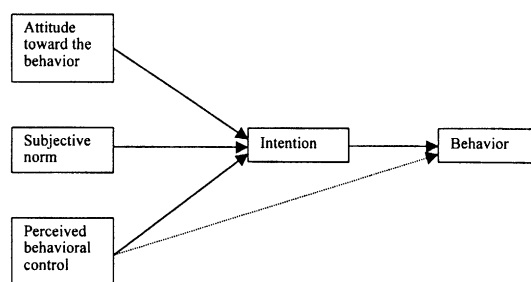
(c)

Theory of Reasoned Action (TRA)



(d)

Theory of Planned Behavior (TPB)



(e)

Figure 1

Causal diagrams for five social cognition models. In some cases these are simplified versions of the original models: (a) HBM; (b) PMT; (c) SET; (d) TRA; (e) TPB

number of experimental tests that have been conducted.

To date, two meta-analyses of PMT studies have been conducted (Floyd et al. 2000, Milne et al. 2000). The analyses used different study inclusion criteria and different effect size measures. Floyd et al. analyzed 65 studies with about 30,000 research participants whereas Milne and colleagues included 27 studies with about 8,000 participants. There were only 12 studies in common. Both analyses found support for each of the main PMT variables as predictors of intentions and/or behavior. Self-efficacy had the strongest, most consistent, and most robust effect.

1.3 Self-efficacy Theory

Self-efficacy theory (SET) is a subset of Bandura's (1986) social cognitive theory. According to this approach, the two key determinants of behavior are perceived self-efficacy and outcome expectancies. The latter construct refers to the perceived positive and negative consequences of performing the behavior. See Schwarzer and Fuchs (1996) for a version of this model that incorporates risk perceptions and behavioral intention, as well as components of the action phase of behavior change. No meta-analysis of SET has been published, though there is substantial evidence for the predictive validity of self-efficacy (Schwarzer and Fuchs 1996).

1.4 The Theory of Reasoned Action and the Theory of Planned Behavior

The theory of reasoned action (TRA; Ajzen and Fishbein 1980) developed out of social-psychological research on attitudes and the attitude-behavior relationship. The model assumes that most behaviors of social relevance (including health behaviors) are under volitional control, and that a person's intention to perform a behavior is both the immediate determinant and the single best predictor of that behavior. Intention in turn is held to be a function of two basic determinants: attitude towards the behavior (the person's overall evaluation of performing the behavior) and subjective norm (the perceived expectations of important others with regard to the individual performing the behavior in question). Generally speaking, people will have strong intentions to perform a given action if they evaluate it positively and if they believe that important others think they should perform it. The relative importance of the two factors may vary across behaviors and populations.

Although not shown in Fig. 1, the TRA also specifies the determinants of attitude and subjective norm. Attitude is held to reflect the person's salient behavioral beliefs concerning the possible personal con-

sequences of the action. For example, a person who believes that performing a given behavior will lead to mostly positive personal consequences will hold a favorable attitude towards the behavior. Specifically, attitude is held to be a function of the sum of the person's salient behavioral beliefs concerning the outcome of the action each weighted by their evaluation of that outcome. An indirect, belief-based, measure of attitude can be created by multiplying each behavioral belief by its corresponding outcome evaluation and then summing over outcomes. In a similar way, subjective norm is a function of the person's beliefs that specific individuals or groups think he or she should, or should not, perform the behavior. A person who believes that most significant referents think he or she should perform the behavior will perceive social pressure to do so. Specifically, subjective norm is held to be a function of the person's salient normative beliefs with respect to each referent, each weighted by their motivation to comply with that referent. An indirect measure of subjective norm can be created by multiplying each normative belief by its corresponding motivation to comply and then summing over referents.

Many behaviors cannot simply be performed at will; they require skills, opportunities, resources, or cooperation for their successful execution. The theory of planned behavior (TPB; Ajzen 1991) was an attempt to extend the TRA to include behaviors that are not entirely under volitional control, for example giving up smoking or using a condom. To accommodate such behaviors, Ajzen added a variable called perceived behavioral control to the TRA. This refers to the perceived ease or difficulty of performing the behavior, and is assumed to reflect past experience as well as anticipated obstacles. According to Ajzen, perceived behavioral control is a function of control beliefs in just the same way as subjective norm is a function of normative beliefs. It is assumed to have a direct influence on intention. For desirable behaviors, greater perceived behavioral control should lead to stronger intentions. Perceived behavioral control may also have a direct predictive effect on behavior, through two different mechanisms. First, holding intention constant, an individual with higher perceived behavioral control is likely to try harder and to persevere for longer than an individual who has lower perceived control. Second, people may have accurate perceptions of the amount of actual control they have over the behavior.

A number of meta-analyses of the TRA/TPB have been conducted. The findings show that when intention is predicted from attitude and subjective norm, or from attitude, subjective norm and perceived behavioral control, between 40 and 50 percent of the variance is explained, on average. When behavior is predicted from intention alone or from intention and perceived behavioral control, between 19 and 38 percent of the variance is explained (Sutton 1998).

1.5 Comparison of Social Cognition Models

The five social cognition models outlined above show a number of important similarities and differences. Some constructs are common to more than one model. For example, perceived susceptibility or perceived vulnerability occurs in both the HBM and PMT. Other constructs appear to be very similar, for example, perceived behavioral control and self-efficacy. Resolution of current controversies concerning the extent of overlap between such constructs requires the development of clear definitions, so that similar constructs can be distinguished on conceptual grounds, and more frequent tests of discriminant validity to investigate whether sets of apparently similar measures are tapping the same or different constructs.

All the models assume that individuals are future oriented and that they weigh up the costs and benefits of possible future courses of action. They all incorporate to a greater or lesser extent the expectancy-value principle which derives from the classical expected utility model, a normative model of decision making. In its weak form, this principle states that expectancies (subjective probabilities) and values are important determinants of behavior. In its strong form, the principle states that expectancies and values are combined multiplicatively, or at least that people behave as if they combine their cognitions in this way. Both the TRA and the TPB employ the strong form of the expectancy-value principle. The most systematic application of the principle is found in the subjective expected utility model (SEUM; Edwards 1954) which is based directly on expected utility theory. Versions of the SEUM have been applied to a number of health behaviors (e.g., Sutton et al. 1987). Researchers who use models such as the TPB and the SEUM need to be aware of the problems that arise from the use of multiplicative composites (Evans 1991).

This raises the thorny issue of rationality. Social cognition models are often criticized for offering an unrealistically rational account of how people form intentions and make decisions. However, the models do not imply that individuals always deliberate carefully and always make optimal decisions. People may not be aware of all the options available to them and of all the consequences that may follow from their actions. They may hold incorrect beliefs about the outcomes. They may make rapid decisions based on a few salient considerations. Having made a considered decision (e.g., to go jogging every Sunday morning), they do not necessarily have to weigh up the pros and cons again unless circumstances change; they may simply retrieve their previously formed intention from long-term memory and act on it. Thus, social cognition models imply a more limited rationality than is sometimes suggested by their critics.

Social cognition models are also sometimes criticized for being static. This criticism is unfounded:

social cognition models summarize dynamic causal processes. In the TRA, for example, changes in behavioral beliefs and/or outcome evaluations are assumed to produce changes in attitude which in turn lead to changes in intention which ultimately produce changes in behavior. Models of health behavior should specify the time lags involved in these causal processes but most of them do not. However, it is often assumed implicitly that effects on intention are almost instantaneous whereas effects on behavior may be delayed.

Social cognition models differ in the degree to which they specify the content of the cognitions they identify. With the TRA, for example, once the behavior of interest has been defined, it is possible to generate questionnaire items for intention and for the direct measures of attitude and subjective norm. However, in order to generate items for behavioral beliefs, outcome evaluations, normative beliefs and motivations to comply, it is recommended that researchers gather information on salient beliefs from members of the target population. These constructs remain 'content-free' until such information is obtained.

Compared with the other models, research on the TRA and the TPB shows a relatively high degree of standardization of measures based on published recommendations (Ajzen and Fishbein 1980). In addition, these authors emphasize the principle of correspondence (or compatibility) which, put simply, states that for maximum prediction the measures of all the constructs in the model should use similar wording. This principle is not widely applied in research using the other social cognition models.

The models also differ with regard to their scope of application. Key constructs in the HBM and PMT include perceived susceptibility and perceived severity with respect to a given health threat. Although these components can be extended to nonhealth-related events, for example the risk of financial loss, the scope of both models is necessarily limited by the nature of these two constructs. By contrast, the other three models are general theories that can be applied to any domain or behavior, for example voting, career, and purchasing decisions. Stated differently, SET, and the TRA and the TPB, regard health behaviors as having the same proximal determinants as other kinds of behavior. Thus, they offer the potential benefit of parsimony.

2. Stage Models

Stage models use similar concepts to social cognition models but organize them in a different way. They are fundamentally different in structure from social cognition models (Weinstein et al. 1998). According to this approach, behavior change involves movement through a sequence of discrete, qualitatively distinct, stages. Different factors are assumed to be important at different stages. Hence, people in different stages are

Table 1

Stage definitions (from DiClemente et al. 1991)

Precontemplation

Currently smoking and 'not seriously considering quitting within the next 6 months'

Contemplation

Currently smoking and 'seriously considering quitting within the next 6 months'; 'were not considering quitting within the next 30 days, had not made a quit attempt of 24h in the past year, or both'

Preparation

Currently smoking, 'were seriously considering quitting in the next 6 months and were planning to quit within the next 30 days,' and 'had made a 24h quit attempt in the past year'

Action

Currently not smoking; quit in last 6 months

Maintenance

Currently not smoking; quit > 6 months ago

assumed to require different interventions to encourage or help them to move to the next stage in the sequence.

Current models of health behavior that incorporate stage assumptions include: the transtheoretical model (TTM; Prochaska et al. 1992) and variants of it (e.g., De Vries and Mudde 1998), the precaution adoption process model (Weinstein and Sandman 1992), the health action process approach (Schwarzer and Fuchs 1996), and the health behavior goal model (Gebhardt 1997). This section focuses on the TTM because it is the dominant model in the field.

The TTM developed from studies of the processes of change in psychotherapy and smoking cessation. Smoking still accounts for the majority of applications of the model but it has been applied to a wide range of other health behaviors, including condom use, exercise, sunscreen use, and healthy eating (Prochaska and Velicer 1997). Although it is often referred to simply as the stages of change model, the TTM includes 15 different theoretical constructs drawn from different theories of behavior change. These include the stages of change (which provide the basic organizing principle), the 10 processes of change, the perceived pros and cons of changing, and self-efficacy and temptation. The TTM was an attempt to integrate these different constructs in a single comprehensive framework—hence the name transtheoretical.

The version of the TTM used most widely in recent years specifies five stages: precontemplation, contemplation, preparation, action, and maintenance (DiClemente et al. 1991). Table 1 gives the operational definitions of the stages as applied to smoking cessation. Using a 'staging algorithm,' participants are classified into stages on the basis of their responses to a small number of questionnaire items. The first three categories contain the current smokers, the remaining two the ex-smokers; people who have never smoked are not represented in this scheme. Precontemplation, contemplation, and preparation are defined in terms

of current intentions and past behavior (whether or not the person has made a 24h quit attempt in the past year), whereas action and maintenance are defined purely in terms of behavior; ex-smokers' intentions are not taken into account.

Prochaska et al. (1992) represented the stages of change as a spiral. People start at the bottom in precontemplation. They then move through the stages in order (contemplation, preparation, action, maintenance) but may relapse to an earlier stage. They may cycle and recycle through the stages several times before achieving successful long-term behavior change.

Of the other constructs in the TTM, the 10 processes of change refer to the things that people think and do to help them move through the stages, the pros and cons are the perceived advantages and disadvantages of changing, self-efficacy is borrowed from Bandura's social cognitive theory, and temptation is a related concept that refers to the degree to which a person expects to feel tempted to lapse in different situations.

The TTM is popular among clinicians and practitioners and there is a large body of evidence which is interpreted by its proponents as supporting the model. However, a closer examination of this literature reveals a number of serious problems with the TTM itself and with much of the research based on it (Sutton 2000). These problems include: lack of standardization of measures, particularly of the central construct of stages of change; logical flaws in current staging algorithms; inadequate specification of the causal relationships among the different constructs; misinterpretation of cross-sectional data on stages of change; and confusion concerning the nature of stage models and how they should be tested.

It is unfortunate that the model that has dominated the field to date is surrounded by a 'thicket of problems,' to use Bandura's (1997) phrase. The notion that behavior change involves movement through a sequence of qualitative stages is an important idea that

deserves further consideration. Health behavior researchers should turn their attention to the other stage models that have been proposed.

3. Conclusion

One focus of current interest is the 'intention-behavior gap.' Researchers in the health behavior field are using Gollwitzer's (1993) concept of implementation intentions (the intention to do X in situation Y) and related ideas to explain why some people who have strong intentions to achieve a goal succeed whereas others fail, and to try to close this gap using simple interventions (e.g., Orbell et al. 1997). Related to this is the increasing interest in models of self-regulation (Abraham and Johnston 1998). Another trend, which has been apparent in social psychology in recent years and is now being imported into research on health behaviors, is to regard attitudes as being activated automatically (e.g., Fazio 1990). Although this seems to pose a challenge to social cognition models, the two approaches are not necessarily incompatible, especially if the idea is extended to automatic activation of previously formed intentions.

There is a plethora of models of health behavior and new models are continually being developed. Although this can be interpreted as a sign of a field rich in conceptual and theoretical development, it makes it increasingly difficult to accumulate research findings into a coherent body of knowledge. The field would benefit from clearer definition of concepts, greater standardization of measures, more tests of convergent and discriminant validity, greater concentration on a small number of models, and more empirical comparisons of models. The vast majority of studies use nonexperimental between-subjects designs. Obtaining repeated measures on the same individuals would allow models to be tested at the individual level (Hedeker et al. 1996). Experimental manipulation of constructs, as has been done in many studies of PMT, would provide much stronger evidence of causal effects.

See also: Control Beliefs; Health Perspectives; Health Behaviors; Health Behaviors, Assessment of; Health Risk Appraisal and Optimistic Bias; Health: Self-regulation; Illness Behavior and Care Seeking; Patient Adherence to Health Care Regimens; Self-efficacy and Health; Vulnerability and Perceived Susceptibility, Psychology of

Bibliography

- Abraham C, Johnston M (eds.) 1998 Special issue on self-regulation and health. *Psychology and Health* **13**: 565-766
 Ajzen I 1991 The theory of planned behavior. *Organizational Behavior and Human Decision Processes* **50**: 179-211

- Ajzen I, Fishbein M 1980 *Understanding Attitudes and Predicting Social Behavior*. Prentice-Hall, Englewood Cliffs, NJ
 Bandura A 1986 *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice-Hall, Englewood Cliffs, NJ
 Bandura A 1997 *Self-efficacy: The Exercise of Control*. Freeman, New York
 Becker M H (ed.) 1974 *The Health Belief Model and Personal Health Behavior*. Slack, Thorofare, NJ
 De Vries H, Mudde A N 1998 Predicting stage transitions for smoking cessation applying the attitude-social influence-efficacy model. *Psychology and Health* **13**: 369-85
 DiClemente C C, Prochaska J O, Fairhurst S K, Velicer W F, Velasquez M M, Rossi J S 1991 The process of smoking cessation: an analysis of precontemplation, contemplation, and preparation stages of change. *Journal of Consulting and Clinical Psychology* **59**: 295-304
 Edwards W 1954 The theory of decision making. *Psychological Bulletin* **54**: 380-417
 Evans M G 1991 The problem of analyzing multiplicative composites. *American Psychologist* **46**: 6-15
 Fazio R H 1990 Multiple processes by which attitudes guide behavior: the MODE model as an integrative framework. *Advances in Experimental Social Psychology* **23**: 75-109
 Floyd D L, Prentice-Dunn S, Rogers R W 2000 A meta-analysis of research on protection motivation theory. *Journal of Applied Social Psychology* **30**: 407-29
 Gebhardt W A 1997 *Health Behaviour Goal Model: Towards a Theoretical Framework for Health Behavior Change*. University of Leiden, Leiden, The Netherlands
 Gollwitzer P M 1993 Goal achievement: the role of intentions. *European Review of Social Psychology* **4**: 141-85
 Harrison J A, Mullen P D, Green L W 1992 A meta-analysis of studies of the Health Belief Model with adults. *Health Education Research* **7**: 107-16
 Hedeker D, Flay B R, Petraitis J 1996 Estimating individual influences of behavioral intentions: an application of random-effects modeling to the theory of reasoned action. *Journal of Consulting and Clinical Psychology* **64**: 109-20
 Janz N K, Becker M H 1984 The health belief model: a decade later. *Health Education Quarterly* **11**: 1-47
 Milne S, Sheeran P, Orbell S 2000 Prediction and intervention in health-related behavior: a meta-analytic review of protection motivation theory. *Journal of Applied Social Psychology* **30**: 106-43
 Orbell S, Hodgkins S, Sheeran P 1997 Implementation intentions and the theory of planned behavior. *Personality and Social Psychology Bulletin* **23**: 945-54
 Prochaska J O, DiClemente C C, Norcross J C 1992 In search of how people change: applications to addictive behaviors. *American Psychologist* **47**: 1102-14
 Prochaska J O, Velicer W F 1997 The transtheoretical model of health behavior change. *American Journal of Health Promotion* **12**: 38-48
 Rogers R W 1983 Cognitive and physiological processes in fear appeals and attitude change: a revised theory of protection motivation. In: Cacioppo J T, Petty R E, Shapiro D (eds.) *Social Psychophysiology: A Sourcebook*. Guilford Press, New York, pp. 153-76
 Schwarzer R, Fuchs R 1996 Self-efficacy and health behaviours. In: Conner M, Norman P (eds.) *Predicting Health Behaviour: Research and Practice with Social Cognition Models*. Open University Press, Buckingham, UK, pp. 163-96
 Sutton S 1998 Predicting and explaining intentions and behavior: how well are we doing? *Journal of Applied Social Psychology* **28**: 1317-38

- Sutton S, Marsh A, Matheson J 1987 Explaining smokers' decisions to stop: test of an expectancy-value approach. *Social Behaviour* 2: 35-49
- Sutton S A 2000 A critical review of the transtheoretical model applied to smoking cessation. In: Norman P, Abraham C, Conner M (eds.) *Understanding and Changing Health Behaviour: From Health Beliefs to Self-Regulation*. Harwood Academic Publishers, Amsterdam, pp. 207-25
- Weinstein N D, Rothman A J, Sutton S R 1998 Stage theories of health behavior: conceptual and methodological issues. *Health Psychology* 17: 290-99
- Weinstein N D, Sandman P M 1992 A model of the precaution adoption process: evidence from home radon testing. *Health Psychology* 11: 170-80
- Wurtele S K, Maddux J E 1987 Relative contributions of protection motivation theory components in predicting exercise intentions and behavior. *Health Psychology* 6: 453-66

S. Sutton

Health Behaviors

1. Introduction

Interest in behaviors that have important impacts on our health and well being is based upon two assumptions; (a) that a significant proportion of the mortality from the leading causes of death is caused by the behavior of individuals, and (b) that such behavior is modifiable (Conner and Norman 1996). Behavior is held to exert its influence on health in three basic ways: by producing direct biological changes, by conveying health risks or protecting against them, or by leading to the early detection or treatment of disease (Baum and Posluszny 1999).

This chapter begins by offering a definition of health behavior and the different types of health behavior. The prevalence of key health behaviors and their relationship to morbidity and mortality is then examined. The chapter then considers who performs these different behaviors, why they might do so, and how health enhancing behaviors might be encouraged and health impairing behaviors discouraged.

2. The Definition of Health Behaviors

Health behaviors have been defined in various ways. For example, Conner and Norman (1996) define them as any activity undertaken for the purpose of preventing or detecting disease or for improving health and well being. Gochman (1997) in the *Handbook of Health Behavior Research* defines them as 'behavior patterns, actions and habits that relate to health maintenance, to health restoration and to health improvement' (Vol. 1, p. 3). Behaviors within this definition include medical service usage (e.g., physician visits, vacci-

nation, screening), compliance with medical regimens (e.g., dietary, diabetic, antihypertensive regimens), and self-directed health behaviors (e.g., diet, exercise, smoking, alcohol consumption). All have received considerable attention from social and behavioral researchers and we now have a good understanding of the factors influencing how and why individuals engage in such behaviors.

In describing health behaviors it is common to distinguish health enhancing from health impairing behaviors. Health impairing behaviors have harmful effects on health or otherwise predispose individuals to disease. Such behaviors include smoking, excessive alcohol consumption, and high dietary fat consumption. In contrast, engagement in health enhancing behaviors convey health benefits or otherwise protect individuals from disease. Such behaviors include exercise, fruit and vegetable consumption, and condom use in response to the threat of sexually transmitted diseases.

3. Key Health Behaviors

Numerous studies have examined the relationship between health behaviors and health outcomes (e.g., Blaxter 1990) and have demonstrated their role in both morbidity and mortality. One of the first such studies identified seven features of lifestyle which were associated with lower morbidity and higher subsequent long-term survival: not smoking, moderate alcohol intake, sleeping 7-8h per night, exercising regularly, maintaining a desirable body weight, avoiding snacks, and eating breakfast regularly (Belloc and Breslow 1972). Health behaviors also impact upon individuals' quality of life, by delaying the onset of chronic disease and extending active lifespan. Smoking, alcohol consumption, diet, gaps in primary care services and low screening uptake are all significant determinants of poor health, and changing such behaviors should lead to improved health. For example, in the USA, Healthy People 2000 (US Department of Health and Human Services [USDHHS] 1990) lists increased physical activity, changes in nutrition and reductions in tobacco, alcohol and drug use as important for health promotion and disease prevention.

3.1 Smoking

Smoking is the health behavior most closely linked with long-term negative health outcomes (see *Smoking and Health*). Morbidity and mortality from coronary heart disease (CHD) are increased among smokers (Doll et al. 1994). Moreover, there is a strong positive relationship between the number of cigarettes smoked per day and the incidence of CHD (Friedman et al. 1979). Smoking has also been linked to a number of cancers, including cancer of the lung, throat, stomach and bowel as well as a number of more immediate

negative health effects such as reduced lung capacity and bronchitis (Royal College of Physicians 1983). Despite the array of negative health outcomes, smokers often report positive mood effects from smoking and the use of smoking as a strategy for coping with stress.

The number of people smoking in the USA and UK has shown a steady decline over the past twenty years. Data from the General Household Survey showed that 28 percent of people over the age of 16 smoke in the UK. Smoking is more common among men and among unskilled manual workers (General Household Survey 1994). A similar pattern is evident in the USA and other First World countries, with smoking more common among less educated, lower income and minority groups (Rigotti 1989). Those who quit smoking reduce the risk to their health, particularly if they quit before 35 years of age (Doll et al. 1994) (see *Smoking Prevention and Smoking Cessation*).

3.2 Diet

The impact of diet upon morbidity and mortality are well established (USDHHS 1988). In the Third World, the problems related to diet and health are ones of under-nutrition; in the First World, the problems are predominantly linked to overconsumption of food. In Western industrialized countries excessive fat consumption and insufficient fiber, fruit and vegetable consumption are related to health problems. In addition, excess consumption of calories combined with insufficient exercise has made obesity a major health problem. Diet has been implicated in cardiovascular diseases (CVDs), strokes and high blood pressure, cancer, diabetes, obesity, osteoporosis, and dental disease.

It is generally agreed that elevated blood cholesterol level is a major risk factor for the development of CVD (Consensus Development Conference on Lowering Blood Cholesterol to Prevent Heart Disease 1985). Nutbeam and Catford (1990) estimate that 26 percent of men and 25 percent of women in the UK have cholesterol levels greater than 6.5 mmol l^{-1} (a level considered to be excessive). While in the USA, it is estimated that 50 percent of the adult population is at risk of CHD by virtue of elevated blood cholesterol levels (Sampos et al. 1989). The reduction of blood cholesterol via dietary change is now widely accepted as an important way of tackling CHD. Dietary recommendations include reducing fat in the diet and increasing soluble fiber intake. However, their impact upon cholesterol levels may be limited.

3.3 Exercise

The potential health benefits of engaging in regular exercise include reduced cardiovascular morbidity and

mortality, lowered blood pressure, and the increased metabolism of carbohydrates and fats, as well as a range of psychological benefits such as improved self-esteem, positive mood states, reduced life stress and anxiety (see *Physical Activity and Health*). Nevertheless, a significant proportion of the population lead a sedentary lifestyle. The General Household Survey (1989) indicated that only one in three men and one in five women in the UK participate in any sport or recreational physical activity. Moreover, the Allied Dunbar Fitness Survey (1992) of 6,000 English adults reported that one in six adults had done no exercise (i.e., for 20 minutes or more at a moderate or vigorous level) in the previous four weeks. Participation in regular exercise is strongly related to a number of sociodemographic variables. In particular, young people and males are more likely to engage in regular exercise. For example, the 1988 Welsh Heart Health Survey (HPAW 1990) reported that among 18–34 year olds, 61 percent of men engaged in moderately vigorous exercise at least two times a week compared with only 35 percent of women. For 35–64 year olds, the percentages drops to 37 percent for men and 17 percent for women. Overall, across First World countries the typical exerciser is likely to be young, well educated, affluent and male.

3.4 Health Screening

Individuals may seek to protect their health by participating in various screening programs which attempt to detect disease at an early, or asymptomatic, stage. In the UK, screening programs have been set up for various diseases, including anemia, diabetes, bronchitis, cervical cancer, and breast cancer. Considering breast cancer, it has been estimated that breast screening programs which include mammograms can reduce breast cancer mortality by up to 40 percent among women aged 50 and over (Strax 1984). However, participation rates in breast screening programs show great variability across different countries, ranging from 25 percent to 89 percent (Vernon et al. 1990). Participation tends to be negatively related to age, and positively related to education level and socioeconomic status.

3.5 Sexual Behaviors

Sexual behaviors are considered health behaviors because of their impact upon the spread of sexually transmitted diseases (STDs) such as gonorrhea and syphilis (see *Sexually Transmitted Diseases: Psychosocial Aspects*). More recently, the role of sexual behaviors in the spread of human immunodeficiency virus (HIV) has been a focus of attention. While early health education campaigns emphasized the need to reduce the number of sexual partners or avoid par-

ticular sexual practices (e.g., anal sex, penetrative sex), more recently the focus has been upon the use of condoms during penetrative sex to reduce the risk of HIV transmission. Condom use is particularly recommended for those with multiple partners or those who do not know their partners' sexual history. For these reasons, much of the health advice concerning condom use has been focused on young people (see *Sexual Risk Behaviors*).

There seems to be considerable variation in the use of condoms in response to the threat of HIV/AIDS. For example, studies in the UK and USA report rates of between 24 and 58 percent (Fife-Schaw and Breakwell 1992, Gerrard et al. 1996). The General Household Survey (1993) in the UK reported changes by age group in the use of condoms for the period 1983 to 1991. Among 16–24 year olds, condom use increased from around 6 percent to around 12 percent during this period, while among 40–49 year olds it dropped from around 18 percent to around 12 percent over the same period.

3.6 Alcohol Use

Moderate alcohol consumption has been linked to positive health outcomes. However, high alcohol consumption has been linked to a range of negative health outcomes including high blood pressure, heart disease and cirrhosis of the liver. High levels of alcohol consumption have also been associated with accidents, injuries, suicides, crime, domestic violence, rape, murder and unsafe sex (British Medical Journal 1982). While many of the adverse effects of high alcohol consumption are due to continued heavy drinking (e.g., cirrhosis of the liver, heart disease), others are more specifically related to excessive alcohol consumption in a single drinking session (e.g., accidents, violence).

The General Household Survey (1992) reported that the average weekly consumption of alcohol in the UK was 15.9 units (1 unit = 1 glass of wine, 1 measure of spirits, or 0.5 pints of beer) for men and 5.4 for women. In addition, 27 percent of men and 11 percent of women were drinking more than the recommended weekly sensible limits (21 units for men, 14 units for women). Heavy drinking is also more likely among younger age groups. In a survey of 12,000 Welsh adults, Moore et al. (1994) reported that 31.1 percent of drinkers aged 18–24 engaged in binge drinking (i.e., drinking half the recommended weekly consumption of alcohol in a single session) at least once a week.

4. Relationship of Health Behaviors to Sociodemographic Factors

A clearer understanding of why individuals perform health behaviors might assist in the development of

interventions to help individuals gain health benefits. A variety of factors have been found to account for individual differences in the performance of health behaviors. Demographic variables show reliable associations with the performance of health behaviors. For example, there is a curvilinear relationship between many health behaviors and age, with high incidences of many health risking behaviors such as smoking in young adults and much lower incidences in children and older adults (Blaxter 1990). Such behaviors also vary by gender, with females being generally less likely to smoke, consume large amounts of alcohol, engage in regular exercise, but more likely to monitor their diet, take vitamins and engage in dental care (Waldron 1988). Differences by socioeconomic status and ethnic group are also apparent for behaviors such as diet, exercise, alcohol consumption and smoking (e.g., Blaxter 1990) (see *Socioeconomic Status and Health*).

Generally speaking, younger, wealthier, better educated individuals, under low levels of stress, with high levels of social support are more likely to practice health protective behaviors. Higher levels of stress and/or fewer resources are associated with health risking behaviors such as smoking and alcohol abuse (Adler and Matthews 1994). Social factors seem to be important in instilling health behaviors in childhood. Parent, sibling and peer influences are important, for example in the initiation of smoking. Cultural values also have a major impact, for instance in determining the number of women exercising in a particular culture. For example, Steptoe and Wardle (1992) report that between 34 and 95 percent of women in their European student sample had exercised in the past 14 days.

5. Understanding the Distribution/Prevalence of Health Behaviors

Perceived symptoms control health behaviors when, for example, a smoker regulates his/her smoking on the basis of sensations in the throat. Access to medical care has been found to influence the use of such health services (e.g., Black Report 1988). Personality factors have also been associated with health behaviors (Adler and Matthews 1994). Cognitive factors also determine whether or not an individual practices health behaviors and may explain how other factors influence behavior. Knowledge about behavior–health links is an important factor in an informed choice concerning health behaviors. Various cognitive variables have been studied including perceptions of health risk, efficacy of behaviors in influencing this risk, social pressures to perform the behavior, and control over performance of the behavior. The relative importance of various cognitive factors in determining who performs various health behaviors constitutes the basis of different models. Such models have been labeled social cognition models (SCMs) because of their focus

on cognitive variables as the primary determinant of individual social behaviors. These SCMs provide a basis for understanding the determinants of behavior and behavior change (see *Health Behavior: Psychosocial Theories*). Each of these models emphasize the rationality of human behavior and assume that behavior is based upon elaborate, but subjective, cost-benefit analysis of the likely outcomes of differing courses of action. It is assumed that individuals generally aim to maximize benefits and minimize costs in selecting a behavior.

5.1 Health Belief Model

The Health Belief Model (HBM) attempts to conceptualize two types of health beliefs that make a behavior in response to illness more or less attractive (Sheeran and Abraham 1996): perceptions of the threat of illness and evaluation of the effectiveness of behaviors to counteract this threat. Threat perceptions depend upon the perceived susceptibility to the illness and the perceived severity of the consequences of the illness. Together these variables determine the likelihood of the individual following a health-related action, although their effect is modified by demographic variables, social pressure and personality. The particular action undertaken is determined by the evaluation of the possible alternatives. This behavioral evaluation depends upon beliefs concerning the benefits or efficacy of the health behavior and the perceived costs or barriers to performing the behavior. Hence, individuals are likely to follow a particular health behavior if they believe themselves to be susceptible to a particular condition or illness which they consider to be serious, and believe the benefits of the behavior undertaken to counteract the condition or illness outweigh the costs. It is assumed that this whole process is set in motion by cues to action. Cues to action include a diverse range of triggers to the individual taking action and are commonly divided into factors that are internal (e.g., physical symptoms) or external (e.g., mass media campaigns, advice from others) to the individual. Other influences upon the performance of health behaviors, such as demographic factors or psychological characteristics (e.g. personality, peer pressure, perceived control over behavior) are assumed to exert their effect via changes in the components of the HBM.

5.2 Theory of Planned Behavior

The Theory of Planned Behavior (TPB) was developed by social psychologists and has been widely employed as a tool to aid our understanding of a variety of behaviors including health behaviors (Ajzen 1991, Godin and Kok 1996). The TPB details how the influences upon an individual determine that indi-

vidual's decision to follow a particular behavior. Within the TPB, the determinants of behavior are intentions to engage in that behavior and perceived behavioral control (PBC) over that behavior. Intentions represent a person's motivation. The construct is conceptualized as an individual's conscious plan or decision to exert effort in order to engage in a particular behavior. Perceived behavioral control is a person's expectancy that performance of the behavior is within his/her control. Intentions are determined by three variables. The first is attitudes, which are an individual's overall evaluation of the behavior. The second is subjective norms, which consist of a person's beliefs about whether significant others think he/she should engage in the behavior. The third measures the extent to which the individual perceives that the behavior is under their personal control and is labeled PBC.

The attitude, subjective norm and PBC components are determined by underlying beliefs. Attitude is a function of a person's salient behavioral beliefs; which represent perceived likely consequences of the behavior (e.g., taking exercise will reduce my risk of heart disease). Subjective norm is a function of normative beliefs, which represent perceptions of specific salient others' preferences about whether one should or should not engage in a behavior (e.g., my family think I should take exercise). PBC is based on beliefs concerning access to the necessary resources and opportunities to perform the behavior successfully (e.g., I have easy access to a place where I can exercise).

So, according to the TPB, individuals are likely to engage in a health behavior if they believe that the behavior will lead to particular outcomes which they value, if they believe that people whose views they value think they should carry out the behavior, and if they feel that they have the necessary resources and opportunities to perform the behavior.

5.3 Perceived Self-efficacy

Self-efficacy is one of the most powerful predictors of health behavior (Bandura 1997, Schwarzer and Fuchs 1996). It has its origins in Bandura's social cognitive theory, which states that behavior is a function of both incentives (i.e., reinforcements) and expectancies. Three kinds of expectancies can be identified, these being: situation-outcome expectancies, which refer to beliefs about how events are connected; outcome expectancies, which refer to beliefs about the consequences of performing a behavior; and self-efficacy expectancies, which refer to beliefs about one's ability to perform the behavior. Thus, in order to perform a health behavior, individuals must value their health (i.e., have an incentive), believe that their current lifestyle poses a threat to their health (i.e., situation-outcome expectancy), believe that adopting the new behavior will reduce the threat to their health (i.e., outcome expectancy), and believe that they are capable

of performing the behavior (i.e., self-efficacy expectancy). While all these beliefs are seen to be important in the initiation and maintenance of health behavior, perceived self-efficacy is seen to be the most important. Individuals with a strong sense of self-efficacy are believed to develop stronger intentions to act, to expend more effort to achieve their goals and to persist longer in the face of barriers and impediments. Perceived self-efficacy is therefore believed to play a crucial role in the determination of health behavior (see *Health: Self-regulation*). According to Bandura (1997), self-efficacy can be conceptualized and measured in terms of three parameters; magnitude, strength and generality, with strength being the key parameter. This is reflected in its measurement (e.g., 'I am confident that I can refrain from smoking, even if someone offers me a cigarette').

5.4 Stage Models of Health Behavior

A number of researchers have suggested that there may be qualitatively different stages in the initiation and maintenance of health behavior, and that to obtain a full understanding of the determinants of health behavior it is necessary to conduct a detailed analysis of the nature of these stages. One of the first stage models was put forward by Prochaska and DiClemente in their transtheoretical model of change (TTM). Their model has been widely applied to analyze the process of change in alcoholism treatment, smoking cessation, and psychotherapy. In its most widely applied form, Prochaska et al. (1992) identify five stages of change: precontemplation, contemplation, preparation, action, and maintenance. Individuals are seen to progress through each stage to achieve successful maintenance of a new behavior. Taking the example of smoking cessation, it is argued that in the precontemplation stage the smoker is unaware that his/her behavior constitutes a problem and has no intention to quit. In the contemplation stage, the smoker starts to think about changing his/her behavior, but is not committed to try to quit. In the preparation stage, the smoker has an intention to quit and starts to make plans about how to quit. The action stage is characterized by active attempts to quit, and after six months of successful abstinence the individual moves into the maintenance stage characterized by attempts to prevent relapse and to consolidate the newly acquired nonsmoking status.

6. Changing Health Behaviors

The above models detail the key cognitive determinants of health behavior. To the extent that these models outline the key social cognitive determinants of health behavior, interventions which target these variables should lead to associated changes in beha-

vior. For example, the HBM would suggest that encouraging health behaviors is best achieved by increasing individuals' perceived susceptibility to negative health outcomes and making individuals aware of the severity of such outcomes. Such approaches have been commonly employed in health promotion messages. In addition, the HBM might suggest the need to focus on the benefits of health behaviors and the fact that barriers to action are easily overcome. However, there has been little systematic evaluation of their effectiveness, perhaps due to the common sense appeal of these approaches.

Relatively few studies have used the TPB as a framework for developing interventions, despite the fact that the theory would suggest a number of interventions focusing on different components of the model. Brubaker and Fowler (1990) did examine the effect of persuasive messages tackling behavioral beliefs upon men's intentions to perform testicular self-examination in response to the threat of cancer. A persuasive message was found to increase intentions to perform testicular self-examination compared to a no message control. The TPB would also suggest the need to tackle normative beliefs and control beliefs in any attempt to change behavior. A number of studies have attempted to use persuasive messages aimed at tackling normative pressures. For example, in relation to preventing adolescent smoking, prevention programs commonly attempt to tackle the perceived pressure from teenage peers who smoke. Tackling control beliefs has been seen to bear many similarities to changing perceptions of self-efficacy (see below).

Another interesting approach has focused directly on the immediate determinant of behavior in the TPB: intentions. Where individuals do have an intention to engage in a health behavior (goal intentions), but are having trouble implementing their intention, forming a specific plan about where and when to act has been found to help (Gollwitzer 1993). For example, Orbell et al. (1997) gave out a questionnaire about breast self-examination in response to the threat of cancer. Half the women were asked to indicate when and where in the next month they intended to perform breast self-examination (a specific plan or implementation intention). A one month follow-up found that 64 percent of these women had performed breast self-examination that month compared with only 16 percent of women who hadn't made an implementation intention, despite having similar goal intentions.

Several studies have focused on enhancing feelings of self-efficacy as a means for encouraging health behavior change. As Bandura (1997) outlines, there are four main sources of self-efficacy, each of which could be addressed in interventions. First, individuals can develop feelings of self-efficacy from personal mastery experience (e.g., it may be possible to split a behavior into various subgoals, such that the easiest subgoals are achieved before more difficult tasks are attempted). Second, individuals may develop feelings

of self-efficacy through observing other people succeed on a task (i.e., vicarious experience). Third, it is possible to use standard persuasive techniques to try to instill feelings of self-efficacy. Finally, one's physiological state may be used as a source of information, such that high levels of arousal or anxiety may indicate to the individual that he or she is not capable of performing a given action (e.g., relaxation techniques may be employed to help maintain feelings of self-efficacy).

Stage models have also been used as the basis for designing interventions. The main distinction from the approaches already presented is that this approach suggests that interventions need to be matched to the needs of each individual. So, for example, an individual at the pre-contemplation stage in relation to exercise needs to be made aware of the health problems associated with a lack of exercise. A different individual in the contemplation stage needs information about the pros of changing their exercise behavior and the cons of not changing. While an individual in the preparation or action stage needs helps with acting on their plans to engage in exercise. Whilst these ideas appear entirely plausible, to date relatively few studies have managed to demonstrate that such stage-matched interventions are more effective than presenting the same intervention to all individuals.

7. Conclusions

Health behaviors have important consequences for both the quality and length of life by influencing disease outcomes. Nevertheless, there is still considerable variation in those individuals who perform these behaviors. Social cognition models provide one approach to understanding the variation in who performs health behaviors. These models are also useful because they suggest ways in order to change health behaviors in order to improve health. Perhaps the greatest challenge for social and behavioral research on health behaviors is the demonstration that such theory-based interventions can produce effective and long-lasting behavior change that will lead to real health benefits for all individuals.

See also: Health Education and Health Promotion; Health Risk Appraisal and Optimistic Bias

Bibliography

- Adler N, Matthews K 1994 Health psychology: why do some people get sick and some stay well? *Annual Review of Psychology* **45**: 229–59
- Ajzen I 1991 The theory of planned behavior. *Organizational Behavior and Human Decision Processes* **50**: 179–211
- Allied Dunbar Fitness Survey 1992 Health Education Authority, London
- Bandura A 1997 *Self-efficacy: The Exercise of Control*. W. H. Freeman, New York
- Baum A, Poslusny D M 1999 Health psychology: mapping biobehavioral contributions to health and illness. *Annual Review of Psychology* **50**: 137–63
- Belloc N B, Breslow L 1972 Relationship of physical health status and health practices. *Preventive Medicine* **9**: 409–421
- Black Report 1988 *Inequalities in Health: The Black Report and the Health Divide*. DHSS, London
- Blaxter M 1990 *Health and Lifestyles*. Routledge, London
- British Medical Journal 1982 *Alcohol Problems*. BMJ Publications, London
- Brubaker R G, Fowler C 1990 Encouraging college males to perform testicular self-examination: evaluation of a persuasive message based on the revised theory of action. *Journal of Applied Social Psychology* **20**(17): 1411–22
- Conner M, Norman P 1996 (eds.) *Predicting Health Behaviour*. Open University Press, Buckingham, UK
- Consensus Development Conference on Lowering Blood Cholesterol to Prevent Heart Disease 1985 Lowering blood cholesterol to prevent heart disease. *J. of the Australian Medical Association* **253**: 2080–6
- Doll R, Peto R, Wheatley K, Gray R, Sutherland I 1994 Mortality in relation to smoking: 40 years' observations on male British doctors. *British Medical Journal* **309**: 901–11
- Fife-Schaw C R, Breakwell G M 1992 Estimating sexual behaviour parameters in the light of AIDS: a review of recent UK studies of young people. *AIDS Care* **4**: 187–201
- Friedman G D, Dales L G, Ury H K 1979 Mortality in middle-aged smokers and non-smokers. *New England Journal of Medicine* **300**: 213–7
- General Household Survey 1989 OPCS, London
- General Household Survey 1992 OPCS, London
- General Household Survey 1993 OPCS, London
- General Household Survey 1994 OPCS, London
- Gerrard M, Gibbons F X, Bushman B J 1996 Relation between perceived vulnerability to HIV and precautionary sexual behavior. *Psychological Bulletin* **119**: 390–409
- Gochman D S (ed.) 1997 *Handbook of Health Behavior Research*. Plenum, New York, Vols. 1–4
- Godin G, Kok G 1996 The theory of planned behavior: a review of its applications to health-related behaviors. *American Journal of Health Promotion* **11**: 87–98
- Gollwitzer P M 1993 Goal achievement: the role of intentions. In: Stroebe W, Hewstone M (eds.) *European Review of Social Psychology*. Wiley, Chichester, UK, Vol. 4, pp. 141–85
- Health Promotion Authority for Wales 1990 *Health in Wales*. Health Promotion, Cardiff, UK
- Moore L, Smith C, Catford J 1994 Binge drinking: prevalence, patterns and policy. *Health Education Research* **9**: 497–505
- Nutbeam D, Catford J 1990 Modifiable risks for cardiovascular disease among general practitioners in Wales. *Public Health* **104**: 353–61
- Orbell S, Hodgkins S, Sheeran P 1997 Implementation intentions and the theory of planned behavior. *Personality and Social Psychology Bulletin* **23**: 945–54
- Prochaska J O, DiClemente C C, Norcross J C 1992 In search of how people change: applications to addictive behaviors. *American Psychologist* **47**: 1102–14
- Rigotti N A 1989 Cigarette smoking and body weight. *New England Journal of Medicine* **320**: 931–3
- Royal College of Physicians 1983 *Health or Smoking: Follow-Up Report*. Pitman, London
- Sampos C, Fulwood R, Haines C et al. 1989 The prevalence of high blood cholesterol levels among adults in the United

- States. *Journal of the American Medical Association* **262**: 45–52
- Schwarzer R, Fuchs R 1996 Self-efficacy and health behaviors. In: Conner M, Norman P (eds.) *Predicting Health Behaviour*. Open University Press, Buckingham, UK, pp. 163–196
- Sheeran P, Abraham S C S 1996 The health belief model. In: Conner M, Norman P (eds.) *Predicting Health Behaviour*. Open University Press, Buckingham, UK, pp. 23–61
- Stephens A, Wardle J 1992 Cognitive predictors of health behaviour in contrasting regions of Europe. *British Journal of Clinical Psychology* **31**: 485–502
- Strax P 1984 Mass screening for control of breast cancer. *Cancer* **53**: 665–70
- US Department of Health and Human Services 1988 *The Surgeon General's Report on Nutrition and Health*. US Department of Health and Human Services, Public Health Service, Washington, DC
- US Department of Health and Human Services 1990 *Health People 2000: National Health Promotion and Disease Prevention Objectives*. US Department of Health and Human Services, Public Health Service, Washington, DC
- Vernon S W, Laville E A, Jackson G L 1990 Participation in breast screening programs: a review. *Social Science and Medicine* **30**: 1107–18
- Waldron I 1988 Why do women live longer than men? *Journal of Human Stress* **2**: 2–13

M. Conner

Health Behaviors, Assessment of

1. Introduction: Health Behaviors as a Multidimensional Construct

In the field of health psychology, health behaviors are most often defined as behaviors that people engage in to maintain or improve their current health. These definitions focus on subjective health beliefs since they include any behaviors performed by a person in order to protect, promote, or maintain his or her health, whether or not such behaviors are objectively effective towards that end. In contrast, most empirical studies in public health and epidemiological research focus on behaviors which are defined by experts as having a substantial relationship to health. Therefore, the term health behavior has two different connotations. First, a subjective one which focuses on which behaviors lay people believe may affect their health, and second, an objective one which focuses on what experts evaluate and define as health-affecting behaviors. However, lay and expert views on which behaviors are related to health may not be consistent. In addition, several studies have found little consistency in people's health habits. Therefore, a person who exercises will not necessarily go for a regular check-up at the dentist. One reason people's current health habits are not more interdependent is the fact that health behaviors differ on a number of dimensions and may be

influenced by different factors. The modest link between health behaviors and the high variability in their diffusion raises the question of whether it makes sense to talk about health behavior as a unitary construct.

Furthermore, health behaviors could be examined as antecedents or causes of health status. Such an approach is taken by social epidemiologists. In contrast, health educators are interested in health behaviors as objective dimensions for systematic interventions directed at producing behavioral changes. Finally, as subjects of scientific study, health behaviors are analyzed as consequences or outcomes, as well as correlates of personal factors. For an accurate and reliable measurement of health behaviors it is essential to distinguish between their different dimensions and to define clearly the topic to be investigated.

2. Measurement of Health Behaviors

There are various methods to assess health behaviors. On rare occasions, physiological methods can be used. This is the most accurate way to measure alcohol consumption (via blood or urine sampling), drug consumption (via immunoassay, hair or sweat bioassay procedures), habitual dietary intakes (via biochemical markers), or physical activity (via doubly labeled water). However, such bioassay methods are only required when a high level of accuracy about recent health behavior is needed (e.g., for workplace drug testing) or used in addition to self-reported data to provide a means of confirming or falsifying self-reported information (e.g., about recent drug use). However, in some circumstances it may be appropriate only to make respondents believe that there is a way of finding out about their behaviors via physiological measures to reduce misreporting.

Other direct methods are behavioral observation techniques, which are used to assess physical activity among children or a driver's choice of speeds, to name two examples.

Another alternative for observing 'true' health behaviors is unobtrusive, seminaturalistic measures, such as video-based simulation which involves participants viewing a video film of various situations relating to the task at hand and measuring responses to that task (e.g., a video-simulation measure of drivers' speed choice). This method could provide ecologically valid measures which simultaneously satisfy the demands of internal and external validity. Another possibility for seminaturalistically measuring health behaviors has been proposed by Rothman et al. (1993), who offered participants the opportunity to order a free sample of suntan lotion and to indicate their preferred protection factor rather than asking for behavioral intention. However, these techniques can only be approximations of 'reality' because they are often confounded with other variables. In the study of

Rothman et al. (1993), one could argue that the participants ordered the suntan lotion because it was for free. In addition, behavioral observation techniques are often impractical for broader samples as well as for complex and habitual health behaviors. Therefore, self-reports are a crucial—and often the only—source of information about health behaviors. There are different kinds of self-reports, which can be distinguished roughly into qualitative and quantitative methods.

3. Qualitative Methods

Unstructured or semistructured interviews are useful qualitative techniques for researchers interested in understanding individuals' cognitive and conceptual models of health behavior and the frames of reference that individuals use to organize health behaviors. Therefore, qualitative methods are mainly concerned with exploration and analysis of health behavior because they allow the interviewee room to address the issues which he or she feels relevant to the topics being raised by the investigator. For instance, detailed interviews and ethnography could give insight into subjective and experiential dimensions of those behaviors related to substance misuse. One major disadvantage of qualitative methods is the fact that generality is, by definition, not quantifiable. Furthermore, since anonymity is not given, self-reports elicited by interviews may be affected by social desirability biases which lead to over-reporting of socially desirable behavior as well as under-reporting of socially undesirable behavior. The results from the National Household Survey of Drug Abuse (NHSDA) showed that the proportion of the respondents reporting marijuana use was higher when questions were self-administered than when they were administered by an interviewer.

Alternatively, one could introduce a 'reality check' on the respondent's account by using the *method of focus groups* (Morgan 1997). A focus group is a collective, discussion-based interview orchestrated by a moderator who uses the small group setting to achieve a group dynamic to construct a supportive social context that allows respondents to reveal thoughts and behaviors they would normally be reluctant to disclose. Hence, this method could be useful for generating information about behaviors that are rarely talked about (e.g., sexual behaviors and drug use).

4. Quantitative Methods

Stone and Shiffman (1994) have labeled strategies for collecting self-reports of respondents' momentary or current state as *Ecological Momentary Assessment*

(EMA). EMA studies usually consist of repeated assessment of participants' momentary state as they go about the tasks of daily life in their natural environment. There are several ways to approach this sampling task. *Interval-contingent* assessments require assessment at regular intervals. One example is the method of *interactive voice response* (IVR), where alcoholics are asked to call in on a regular basis to report their drinking to interviewers. Another way is asking subjects to record every episode of smoking, eating, or some other behavior of interest. This *event-contingent* approach may not lead to a representative sample of the participant's general state and requires a clear definition of the triggering event. In contrast, *signal-contingent* sampling supplies participants with an external signal cue, which is usually timed to be emitted at random, to prompt subjects to complete a written assessment or an electronic diary. As signal devices, beepers, electronic watches, and palmtop computers have been used. Investigations of temporal and contextual patterns of smoking have shown that global self-reports are largely inaccurate in comparison with a computerized EMA. Hence, EMA is a method which precisely assesses recent health behaviors. The major advantage of this method is that it minimizes reliance on recall from memory by relying on respondents' reports of their experience at the moment of the inquiry.

Diary log techniques are data collection strategies which gather information in relation to the passage of time. Therefore, the distinctive feature of this method is that it yields information which is temporally ordered. It tells the sequence of events and the profile of actions across time. Diary techniques can be particularly useful when data from the same person are required over a considerable period of time and/or very frequently, such as for assessing smoking behavior, alcohol consumption, or dietary habits to provide a general estimate of an individual's consumption. For example, alcohol consumption diaries often include questions about the frequency of consumption, the type of drink taken, and the quantity consumed on each occasion. In comparison to questionnaires, the diary log format minimizes recall biases associated with retrospective reporting, but daily reporting may be more reactive. In addition, diaries could be valuable for getting access to so-called 'intimate' information (e.g., sexual behavior).

Timeline Followback Method Reports (TLFB) have been developed by Sobell and Sobell (1995). They provide a detailed insight into health behaviors such as smoking, drug abuse, or drinking over a designated time period. Participants are required to provide retrospective estimates of their daily behavior by using a calendar over a certain time period, which can range up to 12 months prior to the interview. With this method, the pattern, variability, and level of drinking or smoking can be profiled, which is especially useful when relatively precise estimates are needed or when

researchers wish to evaluate specific changes in health behaviors before, during, and after interventions.

A *questionnaire*, which assesses past behavior frequency, is the most commonly used method to measure health behaviors. There are numerous questionnaires which ask for the average or typical quantity and frequency of alcohol use (see Sobell and Sobell 1995), dietary habits, or physical activity (see Pereira et al. 1997). However, the information provided by quantity and frequency measures (QF estimates) is limited by the fact that respondents must abstract an estimate from highly variable experiences. QF estimates often reflect less drinking and tend to misclassify drinkers compared with daily diary or timeline reports. They also provide lower absolute food intake estimates than a longer, interviewer-administered diet history.

5. Distortions in Self-reports of Health Behaviors

There are a series of methodological problems shared by all surveys relying on self-reports which could seriously decrease the internal and external validity of a survey, such as the ambiguity surrounding health behavior terminology. The term 'sexual partner,' for instance, may be interpreted in many different ways by the respondents, ranging from describing someone with whom one has experienced orgasm to someone with whom one has shared a bed while naked. In addition, the target population should be carefully selected because health behaviors have different meanings for adolescents and adults.

Furthermore, short-term fluctuations, for example in substance use, produced by environmental (e.g., social settings) and psychological (e.g., mood or stress) variables, may affect the psychometric properties of use measures. Hence, seasonal effects and short-term fluctuations may lead to superficial behavioral changes which could be misinterpreted by researchers as genuine changes.

Survey questions must be fair and unbiased to obtain valid answers. Asking leading questions, which suggest an expected answer, or loaded questions, which have emotionally laden connotations, can quickly invalidate a survey. In addition, by their very nature, questions about past behaviors assume accurate memory for events as well as willingness to report these to a researcher. However, respondents may not recall the actual events, employing instead various cognitive heuristics (rules of thumb) to estimate frequencies. This could result in certain biases. People use different strategies to answer frequency questions over different time spans. Episodic enumeration (recalling and counting individual incidents) is more likely to be used when shorter time spans are involved in frequency reports, whereas rate-based estimation (projecting the typical rate over the length of the recall period) is more likely to be used when longer time spans are involved. Reported behavioral

frequencies for a year are generally less than 12 times the equivalent frequencies for a month. This phenomenon has been coined as the *time span effect*. Use of different time spans over or within studies may lead to inconsistent or even misleading results.

Beside memory-based information, respondents may use context-based information to make frequency judgments. One context-based heuristic involves the use of information provided by response alternatives. For example, respondents may use the range of the response alternatives provided to them as frame of reference in estimating their own behavioral frequency.

Accurate and reliable measurements of health behaviors, especially drug usage and sexual activity patterns, have proven difficult to obtain because of social desirability influences. People's tendency to under-report smoking and to underestimate their alcohol consumption are two well-established facts. In addition, studies that focus on behavioral frequencies consistently yielded *illusory superiority*: people report lower frequencies of unhealthy behaviors and higher frequencies of healthy behaviors for themselves than for an average peer.

6. Conclusion

If high accuracy is needed, health behavior should be assessed by physiological measures or (if these are not available) by behavioral observation, EMA, or diary techniques. However, these assessment strategies are resource consuming and often not practicable for researchers in the field studying large samples or complex, habitual health behaviors. Although self-reports may have a decreased validity and reliability in comparison with these methods, due to memory or motivational biases, they can nevertheless provide very useful information for scientific studies of the consequences of health behaviors. To ascertain the accuracy of self-reports, researchers should consider different possible connotations of questions in different samples, and construct questions carefully in order to reduce autobiographical memory biases and social desirability influences.

See also: Adolescent Health and Health Behaviors; Health Behavior: Psychosocial Theories; Health Behaviors; Health Risk Appraisal and Optimistic Bias; Illness Behavior and Care Seeking

Bibliography

- Morgan D L 1997 *Focus Groups as Qualitative Research*, 2nd edn. Sage, Thousand Oaks, CA
- Pereira M A, FitzerGerald S J, Gregg E W, Joswiak M L, Ryan W J, Suminsk R R, Utter A C, Zmuda J M 1997 A collection of physical activity questionnaires for health-related research.

- Medicine and Science in Sports and Exercise* 29(6 Suppl.): S1–205
- Rothman A J, Salovey P, Antone C, Keough K, Drake M C 1993 The influence of message framing on intentions to perform health behaviors. *Journal of Experimental and Social Psychology* 29: 408–33
- Sobell L C, Sobell M B 1995 Alcohol consumption measures. In: Allen J P, Columbus M (eds.) *Assessing Alcohol Problems: A Guide for Clinicians and Researchers*. NIAAA Treatment Handbook Series 4. National Institutes of Health, Bethesda, MD, pp. 55–73
- Stone A A, Shiffman S 1994 Ecological momentary assessment (EMA) in behavioral medicine. *Annals of Behavioral Medicine* 16: 199–202

B. Renner

Health Care Delivery Services

Delivery systems for health care services are nested within health care systems, which are in turn nested within health systems. In fact, the topic of this article is best defined with reference to those more macrolevel systems, which are themselves addressed by other articles in the Health section. *Health care delivery services* consist of that segment of any society's *health system* which is specifically focused on the delivery of health care services. A *health system* encompasses a complex set of structural relationships among populations and institutions that have an impact on health, broadly defined. Further, as described by Ruggie (see *Health Care Systems: Comparative*), each country's *health care system* is defined as those specific macrostructures focused on financing, delivering, and evaluating health care, all of which are contained within a singular context of culture, history, politics, and economics.

This article will focus on defining and elaborating the dimensions of *health care delivery services* in a manner that is generally consistent with the literature on health services research, and comparable to discussions of 'health care delivery systems' in the United States and most developed nations. Commonly accepted definitions of health care delivery system emphasize the formal arrangements constructed to deliver health care services to specific populations, identified either by market, geographic location, type of service, or subpopulation characteristics (such as age or gender). Thus, this is a level of structure that is slightly more narrow than Ruggie's treatment of 'health care systems,' and represents a specific substructure within Frenk and Murray's definition of a 'health system.' Further, there are multiple types of health care delivery systems within any health system, depending on the type of health care service being delivered, whether preventative, diagnostic, or thera-

peutic, and the type of population or subpopulation being served.

This article will explore, in order: (a) the principle actors or components within health care delivery systems; (b) the variety of system types often present in highly developed health systems; (c) major characteristics of contemporary health care delivery systems; and (d) strategies for assessing health care delivery system performance. A number of recent trends in the health sector and among health care organizations will be examined, particularly since they highlight the difficulties in assessing delivery system performance in the twenty-first century, using standard measures of performance, most of which were developed during simpler times.

Given the brevity of this article, bibliographic emphasis will focus largely (although not exclusively) on US health care systems. Although the concepts to be explored in this article are applicable across most developed countries, the enormity of changes unfolding in many countries would require several chapters of dedicated comparative analysis. For example, economic and political turmoil in Eastern Europe is currently wreaking havoc with fragile public health and health care delivery systems. A prime example is Russia, where economic resources and basic medical supplies are scarce, government promises to cover basic health care go unfulfilled, and physicians must barter and beg to acquire needed medical equipment and drugs (Shishkin 1999). An 'underground' medical system has developed providing care for those few who can pay privately. Massive change also characterizes Latin American health care systems. In Chile free market reforms adopted during the Pinochet regime privatized much of what had been an extensive and effective government supported health care system. What remains now is a nonsystem with considerable problems of access (Reichard 1996). For more extensive comparative discussions of contemporary changes in health care systems, the reader is encouraged to consult recent work by Field (1989), Light (1999), or Day et al. (1998).

1. Components of Health Care Delivery Systems

The primary actors involved in delivering health care services are the members of the *health care workforce* (physicians and non-MD professionals, technicians, paraprofessionals; see *Health Professionals, Allied*), various *health care organizations* where care is either delivered or coordinated, and the agencies or firms which *finance* the delivery of services. Throughout the last half of the twentieth century, the medical and health division of labor has expanded dramatically, an expansion that has been fueled by the development of sophisticated medical technologies and specializations. Similarly, a variety of health care organizations deliver or coordinate care services along every step of the care continuum, and the number and type of

provider organizations mushroomed during the 1980s and 1990s. Diversification of health care organizational forms has both produced new organizational types (ambulatory surgery centers, women's health centers) and led to the adoption of new product lines by traditional organizational forms. For example, hospitals have diversified by adding a component of geriatric care, or vertically integrated into a comprehensive health care system, complete with ambulatory clinics and long-term care units (Fennell and Alexander 1993).

The *agencies and firms which finance health care service delivery* are also part of this system. In the US, a mix of actors, from both public and private sectors, play this role. Insurance carriers and the federal government have historically been important financiers of health care; more recently private sector corporations have been exerting considerable control over service delivery as major purchasers of health care for their employees. Toward the end of the 1990s the US health system came to embrace the concept of 'managed care,' an arrangement in which the financial partner (the health plan) becomes directly linked to both the providers of care and the client, through risk contracts. These arrangements have obfuscated the roles of the providers of care and the third-party payers of care, health insurance carriers or plans (Scott et al. 2000). Managed care arrangements contractually link providers to health plans, ensuring that an enrolled group of patients will seek care only from contracted (and somewhat controlled) providers. The providers are often paid a preset per capita rate for treating enrolled patients, and agree to 'gate-keep' or control the patient's use of services through utilization management strategies (see *Managed Care*). Large employers seek contracts with health plans that can promise low premiums and a variety of covered services; health plans compete for large employee groups to enroll as they are the most preferred, low-cost, healthy patient panels. Variations on this general strategy have resulted in a range of complex managed care organizational structures, in which the health plan is mediator between patients and providers, and between providers and payment. Managed care organizations combine provider and plan functions into one structure (Flood et al. 1998).

Although health care workers, organizations, and financing agencies constitute the major actors of health care delivery systems, there are a number of external factors which exert very important influences on these actors. Some influences are so important that it is sometimes difficult to separate the system actors from the external influences. Medical technology, for example, is both part of the environment of health care delivery and part of the health care organizations delivering the care. Continued expansion of medical diagnostic and treatment equipment, medications, biological agents, and devices has driven medical costs upward. The new technologies have created new types

of technical and patient care jobs and professions within the health delivery system, which in turn lead to more demand for those technologies (Osterweis 1996). Both medical professionals and patients still expect new technologies to be available, despite their certain impact on costs, and their sometimes uncertain effect on outcomes. The availability of new technologies is used as a primary marketing device by health care organizations to attract patients, professionals, and healthplan contracts.

Several other environmental forces influence the major delivery system actors, including changes in the age structure of the population, shifts from acute to chronic illnesses as the leading causes of morbidity and mortality, the explosion of information and telecommunications technologies, and the role of the state as both purchaser and regulator of health services delivery. Each of these factors has shaped delivery system structures and performances, and interact with system structures and processes so as to change expected outcomes and performances (Fennell and Flood 1998). Although a full discussion of their impact is beyond the scope of this article, the reader is encouraged to consult recent work by Scott and co-authors for a comprehensive case study of change in one local health care system. Using an institutional theory perspective, these authors examine both the material resources (technologies, population characteristics, economic shifts, market characteristics) and the institutional structures and expectations surrounding health care (regulatory changes, norms about care delivery, cultural beliefs) in the San Francisco Bay Area from 1945 through 1995. They conclude that material resources and institutional environments have interacted to change the local health system over time, with different factors more or less dominant during the historical eras of professional dominance, federal involvement in health systems, and the current era of market-oriented delivery. Scott et al. (2000) remind us that it is difficult to draw boundaries around systems in the process of change, as one set of changes begets other changes, which then churn into new combinations of forces, counterforces, and evolving structures (see also Light's discussion of countervailing powers, 1993). Further, change at the local level is influenced by changes in national policy as well as changes in social and cultural norms.

2. *Varieties of Health Care Delivery Systems*

Within any given health system there can be a variety of delivery systems, identified by the type of health service being delivered. For example, care systems can be defined by where in the continuum of care the service is located, such as primary care, secondary, tertiary, rehabilitative, and/or custodial care. Two frequently made dichotomous distinctions, based on the general type of care, are acute vs. chronic

care and mental vs. physical health systems. Additionally, care systems have been described on the basis of *where* that care is provided, and by which actors. Categorization by location used to be a fairly reliable and straightforward method. For example, complex surgical care was once always provided in the hospital. Ambulatory care meant that care happened in clinics and doctors' offices. Long-term care always occurred in nursing homes. Hospice care at the end of life was most often provided in freestanding hospice centers or by a home care team working with the family (Watt 1996).

More recently, however, those clear boundaries of care location have faded out of focus, and the sectors of care often blend into each other. What used to be considered inpatient care is now routinely delivered on an outpatient basis (many surgical procedures, for instance, and their adjuvant treatments), and significant rehabilitative care following an acute hospital episode is delegated to either the skilled nursing facility, the rehabilitation facility, or home with nursing visits. Early innovations in hospice care as an alternative setting for end of life care, which moved the dying patient away from the technically-oriented hospital setting, first developed in the United Kingdom in the late 1960s, and then spread to the US in the 1970s (Boling and Lynn 1998, Keay and Schonwetter 1998). More recently in the US, hospice services are increasingly provided by special units embedded within hospitals or nursing homes. Efforts to evaluate the quality of hospice care within such a broad range of care settings are, of course, made more complex given the inevitable clash between the goals of acute care settings and end-of-life care (Mor 1987).

Similarly, the traditional definitions of acute, sub-acute, or postacute care services can no longer be made on the basis of care setting. Hospitals are increasingly providing more long-term care related services, and nursing homes have been providing care for more acute and subacute patients. It has even become more difficult to clearly differentiate the type of care services provided in particular care settings on even the most basic of distinctions: medical vs. mental health care. The prevalence of depression, for example, among older patients, whether located in primary care settings, inpatient settings, or long-term care settings has come under closer scrutiny. Some estimates suggest that up to 65 percent of elderly patients in nursing homes suffer from depression. Whether those institutionalized elderly actually receive appropriate mental health diagnoses or mental health care, however, is not at all certain.

Diversified health systems market their superiority over other health care providers, of course, on the basis of the 'seamless delivery system,' where all needed services are available in one location or from one large, multiunit provider (Shortell et al. 1996). Theoretically, the diversified system should be able to ensure better coordination of care and a much

smoother transfer of information from one division to another. But whether or not this type of delivery system structure actually results in net gains for care quality or continuity has yet to be demonstrated in any conclusive fashion. In fact, whether diversified delivery systems actually result in shared or compatible patient record systems, thus, assuring coordination of information about the patient within the system, still needs to be examined. Recent work on the difficulties of guaranteeing security and privacy of electronic patient records revealed that the systems presumed to coordinate electronic information across health care organizations are much further behind in development than expected. When health care organizations merge or are acquired by corporate systems, the details of developing shared information systems through which patient care can be coordinated across care settings are only attended to as an afterthought or far later in the merger process (National Research Council 1997).

3. *Characteristics of Contemporary Health Care Delivery Systems*

Traditional studies of the structure and process of health care organizations have provided a model for the study of health care delivery systems. The basic *structural characteristics* of size, ownership, profit orientation, range of service provision, and primary governance structure provide a template for plotting variation in delivery systems. Similarly, *process characteristics* describe systems in terms of what is done to and for the patient over the course of treatment. Key variables of process include system coordination mechanisms, information flows, decision-making processes, quality control, and communication flows between practitioners and patients. In the past two decades, health care delivery systems have been characterized by three major trends, the first of which was mentioned earlier: (a) an increase in the diversification of structural types and products; (b) change in traditional ownership and management configurations; and (c) the development of new inter-organizational patterns of care delivery and complex multitiered governance structures.

Changes in ownership and management patterns of delivery systems and their component organizations over the past two decades have drastically changed how we think about the delivery of health care services. There have been basically four major facets of ownership/management reconfiguration, affecting different health care organizations and systems in various ways: (a) an increased concentration of the acute care delivery system, through hospital closures and mergers; (b) a rise in the prevalence of corporate restructuring across both acute and long-term care sectors, and the spread of managed care philosophy and organizational structures; (c) a blurring of the once basic distinction between profit and not-for-

profit ownership (see *Health Care Organizations: For-profit and Nonprofit*); and (d) a drastic and pervasive change in the management 'climate' of most health care organizations, from an emphasis on nonprofit service and community responsibility to an emphasis on competitive business strategies. These trends have been discussed at length elsewhere (see Bibliography).

The development of new interorganizational patterns of care delivery and more complicated governance structures is best exemplified by the growth in multihospital systems, loosely structured provider networks or coalitions, and formally structured ownership arrangements linking hospitals, physician groups, health plans, and other care settings (best exemplified by the proliferation of managed care organizations, as discussed above). Both horizontally and vertically integrated forms have made substantial inroads in most health care markets, and they vary by size, composition, and level of integration. The governance structures that have developed along with these complicated multiunit structures are generally described as business-oriented, no matter what their profit status. Governing boards within integrated delivery systems often take a two-level form: the separate units (such as a hospital) may retain an internal board, with at least some members drawn from the community. But these local boards tend to focus only on community fund raising and local public relations. Significant strategic planning and investment decisions are usually under the control of the 'corporate' board of the system.

4. Assessing Health Care Delivery System Performance

Two criteria for evaluating the performance of health care delivery systems and the organizations within them have dominated the health services literature: the quality of care received by patients (including their satisfaction with services) and the cost efficiency with which health care services are delivered. Quality of care can be evaluated by examining actual care outcomes (changes in health or functioning) or by benchmarking, that is, comparing actual performance or structures to state-of-the-art standards. Evaluations of care delivery focused on processes or structures assume, of course, that an excellent structure in place enhances the likelihood of high-quality processes being performed and, consequently, the likelihood of better outcomes occurring. These assumptions therefore rest on correctly identifying state-of-the-art standards or validating that their use does in fact lead to improved outcomes.

Fueled by physician profiles and public report cards, the 1990s have brought about an increased sophistication in measuring outcomes. Most health services researchers are also careful to control for the myriad of patient-level differences that can affect appropriate

accountability for care delivered within a single organization within the delivery system, such as adjusting for patient health factors while assessing outcomes following surgical care. Very sophisticated measures attempt to include posthospital outcomes to assess hospital care. Nonetheless, most of these outcome measures have been designed to focus on health statuses typical of young and middle-aged patients, and rarely do they focus on the most commonly seen age group of health service users—the elderly. Outcome measures are also most likely to be built around acute medical problems, and rarely are they modeled on the health problems of the elderly or of the chronically ill, who typically have several concurrent, complex, interrelated chronic diseases with the potential for acute episodes. Nor are they able to measure accurately the outcomes for people with interrelated mental and physical health problems, people who are facing end-of-life scenarios, or people for whom the goals of care delivery need to emphasize pain relief, palliation, or functional maintenance (Lamers 1990).

4.1 Problems with Outcome Measure Validity

Indeed, the convergence of several recent trends in health care delivery structures and environments (discussed briefly above) are of particular concern when assessments of service delivery system performance are considered: the graying population of most developed countries, shifts in the prevalence of acute and chronic diseases, the proliferation of new inter-organizational patterns of care delivery, and the spread of managed care. A variety of methodological and design issues need to be addressed in order to properly model and evaluate health care delivery system performance under this confluence of conditions. As an example, consider the situation of terminally ill patients: what outcomes should be used to evaluate hospices, which are designed to provide psychological, social, and spiritual care for dying persons and their families? Evaluations in the past have typically focused on the costs of care and the range of services provided (Mor 1987). Quality of hospice care is usually defined in terms of its effect on the quality of life for the patient receiving hospice services and is often 'customized' to reflect differences in each state's goals for the hospice program. But what truly constitutes a 'positive outcome' for dying patients? Whose care needs to be assessed—should we also consider the care of the dying patient's family? How should 'quality of life' be modeled? What is the set of 'costs' that needs to be included in evaluating the cost-effectiveness of hospices themselves or hospices compared to other sites where people die?

In addition to these issues regarding the proper 'outcomes' to measure, terminally ill patients receive their care from a wide range of providers and in a variety of settings. Despite the typical complexity of

care for any given patient, evaluations of the performance of health care delivery systems almost always have focused on a particular portion of that care attributable to a single provider or organization. Common examples include the Resident Assessment Instrument (RAI) for nursing homes (Morris et al. 1990), accreditation standards for hospitals, and NCQA standards for managed care organizations (Zinn and Mor 1998). How well, in fact, have these measures identified the portion of care that can be attributed to a single provider, when the true process of health care delivery involves multiple actors?

What is needed most is a reliable and valid measure of continuity of care, which should take into account the extent to which care is coordinated across care settings, care providers, and transitions between providers and settings. Even outcome measures based upon an illness episode (rather than a specific and standard unit of time, or service consumption event) do not necessarily capture care outside of a particular setting, such as when a hospitalization is the unit of analysis for an episode.

4.2 Problems in Defining the Unit/Level of Analysis for System Performance

Although there are obviously many possibilities for defining levels of systems of care delivery, it is conventional to identify at least three: (a) a single health care organization; (b) a larger, socially defined unit that contains the health care organization, such as a community or a system of health care organizations; and (c) the subunits contained within the organization, such as individual departments or practitioners. Researchers examining organizational issues typically inspect only one facet of care delivery at a time, or one level of analysis. All complex systems, however, tend to be nested units, that is, systems within systems within systems. While most analyses of delivery system performance attempt to bypass these issues by focusing on one or more of the standard three levels, they can instead fail to identify fully the processes and outcomes they are intended to study. The critical point is not that all nested levels must always be studied, but that one should be as clear as possible in identifying the level of analysis selected and about ensuring that the choice of level is aligned with the characteristics of the type of care delivered, the subpopulation of patients, and the goals of the delivery system under examination.

It is also important to recognize that system performance at any given level may not be analyzable as a simple aggregation of performance at lower levels. This is one of the principal features of any system: its performance is determined as much (if not more) by the arrangement of its parts, the types of coordination and integration achieved across units, as by the performance of the individual components. Specifying

the appropriate levels within the model, however, is not a simple matter, since multiple contexts can influence patient outcomes simultaneously and those effects may be difficult to distinguish due to nesting or overlapping. For example, a physician's clinical practice behavior can be nested within his/her clinic's setting, which itself may be nested within a particular diversified health care system, and a particular region or market with characteristic practice patterns.

5. Conclusion

Continuity of care, or clinical integration, has been recommended by Shortell and other researchers as the model on which managed care should be based, so that providers are explicitly recognized as able to bridge care settings along with the patient, so that information is coordinated and shared explicitly across care setting transitions, and so that the care manager may change with these transitions, even if the same actors are involved. Managed care at its best would embrace such a model, and by creating effective mechanisms such as case managers, real integration and coordination would result, presumably leading to better delivery outcomes. Similarly, the multi-institutional, multiprovider, diversified health care delivery system should, on the face of it, make the sharing of information a matter of routine, through uniform patient records and centralized record systems across care settings and providers, thus improving care continuity and a variety of outcomes.

This is the basis of the 'seamless delivery system.' But whether or not such changes in care structures actually result in net gains for care quality or continuity has itself yet to be demonstrated in any conclusive fashion. Understanding the complexity of service delivery systems, and the other systems within which they are embedded, will not simplify this task. But it will lead to better measurement and modeling, and thus, perhaps, to a more realistic assessment of what is needed to achieve 'seamless' system functioning.

See also: Caregiver Burden; Caregiving in Old Age; Gender and Health Care; Health Care: Legal Issues; Health Care Markets: Theory and Practice; Health Care Organizations: For-profit and Nonprofit; Health Care, Rationing of; Health Care Systems: Comparative; Health Care Technology; Illness Behavior and Care Seeking; Managed Care; Minority Access to Health Services: United States

Bibliography

Boling A, Lynn J 1998 Hospice: Current practice, future possibilities. *Hospice Journal* 13(1-2): 29-32

- Day P, Klein R, Miller F 1998 *Hurdles and Levers: A Comparative US-UK Study of Guidelines*. Nuffield Trust, London
- Fennell M, Alexander J 1993 Perspectives on organizational change in the US medical care sector. *Annual Review of Sociology* **19**: 89-112
- Fennell M L, Flood A B 1998 Key challenges in studying organizational issues in the delivery of health care to older Americans. *Health Services Research* **33**(2): 424-33
- Field M G (ed.) 1989 *Success and Crisis in National Health Systems: A Comparative Approach*. Routledge, New York
- Flood A B, Fremont A M, Jin K, Bott D M, Ding J, Parker R C Jr 1998 How do HMO's achieve savings? The effectiveness of one organization's strategies. *Health Services Research* **33**(1): 79-99
- Friedson E 1994 *Professionalism Reborn: Theory, Prophecy and Policy*. University of Chicago Press, Chicago
- Gold M R (ed.) 1998 *Contemporary Managed Care: Readings in Structure, Operations, and Public Policy*. Health Administration Press, Chicago
- Gray B H 1999 The changing face of health care. In: Clotfelter C F, Ehrlich T (eds.) *Philanthropy and the Nonprofit Sector in a Changing America*. Indiana University Press, Bloomington, IN
- Keay T J, Schonwetter R S 1998 Hospice care in the nursing home. *American Family Physician* **57**(3): 491-4
- Lamers W M Jr 1990 Hospice: Enhancing the quality of life. *Oncology-Huntington* **4**(5): 121-6
- Light D 1993 Countervailing Power: The changing character of the medical professions in the United States. In: Hafferty F W, McKinlay J B (eds.) *The Changing Medical Profession: An International Perspective*. Oxford University Press, Oxford
- Light D 1999 *Future Challenges for the NHS: An International Perspective on the 50th Anniversary*. Nuffield Trust, London
- Mor V 1987 *Hospice Care Systems: Structure, Process, Costs, and Outcomes*. Springer, New York
- Morris J N, Hawes C, Fries B E, Phillips C D, Mor V, Katz S, Murphy K, Drugovich M L, Friedlob A S 1990 Designing the National Resident Assessment Instrument for nursing homes. *Gerontologist* **30**: 293-307
- National Research Council: Committee on Maintaining Privacy and Security in Health Care Applications of the National Information Infrastructure 1997. *For the Record: Protecting Electronic Health Information*. National Academy Press, Washington, DC
- Osterweis M, McLaughlin C J, Manasse H R, Hopper C L (eds.) 1996 *The U.S. Health Workforce: Power, Politics, and Policy*. Association of Academic Health Centers, Washington, DC
- Reichard S 1996 Ideology drives health care reforms in Chile. *Journal of Public Health Policy* **17**(1): 80-98
- Shishkin S 1999 Problems of transition from tax-based system of health care finance to mandatory health insurance model in Russia. *Croatian Medical Journal* **40**(2): 195-201
- Scott W R, Ruef M, Mendel P J, Caronna C A (eds.) 2000 *Institutional Change and Health care Organizations: From Professional Dominance to Managed Care*. University of Chicago Press, Chicago
- Shortell S M, Gillies R R, Anderson D A, Erickson K M, Mitchell J B 1996 *Remaking Health Care in America: Building Organized Delivery Systems*. Jossey-Bass, San Francisco
- Watt K 1996 Hospice and the elderly: a changing perspective. *American Journal of Hospital Palliative Care* **13**(6): 47-8
- Wennberg J (ed.) 1998 *The Dartmouth Atlas of Health Care in the United States*. American Hospital Association, Chicago
- Wholey D R, Burns L R 2000 Tides of change: the evolution of managed care in the United States. In: Bird C, Conrad P, Fremont A (eds.) *Handbook of Medical Sociology*, 5th edn. Prentice-Hall, Upper Saddle River, NJ
- Zinn J S, Mor V 1998 Organizational structure and the delivery of primary care to older Americans. *Health Services Research* **33**(2): 354-80

M. L. Fennell

Health Care: Legal Issues

After years of relative stability, the relationship between law and health care in the USA is changing rapidly. These changes are both a response to, and a reflection of the transition to, a market-based health care system. How the courts and legislatures respond to these changes will, in part, determine the success of managed care's cost containment initiatives. This article broadly examines the relationship between law and health care, with a particular focus on how the courts have responded to challenges to managed care's cost containment innovations.

1. Law and Health Care: Framing the Issues

It should not be surprising that two of the USA's most important institutions, the legal and health care systems, are intertwined and intersect at various points. Since the early 1800s, medical jurisprudence, where physicians are called as expert witnesses and where the courts act to regulate the quality of medical care, has been an integral part of the growth of the health care enterprise. As Mohr noted, '... medical jurisprudence ... figured prominently in early American medical history ...; neither subject can be fully understood in its modern context without taking the history of legal medicine into account' (Mohr 1993, p. 251). As this quote indicates, the involvement of the courts in health care delivery is not new. Yet the relationship between law and health care has never been more visible or manifest than it is today.

Despite the inevitable ebbs and flows of the relationship between law and medicine, it is arguable that the relationship was essentially stable from the mid-nineteenth century until the mid-1960s (Stryker 1932). As recently as the late 1960s and early 1970s, health care attorneys and health care executives had a relatively easy set of legal rules to live with. Both sides could predict the nature and scope of the litigation since the essential rules establishing the standard of care and the types of litigation initiated varied little during this period. A typical court case involved one patient suing one physician, guided by liability rules that reflected judicial deference to physicians in setting the standard of care. That relative stability was largely shattered by the advent of managed care. At both the state and Federal levels, all three branches of the legal

system, the legislatures, regulatory agencies, and courts, have persistently been involved in setting health care policy and monitoring health care delivery since the mid-1960s. The framing of the legal issues in the managed care era thus reflects the dramatic changes engulfing the health care system since the mid-1980s and the resulting policy conflicts that have emerged.

1.1 The Litigation Context

For many years, courts have actively monitored quality of care through determining the standard of care for medical liability (Jacobson 1989, Furrow 1997) and interpreting contractual definitions of medical necessity (Eddy 1996, Hall and Anderson 1992). Courts have also been called upon to interpret the vast array of state and federal legislation regulating health care delivery.

In responding to litigation challenging managed care cost containment practices, judges face a threshold question of what role courts should play in monitoring the managed care environment relative to legislatures. Courts face the tension of applying common law liability principles to new situations (perhaps by depicting public policy considerations) or deferring to elected representatives to set public policy. In the new health care environment, courts will be asked to distinguish the economic aspects of managed care that order relationships between health plans, physicians, and patients, from the incentives that contribute to below-standard care. The former are more traditionally legislative prerogatives while the latter are traditionally within the judicial purview. For those cases resolved in the courts, judges will need to decide whether they should be resolved by contract or tort (civil wrongs such as negligence) law. The difference is that recovery under contract is more difficult and damage awards are lower. What complicates matters is that in managed care, the insurance (i.e., contract) and medical care (i.e., potential negligence) are combined in one entity.

1.2 Liability

Common law liability doctrine has traditionally emerged incrementally to adapt to changing circumstances. Thus, it will take some time before the actual contours of liability can be assessed. Nevertheless, certain themes have emerged in the litigation so far.

Despite variation in the state cases decided to date, the direction appears to be that courts are applying traditional liability doctrine to the new organizational forms, including health maintenance organizations (HMOs), independent practice associations (IPAs), and preferred provider organizations (PPOs) (Furrow 1997). For staff or group model HMOs, the law had

already developed holding them liable for torts (civil wrongs) committed by their employees. The primary issue is whether courts will apply indirect liability doctrine (i.e., vicarious liability and agency principles) to the other MCO organizational forms for malpractice committed by an independent physician.

In determining vicarious liability, the courts look to the MCO's control over the physician or how the plan markets its physicians. The greater the indications of control and the more the plan markets the quality of its physician panel, the greater the likelihood that the MCO will be held vicariously liable. In the leading case of *Boyd v. Albert Einstein Medical Center*, 547 A.2d 1229 (Pa. Super. 1988), for example, the court held that an IPA could be found liable for an individual physician's medical malpractice because the patient reasonably believed that the IPA controlled the physician. The court rejected the plan's argument that it was not liable because the physician was an independent contractor, holding that the physician could still be the plan's agent with respect to the patient. A key element is whether the MCO has sufficient control, such as through utilization management arrangements, to override a physician's clinical decision.

1.3 Challenges to Cost Containment Practices

The changes in the health care environment have inevitably resulted in a new set of issues to be litigated, particularly the conflict between population-based cost containment and access for individual subscribers (sometimes to experimental procedures), the multiplicity of actors in a given case, and the evolving nature of the organizational structures. Just as the courts had to adapt medical liability principles to the emergence of hospitals as central to health care delivery in the 1960s, the same process is now underway with regard to the emerging organizational forms comprising managed care. There is inevitably a learning curve as courts become educated about the underlying changes and begin to adopt principles derived for a different model to the new arrangements (Jacobson and Pomfret 1999). For example, instead of the simple model of litigation involving one physician and one patient, managed care litigation may include the patient, the physician, the health plan, the utilization review firm, and the administrator of an ERISA-covered plan. Sorting out which party is legally responsible for delayed or denied care and the acceptable level of care resulting from cost containment strategies remains a work in progress.

Nonetheless, the cases decided so far suggest that courts are not systematically impeding the implementation of cost containment initiatives (Jacobson 1999). While it may be that many of the most challenging cost containment innovations have yet to be litigated, courts seem willing to uphold cost containment programs, except, perhaps, if such programs are seen as

limiting access to life-sustaining treatment. For example, courts have generally rejected challenges to financial incentives (though some courts have allowed such challenges to proceed to trial), utilization management programs, and many contractual challenges to benefit denials.

Few courts seem willing to usurp legislative choices in formulating health policy or to obstruct the market in organizing and delivering health care services, with several courts specifically deflecting allegations that managed care incentives violate public policy to the legislatures. One reason for this is that courts have interpreted the Employee Retirement Income Security Act (ERISA) narrowly to block state court challenges to cost containment programs that deny benefit coverage. Another reason why cost containment initiatives have not been overturned is that courts are looking to contract law to determine the extent of the parties' obligations and responsibilities, especially in physicians' litigation against MCOs. Though the trend toward contract is neither unlimited nor uniform, it represents a potentially profound shift from the early 1970s, where courts were perceived as protectors of individuals, to courts as deferring to market-based changes.

1.4 Physician Autonomy

The initial litigation trends suggest that courts will not subordinate cost containment programs to physician autonomy. Courts are ruling that people who are dissatisfied with managed care, including physicians, should take their case to the legislatures. This represents a dramatic departure from the era when courts deferred to physician dominance.

It would be understandable if physicians expected that courts would protect the physician-patient relationship from interference by managed care plans. After all, the essence of the health care system lies with the physician-patient relationship, and courts have protected physician dominance in the past (Kapp 1985, Hall 1988). But courts have not done so in the managed care era. To the extent that physicians have been able to secure some gains through the courts, it has largely been through fair process protections and retaining the often-maligned legal standard of care.

The question is why courts are not deferring to physicians; what has changed over the past decade or so to cause the courts' shift from protecting physician dominance to reinforcing managed care's dominance? At this point, the answer is somewhat speculative because of the limited number of cases. One potential answer is that courts respond to changes in the underlying environment. As the nature of health care delivery has changed, it stands to reason that courts would begin to look at and incorporate those changes. Slowly but surely, courts are internalizing cost con-

tainment goals in developing legal doctrine, rather than ensuring the primacy of the physician-patient relationship. Since the dominant force in health care is institutional, it is not surprising that the legal dominance physicians maintained is now waning.

2. Future Directions and Challenges

During the next few years, the legal system faces four key challenges regarding health care: (a) how to allocate liability given the multiplicity of decision makers involved in managed care; (b) how to balance between the health care needs of individual patients and the cost containment goals of the health care system; (c) how to deal with the consequences of ERISA pre-emption, which now prevents state courts from resolving challenges to how cost containment goals are implemented; and (d) how can conflicting social policy goals (i.e., the cost/quality/access trade-offs) be reconciled.

2.1 Allocation of Liability and ERISA Pre-emption

The allocation of liability and ERISA pre-emption issues are both about imposing accountability on a system that has managed to shift legal accountability to physicians while maintaining the fiction of non-control. Either voluntarily or through legislation, MCOs should be held accountable for their actions.

2.2 Balancing

There is no easy solution to balancing the individual patient's access to health care with conserving the plan's assets for the entire patient population. One approach might be to articulate a set of fiduciary obligations that a plan must adhere to in balancing the two. The burden would be on the plan to justify denial of care to the individual patient by a preponderance of the scientific evidence and to demonstrate that the individual's medical care is not being subordinated to the patient population (Jacobson and Cahill 2000).

2.3 Social Policy Conflicts

The plethora of legislation has placed health care executives in somewhat of a bind. Some regulations, such as fraud and abuse, are specifically designed to reduce health care costs. Others, such as EMTALA, are designed to improve access regardless of cost. Taken alone, each individual law is defensible. Taken together, however, the social policy goals of cost containment, improving access to health care, and improving quality of care conflict across statutes and

regulations. Complying with all of the laws may be impossible, and may reflect incoherent social policy.

3. Implications for Other Countries

While other countries may not, for domestic social policy reasons, want to follow the managed care model, it seems likely that most countries will face the need to impose cost constraints on their national health care systems. These choices are already evident in the British and Canadian systems. What this brief survey of US health law trends suggests is that reconciling cost containment with various institutional objectives and interests will take careful planning to ensure that the changes are implemented fairly and equitably.

See also: Doctor–Patient Interaction: Psychosocial Aspects; Health Care Markets: Theory and Practice; Health Care Organizations: For-profit and Nonprofit; Health Care, Rationing of; Liability: Legal; Litigation; Litigation: Effects on Medical Research; Psychotherapy: Legal Issues (United States)

Bibliography

- Eddy D M 1996 Benefit language: Criteria that will improve quality while reducing costs. *Journal of the American Medical Association* **275**: 650–57
- Furrow B R 1997 Managed care organizations and patient injury: Rethinking liability. *Georgia Law Review* **31**: 419–509
- Hall M A 1988 Institutional control of physician behavior: Legal barriers to health care cost containment. *University of Pennsylvania Law Review* **137**: 431–536
- Hall M A, Anderson G F 1992 Health insurers' assessment of medical necessity. *University of Pennsylvania Law Review* **140**: 1637–712
- Jacobson P D 1989 Medical malpractice and the tort system. *Journal of the American Medical Association* **262**: 2230–7
- Jacobson P D 1999 Legal challenges to managed care cost containment programs: An initial assessment. *Health Affairs* **18**(4): 69–85
- Jacobson P D, Cahill M T 2000 Redefining fiduciary responsibilities in the managed care context. *American Journal of Law and Medicine* **26**(2 & 3): 155–74
- Jacobson P D, Pomfret S D 1999 Establishing new legal doctrine: A model of judicial response to industrial change. *University of Michigan Journal of Law Reform*
- Kapp M B 1985 Medicine and law: A symbiotic relationship? *American Journal of Medicine* **78**: 903–7
- Mohr J C 1993 *Doctors and the Law: Medical Jurisprudence in Nineteenth-century America*. Oxford University Press, New York
- Rosenbaum S, Frankford D M, Moore B, Borzi P 1999 Who should determine when health care is medically necessary? *New England Journal of Medicine* **340**: 229–332
- Stryker L P 1932 *Courts and Doctors*. Macmillan, New York

P. D. Jacobson

Health Care Markets: Theory and Practice

This article describes the origins, development, and proliferation of health care markets. It then reviews their performance, summarizes recent theory and research on their limitations, and concludes by suggesting new ways to conceptualize and research economic behavior in health care.

1. Origins

'Health care markets' is a relatively new term, framed by economical, structural, and social forces. It hardly existed 25 years ago but rose to prominence in response to the years of recession and high inflation ('stagflation') in the 1970s. The costs of social and welfare programs, such as health care, kept rising as revenues stagnated, prompting a 'crisis in the welfare state' (Esping-Andersen 1994). In addition, there were increasing complaints about government-run programs and about hospital-based specialists running up large bills without regard to the needs of payers or the population (Bengoa and Hunter 1991). In response, health economists recommended that payers (governments, employers, and insurers) apply the well-established principles and techniques of price competition theory to health care in order to constrain costs, increase efficiency, and make providers more responsive to the priorities of payers. This constituted a historical shift from professional dominance to buyer dominance among the countervailing powers of providers, purchasers, patients, and the state, over the control of power, costs, information, and the distribution of risks (Light 1995a).

1.1 Problems of Market Failure

Applying price competition theory to health care was unprecedented, for medicine had long been regarded by leading economists as lacking the basic qualities for markets (Arrow 1963, Fox 1979). The idea of health care as a commodity to be traded in a market was anathema to them. Clinical medicine is rife with uncertainties and too often contingent on diagnosis and response. Thus 'products' often cannot be defined, and property rights may be unclear. Physicians vary substantially in how they treat the same problem. Good market information is poor and highly asymmetrical. The absence of these requirements for beneficial competition can mean that competition is harmful to patients and society.

1.2 Managed Competition as a Possible Solution

These serious misfits between the attributes of clinical medicine and the requirements for competition led

Enthoven 1988) to develop the theory of managed competition. To overcome these forms of market failure, Enthoven emphasized competition between comprehensive health care plans and stipulated several requirements for a fair and beneficial market, such as universal health insurance or access, coverage of a common set of benefits, risk-adjusted contracts and good comparative information on quality, efficiency, and costs. He also held that health care markets had to be actively managed, because providers had proved so clever in manipulating or circumventing rules. This he called 'managed care' rather than 'regulated competition.'

Managed competition won critical acclaim and spawned international efforts to proselytize it as the 'only practical solution' to soaring costs, though the model had serious flaws. It assumed that some people (managers) can be trusted and others (physicians) cannot; it did not solve the problems of uncertainty or contingency but rather tucked them inside the health plans; and it led to oligopolies, which minimize price competition (Light 1995b). New businesses and a new industry developed to create health care markets and varieties of managed competition by monitoring services and bills, changing clinical behavior, and organizing providers into 'preferred provider organizations' and 'network HMOs.'

1.3 Proliferation and Variations

A signal event occurred when Prime Minister Margaret Thatcher decided in 1989 to transform the world's largest health care organization, the National Health Service, from an administered service to a market model of competitive contracts between purchasers and providers. Other governments, notably New Zealand and the Netherlands, followed suit. By the mid-1990s, however, most had pulled back from implementing markets in health care, as they came to realize that price competition could cause service dislocations, cost shifting, increased inequality, reduced choice, reduced care, and increased costs over government-run programs (Light 1997, *Social Science & Medicine*, special issue 2001). Some countries such as the United States did little to prevent adverse selection, service dilution, and cost shifting. In the meantime, important European variations on American models developed such as 'planned markets' and 'quasi-markets' (LeGrand and Bartlett 1993, Saltman and von Otter 1992). Behind them was the realization that markets themselves are never 'free' but rather structured and regulated to reflect a society's values and judgments about the limits of competition.

2. Performance Assessment

It has proven difficult to assess the performance of health care markets and different competitive schemes

because changes are introduced wholesale with no controls, or markets are introduced simultaneously with changes in coverage. American employers, for example, have simultaneously shifted costs to employees, increased forms of risk-rating and exclusions in policies, used policy churning and within-group underwriting, reduced the amount of service or reimbursement in areas like mental health, and limited access to specialty services and then credited the savings to 'competition.' Most stakeholders do not seem interested in well-functioning markets like the one developed by Enthoven. They want to save or make money, as the case may be. Reflecting a policy bandwagon more driven by conviction than scientific evaluation, the most systematic external analysis of research on US health care markets found that among hundreds of research articles, only a small percent systematically compared HMO performance to some alternate delivery model (Robinson and Steiner 1998). Evaluation of the market reforms in New Zealand also found the measures seriously flawed, efficiency no greater, health gain no greater, and fewer services leading to greater unmet need (Ashton 1999).

Operationally, the best tracking system in the United States reported that 'few private employers offer employees a choice of plans, give employees financial incentives to choose economical plans, or provide information so that employees can evaluate the quality of the care they receive—three hallmarks of managed competition' (Long and Marquis 1998, p. 1). The well-designed studies showed that managed care organizations reduced the high utilization rates and prices found in the United States, but there were no controls for better health status of enrollees which could explain lower utilization. Reinhardt (1996) summarized the American scene as employers hiring corporate health plans as 'bounty hunters' to reduce costs—no questions asked. Buyers have paid little attention to effects on quality. Zelman and Berenson (1998), two of the architects of managed care markets, concluded that the corporations took early large profits by selecting healthier and wealthier markets and patients, while compromising on quality. Now that most of the excess has been squeezed out of prices and utilization, it is unclear how corporate health plans will reward investors. Stung by widespread criticism and political backlash, competing plans have eased up on managing utilization to such a degree that they resemble the open choice of fee-for-service that originally prompted the development of health care markets and managed care organizations (Freudenheim 2000).

In other countries, Hsiao (1994) evaluated the impact of introducing health care markets in several Latin American and Asian systems and reported higher costs, reduced access, increased inequality, and profiteering. Evans' (1997, p. 427) international assessment concluded that '...greater reliance on the market is associated with inferior system perform-

ance—inequality, inefficiency, high cost, and public dissatisfaction...’ Stocker et al. (1999) reported that market-based reforms in Latin America are driven largely by American health insurance or other corporations who have taken the early profits out of American markets and now seek high profits in new countries by buying government support. Meantime, backlash to health care markets in the UK and New Zealand led to ousting the ruling party that installed health care markets, and substantially dismantling the health care markets in order to foster cross-sector partnerships and develop community-based programs that integrate public health with clinical services. Competitive markets are regarded as inherently anathema to community and public health.

A central challenge posed by health care markets is that they are costly to start up and operate. Good market information, contracting, monitoring, and managing together cost about 10 percent of revenues, so that comparable savings on clinical services must be achieved just to stay even. Another problem, related to what Hsiao found in several other systems, is that proper markets are often not set up with the necessary information and safeguards against inefficient but profitable strategies.

2.1 Contributions

In response to criticisms, market-oriented research has made a number of contributions. Health care markets evolved earliest and most extensively in the United States, as employers self-insured and began to ask new questions such as, ‘Which services are most cost-effective? Which providers have the lowest costs and prices?’ To answer such questions, new statistical techniques, methods, and theory were developed to measure effectiveness and efficiency, identify low-cost providers, estimate risk adjustments, and compensate for imperfect information (Folland et al. 1997). Accountability, performance, consumer satisfaction, and quality have become much more visible and have led to serious efforts to measure them. Health care market thinking has spawned a new and wide range of strategies for delivering more cost-effective care, even if those strategies do not themselves entail much market-like behavior. The major reports from WHO-Europe and the London School of Economics, for example, describe new developments in priority setting, contracting, capital formation, performance-related pay, decentralized management structures, and consumer involvement that would not have occurred without the radical challenge posed by market-like thinking (Saltman et al. 1998, Mossialos and LeGrand 1999).

2.2 Limitations of Health Care Markets

Restructuring clinical services into markets of corporations bent on growth and profit is likely to drive up

costs in the long run, abetted by the important natural advantages of sellers (providers). They diagnose, prescribe, and treat. They control the clinical record that determines charges. They also have the patient’s loyalty, if not dependency. A second structural problem in many health care markets is that each insurer or plan contracts with most of the providers in a given market in order to offer subscribers the widest choice possible. But that means no real competition is possible, because each provider in effect ‘competes’ against himself. This practice can lead to the third limitation, that there are so many plans and policies that no one can identify or evaluate real differences. In a sample of 2,277 patients, Grembowski et al. (2000) found there were 189 different provider organizations and 755 different policies. Behind so many policies and plans are high costs for designing, marketing, and managing them, and such numbers far exceed what one finds in real competitive markets.

At a micro-, empirical level, Rice (1998) has summarized studies documenting several practical problems. First, the preferences of patients and employers are not fixed but are affected by prior experience, current interactions, and advertising. This poses ‘a very big problem’ for researcher and competitors as well as for economic theory, as do problems posed by uncertainty and contingency. Second, consumers’ choices do not necessarily reveal their preferences and may be poor judges of them. Third, Rice observes, consumers have poor market information, and fourth, when they are given it, studies find they sometimes ignore the most relevant data. Finally, a majority of users are also unable to understand basic labels, instructions, and policies (Parker et al. 1999). What, then, actually takes places in health care markets, especially when ‘consumers’ do not consume or buy, even though standard texts characterize them that way (Follard et al. 1997, p. 21)?

3. Casting a Wider Research Net

Studies reviewed by Rice indicate there are variables not now being captured in economic models and research. These have been the focus of the burgeoning field of economic sociology, which attends to the influence of affiliations, networks, institutions, and culture, as well as to the role of power, hierarchy, and the past in shaping ‘path-dependent’ behavior (Smelser and Swedberg 1994). In a classic article, Granovetter (1985, p. 481) showed how economic behavior and institutions ‘are so constrained by ongoing social relations that to construe them as independent is a grievous misunderstanding.’ What passes for ‘rational decision-making’ differs substantially from one nation to another (Dobbin 1994) and even from one corporation to another.

Economic behavior begins with cognitive and cultural categories that get combined into scripts which actors routinely use in deciding what they value, buy, or sell (DiMaggio 1997). These scripts allow actors whose capacity for rationality is bounded by their ability to filter, sort, and apply what they know through 'selective attention' (Lindenberg 1998). The sociology of microeconomics, or actual market behavior closely observed, looks more like a negotiated construction of a reality that the parties can live with, even if front-stage behavior conforms to the scripts and symbols of market competition (Hughes et al. 1997). Institutionalized routines, once established, define paths which others follow, making future actions 'path dependent.' Most industries (and certainly health care) are 'oligopolies guarding carefully differentiated niches through strategies that aim to preserve rents by avoiding ruinous competition' (DiMaggio 1998, p. 700).

Key dimensions for analyzing economic interactions in health care are: (a) the relative frequency of transactions; (b) their relative complexity; (c) their relative specificity; (d) their relative size and risk; (e) their relative quality of information; and (f) the degree of social relationship over time. Each of these continua must be assessed in terms of each major party affected. Sociological dimensions of the relationships between parties in an economic transaction include: (g) the degree of autonomy vs. mutual obligations between the; (h) the degree of tightness or looseness in the network or organization; (i) the size of the organizations or networks; and (j) the structure of power relationships. This approach is likely to deepen our understanding of health care markets in the future.

At a second level, the concept of 'health care markets' and its recent rise to dominance in certain countries, but not others, needs explaining. In a historical study of the 'conceptions of control' that prevail in a given industry from one period to another, Fligstein (1990) has described how a crisis in an industry prompts its leaders to construct or select a model of their business and its realities that will provide them with a coherent account of what went wrong and how to succeed in the future. The theories of price competition and managed competition in the late 1970s and 1980s seemed to have been embraced by payers for these reasons. Fligstein notes that the prevailing conception is interpreted as suits individual leaders, an observation that conforms to the selective and partial ways in which corporations and governments implement theories for how to have cost-effective models of health care markets. As a result, health care markets has become an international business and policy movement, spurred by international consulting firms, even as some countries have spurned them. The World Bank, the International Monetary Fund, and other major lenders began to require, or strongly urge, scores of developing countries to restructure their services into health care

markets as a condition for receiving financial help with their economies (World Bank 1993, Stocker et al. 1999).

Selective and partial implementation probably explains why managed care and managed competition have not resulted in greater efficiency or better services in some countries and why they have held cost increases at bay in the United States for only a few years. As the crisis of sharply increasing costs during the 1980s resumed in 2000, employers were concluding that health care markets as a 'conception of control' could not hold down costs and were losing faith in it (Freudenheim 2000). Like leaders in some other countries, they may start to search for a new conception of control, one that will explain why the health care markets of the previous 20 years were only partially successful and how to manage costs more effectively.

See also: Health Care Delivery Services; Health Care: Legal Issues; Health Care Organizations: For-profit and Nonprofit; Health Care, Rationing of; Health Care Systems: Comparative; Health Care Technology; Health Insurance: Economic and Risk Aspects; Managed Care; Minority Access to Health Services; United States; National Health Care and Insurance Systems; Pecuniary Issues in Medical Services: Ethical Aspects; Psychiatric Care, Economics of; Services Marketing; Socioeconomic Status and Health Care

Bibliography

- Arrow K 1963 Uncertainty and the welfare economics of medical care. *American Economic Review* 53: 941-73
- Ashton 1999 The Health reforms: To market and back? In: Boston J, Dalziel P, St. John S (eds.) *Redesigning the Welfare State in New Zealand: Problems, Policies and Prospects*. Oxford University Press, Auckland, pp. 134-53
- Bengoa R, Hunter D J (eds.) 1991 *New Directions in Managing Health Care*. World Health Organization Regional Office for Europe, Copenhagen
- DiMaggio P 1997 Culture and cognition. *Annual Review of Sociology* 23: 263-87
- DiMaggio P 1998 The new institutionalism: Avenues of collaboration. *Journal of Institutional and Theoretical Economics* 154: 696-705
- Dobbin F 1994 *Forging Industrial Policy*. Cambridge University Press, New York
- Enthoven A 1988 *Theory and Practice of Managed Competition in Health Care Finance*. North-Holland, Amsterdam
- Esping-Andersen G 1994 *After the Golden Age: The Future of the Welfare State in the New Global Order*. United Nations Research Institute for Social Development, Geneva
- Evans R G 1997 Going for the gold. *Journal of Health Politics, Policy and Law* 22: 427-65
- Fligstein N 1990 *The Transformation of Corporate Control*. Harvard University Press, Cambridge, MA
- Folland S, Goodman A C, Stano M 1997 *The Economics of Health and Health Care*. Prentice-Hall, Upper Saddle River, NJ

- Fox D M 1979 *Economists and Health Care: From Reform to Relativism*. Prodist, New York
- Freudenheim M 2000 H.M.O. costs spur employers to shift plans. *New York Times* 6 September: A1
- Granovetter M 1985 Economic action and social structure: The problem of embeddedness. *American Journal of Sociology* **91**: 481–510
- Grembowski D E, Diehr P, Novak L C et al 2000 Measuring the 'managedness' and covered benefits of health plans. *Health Services Research* **35**: 707–34
- Hsiao W C 1994 'Marketization'—the illusory magic pill. *Health Economics* **3**: 351–57
- Hughes D, Griffiths L, McHale J V 1997 Do quasi-markets evolve? Institutional analysis and the NHS. *Cambridge Journal of Economics* **21**: 259–76
- LeGrand J, Barrett W (eds.) 1993 *Quasi-Markets and Social Policy*. Macmillan, London
- Light D W 1995a Countervailing powers: A framework for professions in transition. In: Johnson T, Larkin G, Saks M (eds.) *Health Professions and the State in Europe*. Routledge, London pp. 25–41
- Light D W 1995b *Homo Economicus*: Escaping the traps of managed competition. *European Journal of Public Health* **5**: 145–54
- Light D W 1997 From managed competition to managed cooperation: Theory and lessons from the British experience. *Milbank Quarterly* **75**: 297–341
- Lindenberg S 1998 The cognitive turn in institutional analysis: Beyond NIE and NIS? *Journal Institutional and Theoretical Economics* **154**: 716–27
- Long S H, Marquis M S 1998 How widespread is managed competition? *Health System Change Data Bulletin*. Center for Studying Health System Change, Washington, DC
- Mossialos E, LeGrand J (eds.) 1999 *Health Care and Cost Containment in the European Union*. Ashgate, Aldershot, UK
- Parker R M, Williams M V, Weiss B D 1999 Health literacy: Report of the Council on Scientific Affairs. *Journal of the American Medical Association* **281**: 552–7
- Reinhardt U E 1996 A social contract for 21st century health care: three-tier health care with bounty hunting. *Health Economics* **5**(6): 479–99
- Rice T 1998 *The Economics of Health Reconsidered*. Health Administration Press, Chicago
- Robinson R, Steiner A 1998 *Managed Health Care*. Open University Press, Buckingham, UK
- Saltman R E, Figueras J, Sakellariades C (eds.) 1998 *Critical Challenges for Health Care Reform in Europe*. Open University Press, Buckingham, UK
- Saltman R E, vonOtter C 1992 *Planned Markets and Public Competition: Strategic Reform in Northern European Health Systems*. Open University Press, Buckingham, UK
- Smelser N J, Swedberg R (eds.) 1994 *The Handbook of Economic Sociology*. Princeton University Press, Princeton, NJ
- Social Science and Medicine* 2001 **52**(8) (special issue)
- Stocker K, Waitzkin H, Iriat C 1999 The exportation of managed care to Latin America. *New England Journal of Medicine* **340**: 1131–6
- World Bank 1993 *World Development Report 1993: Investing in Health*. Oxford University Press, Washington, DC
- Zelman W A, Berenson R A 1998 *The Managed Care Blues & How to Cure Them*. Georgetown University Press, Washington, DC

D. W. Light

Health Care Organizations: For-profit and Nonprofit

This article must begin with an explanation of the meaning and boundaries of its subject. First, its focus is on organizations, not on the motivation of individuals. Second, it pertains to two ownership categories that by convention, partly rooted in the US experience, stand in distinction to public or governmental ownership. That is, this article pertains to non-governmental organizations, which may be established as for-profits or as nonprofits.

Third, it pertains to a legal characteristic of organizations that is defined under the law of the jurisdiction in which they exist. In the USA, health care organizations are typically incorporated under state laws. Variations among these state laws are not important for our purposes, but the fact that fundamental aspects of the ownership and control of health care organizations (as well as the roles they play in the health care system) are defined by the laws and traditions that are specific to particular countries is. A recent effort to measure and compare the size of the nonprofit sectors in different countries had to define 'nonprofits' in terms of certain organizational characteristics (not profit-distributing, self-governing, voluntary) rather than legal form, because nations do not make the same legal distinctions between private and governmental or between for-profit and nonprofit (Salamon et al. 1999). To cite an example from health care, ownership by identifiable individuals is incompatible with the nonprofit form under US law, while Japan, where most hospitals are owned by physicians, is often said to prohibit for-profit ownership of hospitals because of a prohibition of ownership by corporations that pay dividends (Rodwin and Okamoto 2000).

Fourth, this article pertains primarily to types of health care organizations in which a mix of for-profit and nonprofit ownership is found. Thus, attention will not be directed to organizations that manufacture pharmaceuticals or hospital supplies (which are typically all for-profit corporations), to physicians' practices (which are seldom incorporated as nonprofit organizations), or to charitable fundraising organizations (such as the American Cancer Society or the Royal Commonwealth Society for the Blind), which are all nonprofit. Our focus is on the types of organizations among which both for-profits and nonprofits can be found, as is the case with hospitals, nursing homes, health maintenance organizations, and home care agencies in the USA.

Most literature on for-profit and nonprofit health care comes from the USA, which has an unusually pluralistic health care system on both the provider and purchaser sides (Starr 1982). Although the health care systems of other countries are much more typified by governmental or quasigovernmental ownership, the private ownership of important components, such as

hospitals, is not peculiar to the USA. In some countries there are doctor-owned clinics, small hospitals, or nursing facilities that are outgrowths or extensions of private physicians' offices. Some countries that have universal health insurance systems nonetheless have private hospitals. These countries may either prohibit for-profit ownership, as does Canada, or they may exclude private hospitals (including for-profits) from participation in the national health insurance system, while permitting for-profit hospitals to serve people who have the desire and wealth to opt out of that system. Several investor-owned hospital companies from the USA also own hospitals in other countries, serving such niches. And at least one German hospital company owns some US hospitals. In recent years, American companies have begun to export versions of managed care to other countries, notably in Latin America (Stocker et al. 1999).

In the main, however, in Western countries other than the USA for-profit organizations that provide health services are extensions of doctors' practices or constitute a sector that serves a small market outside of the governmental health system (and is often not included in statistical information about that system). This, plus the nation-specific nature of the laws that define the legal form of health care organizations, probably explains the paucity of research comparing for-profit/nonprofit organizations in countries other than the USA or making cross-national comparisons of ownership forms. Policy debates about privatization have stimulated research on public versus private provision of services in a number of countries, but these are outside the scope of this article.

1. The Nonprofit/For-profit Distinction

In this article, the terms 'for-profit' and 'nonprofit' (or its less economical synonym, 'not-for-profit') are used as they are commonly defined under state statutes and federal tax law in the USA. The term 'nonprofit' invites confusion, since nonprofit organizations are not forbidden from generating surpluses of revenues over expenses. Indeed, such surpluses, which are called profits in the for-profit world, are essential to economic health and survival of nonprofit organizations. It is important, therefore, to be clear about the shared characteristics of for-profit and nonprofit organizations and the ways that they differ.

Table 1 shows that there are important similarities and differences between for-profit and nonprofit health care organizations in the USA (Gray 1992). Two key families of legally defined characteristics are particularly important in distinguishing these organizations. The first pertains to ownership, governance, and purpose. For-profit organizations have owners (proprietors, partners, or the shareholders of corporations), who share in the profits and increases (or decreases) in the value of the enterprise and who, in the corporate form, elect the board of directors. Those who operate the for-profit organization are thus either owners or are managers who are directly accountable to owners. Thus, those who control the organization have a direct stake in the financial performance of the organization.

By contrast, nonprofit organizations do not have owners (except insofar as they are subsidiaries of other nonprofit organizations) and are legally prohibited

Table 1

Basic legal and economic differences between for-profit and nonprofit hospitals

For-profit	Nonprofit
Corporations owned by investors	Corporations without owners or owned by 'members'
<i>Purpose:</i> Management has legal obligation to promote wealth of shareholders within the boundaries of law; does so by providing services	<i>Purpose:</i> Has legal obligation to fulfill a stated mission (provide services, education, research, etc.); must maintain economic viability to do so
Can distribute some proportion of profits (net revenues less expenses) to owners as dividends; owners also share in increases in the value of the enterprise	Cannot distribute surplus (net revenues less expenses) to those who control the organization; there are no owners entitled to proceeds if organization is sold
Management ultimately accountable to owners (stockholders) through boards of directors elected by stockholders	Management accountable to voluntary, often self-perpetuating boards of trustees
Sources of capital include:	Sources of capital include:
Equity capital from investors	Charitable contributions
Debt	Debt (generally tax-exempt)
Retained earnings (plus depreciation and deferred taxes)	Retained earnings (plus depreciation)
Revenues derived from sale of services	Government grants
	Revenues derived from sale of services and from charitable contributions
Pay property, sales, and income taxes	Generally exempt from most taxes

from distributing surplus revenues to those who control the organization—a characteristic that Hansmann (1980) called the nondistribution constraint. Nonprofit organizations are governed by boards of trustees whose members are typically chosen by the existing board and who are responsible for assuring that the organization operates soundly and adheres to the mission (charitable, educational, religious, etc.) for which it was chartered. Importantly, the boards to which management is accountable have no direct stake in the profitability of the institution.

The second family of differences pertains to the tax treatment of for-profit and nonprofit organizations (Simon 1987). Nonprofit organizations that meet qualifying standards have a variety of tax-related benefits under federal, state, and local laws. The tax treatment of nonprofits varies according to the kind of organization it is (for example, there are differences in how churches, grantmaking foundations, charitable organizations, and professional associations are treated) and according to governmental jurisdiction (criteria for exemption from local property taxes may differ from criteria for exemption from federal corporate income tax). However, most nonprofit health care organizations are exempt as charities from federal corporate income tax under Section 501(c)(3) of the Internal Revenue Service Code. Charity in this case is defined in terms of benefiting the community at large, not in terms of providing free care (Fox and Schaffer 1991). (Some health maintenance organizations and insurers are exempt under Section 501(c)(4) as social welfare organizations and receive less-favorable tax treatment than do the charitable organizations.) Other tax benefits also apply to organizations that are exempt in the eyes of the Internal Revenue Service. Charitable nonprofits are generally able to receive contributions that are tax deductible to the donor, to borrow at favorable rates under bond issues whose interest is not taxable to the lender, and to be exempt from local and state property taxes and certain excise taxes.

However, in health care there are important financial similarities between for-profits and nonprofits. First, the revenues of both for-profit and nonprofit health care organizations come overwhelmingly from the sale of services. For most nonprofit health care organizations, charitable contributions represent no more than one percent of revenues. To survive, these nonprofit organizations must compete successfully for business in a very crowded and heavily regulated marketplace. Though not businesses, they must operate in a very businesslike fashion.

Second, although nonprofit hospitals and HMOs have benefited in earlier periods from governmental grant and lending programs, today's nonprofits and for-profits depend upon similar sources of capital—primarily retained earnings and debt (Institute of Medicine 1986). Thus, just as with for-profit organizations, the primary capital sources on which nonprofit health care organizations depend are a function of the

organization's economic performance. The fact that the typical nonprofit health care organization is tax exempt as a charitable organization under the Internal Revenue Service code suggests that they are subject to complex and somewhat conflicting social expectations.

Notwithstanding the importance of their common reliance on retained earnings and debt, for-profit and nonprofit organizations each have access to unique sources of capital. Unlike nonprofits, for-profits have the paid-in equity capital of individual and organizational investors or stockholders. Access to equity capital provides a substantial advantage during periods of growth, particularly for publicly traded firms whose stock price in relationship to earnings may be very high. The unique form of capital to which nonprofits have access is charitable contributions, as well as debt financing from instruments whose interest payments are not taxable for the lender.

These differences in sources of capital are enormously important regarding the expectations placed on these organizations—with stockholders seeking value, economic returns, and capital growth, and contributors seeking charitable behavior. The two families of differences have driven theoretical and empirical work in the social sciences (Powell 1987, Powell and Clemens 1998) and stimulated an array of public policy debates.

2. The Role and Performance of Nonprofit Organizations

The question of why some segments of our capitalist economy are predominantly for-profit while others are predominantly nonprofit has intrigued social theorists and economists, who generally start with the premise that the incentives internal to the for-profit structure create a drive toward efficiency. So, what is the role of nonprofits? Several answers have been offered. Weisbrod (1990) suggests that nonprofits occupy a niche created by, on the one hand, the failure of for-profit organizations to provide services that benefit the community at large, not just the purchasers, and, on the other hand, the failure of governmental programs to meet the demand for other than average services. Hansmann (1980) posits that the nonprofit form provides a comparatively efficient alternative in situations of contract failure in which purchasers are not well situated to monitor the performance of sellers of services. The nondistribution constraint makes nonprofits less likely than for-profits to exploit the purchaser's vulnerability. For services to be financed by donations, the nonprofit form gives potential donors some confidence that their funds will be used for the promised purpose (e.g., for charitable services). For services that are difficult for purchasers to monitor (and medical care is an example), the nonprofit form may provide some assurance that vulnerabilities of patients and payers will not be exploited.

Such theories suggest three directions in assessing the performance of nonprofit organizations—their costliness (in the absence of accountability to owners, are they less efficient?), their charitability (very important for tax-exempt organizations in a health care system with more than 40 million uninsured patients), and their effectiveness or trustworthiness as providers of services/care. The empirical literature that has attempted to assess these topics in different types of health care organization is large, growing, and impossible to summarize adequately in a brief article. Citations to most studies can be found, however, in the Institute of Medicine (1986) report on for-profit and a 1999 literature review that is available on-line (Gray et al. 1999) and that summarizes some 75 studies.

First, in health care, for-profit organizations are not necessarily less costly to purchasers than are nonprofits. Most studies of hospitals have found that costs to purchasers are higher in for-profits (Gray et al. 1999). This appears to be a response to 'reimbursement maximization' strategies based on exploiting the weaknesses of third-party payment systems, not to some inherent inefficiency disadvantage of for-profits vis-à-vis nonprofits. For example, several studies have shown that expenses incurred in the production of hospital services are similar in for-profit and nonprofit hospitals, but that markups of price over expenses are higher in for-profits.

Second, taken as a group, nonprofits are more likely to engage in charitable activities and public goods than are for-profits (Gray 1992, Claxton et al. 1997, Gray et al. 1999). For example, nonprofits care for more uninsured patients (Frank et al. 1990), provide more uncompensated care (Lewin et al. 1988), provide more services focused on vulnerable populations (LeBlanc and Hurley 1995, Shortell et al. 1987), and engage in education and research (Institute of Medicine 1986), and they are less likely to discourage admission of uninsured patients (Schlesinger et al. 1987) and less likely to avoid location in areas with large numbers of the uninsured (Norton and Staiger 1994). However, the charitable performance of nonprofits is quite variable, with charitable activities and public goods heavily concentrated in urban teaching hospitals. The performance of nonprofit hospitals as charitable organizations has come under criticism in recent decades, and in the 1990s eight US states established requirements for nonprofit hospitals to report on their community benefit activities. There is evidence that the charitable performance of nonprofits is enhanced by articulated community expectations and active public oversight (Schlesinger 1998).

Third, although the Institute of Medicine (1986) study found no clear pattern of ownership-related quality differences among hospitals, several more recent studies have found a quality advantage for nonprofits regarding hospital mortality (Hartz 1989, Kuhn et al. 1994, Sloan et al. 2000) and post-hospital recovery (Sloan et al. 2000). Quality advantages have

been reported among nonprofit nursing homes (Hawes and Phillips 1986) and HMOs (Himmelstein et al. 1999), two fields characterized by serious contract-failure problems. In nursing homes, many residents are unable to look after their own best interests and have no one else to do so and so must rely on their notoriously difficult-to-monitor relationship with the facility to protect their interests. In HMOs, the complexity of the incentive relationship between health plans and physicians may make it difficult for patients to be certain that contracted-for services are actually being provided.

At a more general level, the research findings to date suggest that for-profits are indeed more responsive to economic incentives than are nonprofits, which can be a problem in a field with significant information asymmetries between sellers and purchasers of service (Arrow 1963). In view of the similarities of for-profits and nonprofits in revenue capital and revenue sources and the regulatory regimes under which they must operate, the overall pattern of performance differences that are often small but generally in the predicted direction is notable. It is clear that the comparative performance of nonprofits and for-profits is not a simple function of ownership form, but of an interaction of ownership form with market forces, payment systems, and regulatory environments.

3. Ownership Trends in Health Care

The concentration of different forms of ownership varies in the USA across fields (nursing homes are predominantly for-profit; hospitals are mostly nonprofit) and over time (Marmor et al. 1987, Gray 1999). Among acute care community hospitals, surveys conducted early in the twentieth century reported that about half of all hospitals were for-profit (Stevens 1989, Starr 1982). The for-profit share of hospitals declined through the century, falling to 15 percent in 1965, with nonprofits comprising 60 percent and the remainder governmental—a level that has remained roughly the same, with a slight increase in for-profit and a slight decrease in governmental. However, the character of the for-profits has changed since the mid-1960s, with ownership by large, powerful, publicly traded investor-owned firms replacing ownership by local entrepreneurs (often doctors) as the predominant for-profit form, and with the average for-profit facility being larger as measured by the number of beds.

There are, however, a number of signs that the US health care system is moving toward an increased for-profit character. Even among hospitals, where the nonprofit share has been stable at 60 percent for 30 years, a shift toward for-profit has occurred beneath the surface. First, if beds rather than hospitals is used as the measure, the for-profit share of acute-care hospitals doubled between the mid-1960s and the mid-1990s (from 6 percent to 12 percent, largely at the

Table 2
Health care organizations by form of ownership, USA 1985 and 1997

Provider type	1985		1997		% of change > 85–> 97
Community hospitals ^{a,b}					
For-profit	805	14%	797	16%	– 1
Nonprofit	3,349	58%	3,000	59%	– 10
Government	1,578	28%	1,260	25%	– 20
Private psychiatric hospitals ^{a,c}					
For-profit	149	64%	233	77%	56
Nonprofit	84	36%	69	23%	– 18
Private rehabilitation hospitals ^{a,c}					
For-profit	11	18%	91	58%	727
Nonprofit	51	82%	65	42%	27
Home health agencies ^d					
For-profit	2,055	36%	6,290	67%	206
Nonprofit	3,653	64%	3,123	33%	– 15
Nursing homes ^{e,f}					
For-profit	14,300	75%	11,250	65%	– 21
Nonprofit	3,800	20%	4,792	28%	26
Government	1,000	5%	1,134	7%	13
Dialysis centers ^g					
For-profit	616	42%	2,322	68%	277
Nonprofit	847	58%	1,101	32%	30
HMOs ^{h,i}					
For-profit	137	35%	481	74%	251
Nonprofit	256	65%	167	26%	– 35

a American Hospital Association (1986) American Hospital Association Hospital Statistics, Chicago. b American Hospital Association (1999) American Hospital Association Hospital Statistics, Chicago. c American Hospital Association (1998) American Hospital Association Annual Survey of Hospitals, Chicago. d Health Care Financing Administration (1998) 1998 Profile of Medicare Chart Book, Rockville. e National Center for Health Statistics (1995) NCHS National Nursing Home Survey, Rockville. f National Center for Health Statistics (1997) NCHS National Nursing Home Survey, Rockville. g HCFA Office of Standards Planning and Quality (1999) National Listing of Medicare Providers Furnishing Kidney Dialysis Transplant Services, Baltimore. h InterStudy Publications (1985) InterStudy Guide to National Firms, Minneapolis. i InterStudy Publications (1999) InterStudy Competitive Edge 8.1, Minneapolis

expense of public hospitals). Second, even with hospitals as the unit of analysis, if one includes nonprofit and public hospitals that are managed by for-profit firms, as well as non-acute care hospitals (e.g., psychiatric and rehabilitation facilities), the for-profit sector among hospitals grew to about 25 percent of the total by the late 1990s.

Great increases in for-profit ownership have occurred elsewhere in the US health care system since the 1980s, as is shown in Table 2. Health maintenance organizations (HMOs), which were almost exclusively nonprofit in 1980, are now predominantly for-profit (Gray 1999). Strong trends in the for-profit direction also occurred among home health agencies and dialysis treatment centers. Within the trend toward for-profit ownership has been a second trend—ownership not by individual proprietors but by investor-owned corporations that own multiple units (Institute of Medicine 1986, Corrigan et al. 1997). Similar trends toward multi-institutional arrangements have also occurred among nonprofits. This, along with increased entrepreneurial activity by nonprofits (Sloan 1998) has led to concerns among many observers that the

character of many nonprofit health care organizations has changed in ways that mirror the for-profits.

Comparing the cases of community hospitals and HMOs, some of the factors that are involved in the persistence and change in ownership patterns among health care organizations can be seen (Gray 1999). On the side of persistence, many nonprofit hospitals, with their multiple important community roles (care giver, employer) and their governing boards often made up of prominent citizens, have received both charitable and governmental support in bad economic times. The difficulty of closing a nonprofit hospital has been frequently noted in health policy circles as hospital occupancy has declined since the 1980s. Nonprofits have also engaged in a variety of strategies to compete and grow—entering into multihospital systems, merging, joint-venturing, and establishing for-profit subsidiaries. Use of these strategies has carried a price, however. The increased commercialism of the nonprofits has attracted criticism and fanned doubts about their deservedness of tax exemption.

The most important factor affecting the ownership composition of the health care field is public policy. It

could hardly be otherwise since the basic corporate structures of for-profit and nonprofit organizations are themselves products of public policy. But the role that public policy plays in influencing the composition and performance of health care organizations has become increasingly apparent since the 1950s. The creation and growth of the investor-owned hospital industry was a direct result of capital reimbursement policies built into the Medicare legislation in 1965 (Gray 1992). The interest costs incurred in borrowing money to purchase hospitals was a reimbursable expense under Medicare, and gave purchasers annual payments for the noncash expense called depreciation. Since Medicare payments to hospitals were based on reimbursing them for costs incurred, the legislation also provided for giving for-profit hospitals generous 'return-on-equity' payments so that investors would receive a return on their investments. All of these provisions stimulated the growth of the investor-owned companies into the early 1980s, when, in separate actions, Congress reduced return-on-equity payments and changed the depreciation rules so that Medicare would pay for the depreciation of any particular asset only once. After these changes were made, the leading investor-owned companies all stopped growing at a time in which it was being predicted that they were going to take over the system.

Similarly, the growth of the nonprofit HMO industry in the 1970s and the for-profit transformation of the field in the 1980s were direct results of public policy. First, provisions of the HMO Act of 1973 made capital available to nonprofit HMOs. But the field was transformed in the for-profit direction in the 1980s as a result of the Reagan administration's decision to phase out the HMO Act's support of nonprofit HMOs and to undertake a program to stimulate investment in for-profit HMO companies, which had not previously existed; a surge in demand for managed care from the corporate sector which was experiencing large increases in health care costs; and inattentive state regulators who allowed many nonprofit HMOs to be acquired for a pittance by their leadership and converted into for-profit organizations (Gray 1999).

Permissive state regulation also created the environment for a surge in the acquisition of nonprofit hospitals by large investor-owned firms (Columbia/HCA, Tenet) in the early and mid-1990s (Gray 1997). This, along with conversions involving HMOs and state Blue Cross insurance companies, led more than 25 states to pass legislation in the late 1990s to regulate nonprofit conversions. Some 120 new foundations were created in the 1980s and 1990s from the proceeds of the sale of nonprofit health care organizations (Isaacs et al. 1997). Studies of the consequences of hospital conversions have generally shown them to have occurred among financially troubled institutions and that few negative consequences could be documented, either because the institutions had previously engaged in only low levels of charitable activity or

because strong contractual protections had been built into the sales agreement (Mark 1999, Young and Desai 1999). Interestingly, there have also been examples of for-profit hospitals that reverted to the control of nonprofit organizations, though the trend remains in the direction of greater for profits.

4. Conclusion

The American health care experience has shown that for-profits and nonprofits can coexist in the same field. It has also shown that the comparative performance of the two forms of organizations are influenced by contextual factors such as the incentive structures built into the systems by which they are paid, regulatory regimes, professional values, and public expectations. Although for-profits generally appear to be more responsive to economic incentives than are nonprofits, nonprofits cannot be indifferent to economic realities. It is increasingly apparent that policy-makers need to pay attention to the characteristics of the organizations that provide health services as they design programs, payment systems, and regulatory structures to achieve desired health policy goals.

The US experience also shows that the for-profit/nonprofit composition of fields can change, sometimes quite rapidly, as can the ownership form of particular organizations. In a policy-sensitive field such as health care, the future of for-profit and nonprofit health care organizations is likely to be affected quite substantially by public policies. These policies may be focused specifically on some aspect of the ownership issue, but the most important ones may have an ownership-neutral motivation, such as the desire to control costs in large governmental programs. For segments of the health care industry in which government is the dominant payer, it may be difficult for investor-owned companies to generate revenues in the long run that will be satisfactory to investors. High profitability is likely to lead to cuts in the programs that fund that profitability. For-profits also face suspicion that has been driven by highly publicized fraud-related scandals involving some leading health care companies, plus widespread distrust of the predominantly for-profit HMO industry.

For their part, nonprofits face a twin challenge. In a highly competitive health care system dominated by large purchasers and without direct governmental subsidies, they must perform well enough economically to be able to continue paying their bills and to maintain their access to capital, which depends upon lenders' confidence in their ability to repay their debts. As they do this, they must also contend with a growing skepticism, fed in part by their own economically driven policies and practices, regarding whether they engage in sufficient charitable activity to deserve the tax-related benefits they receive.

With health care accounting for almost 15 percent of the US economy, investor money is likely to be available to fund new ideas for making money from the provision of health services, and there is interest among American companies in making investments in other countries as well. The clear advantages that the for-profit form might offer are the ability to raise capital, a willingness to take risks and innovate, and, perhaps, a reservoir of experience that might be transferable into new contexts. But will countries that have had little experience with this organizational form in their health care system find such factors outweigh possible costs in terms of a changed ethos of the health care system, major shifts in power and control, and potential serious problems of exploitation of the informational asymmetries that typify health care?

See also: Alternative and Complementary Healing Practices; Gender and Health Care; Health Care Delivery Services; Health Care: Legal Issues; Health Care Markets: Theory and Practice; Health Care, Rationing of; Health Care Technology; National Health Care and Insurance Systems; Psychiatric Care, Economics of

Bibliography

- Arrow K 1963 Uncertainty and the welfare economics of health care. *American Economic Review* **53**: 941–73
- Claxton G, Feder J, Shactman D, Altman S 1997 Public policy issues in nonprofit conversions: An overview. *Health Affairs* **16**: 9–28
- Corrigan J M, Eden J S, Gold M R, Pickreign J D 1997 Trends toward a national health care marketplace. *Inquiry* **34**: 11–28
- Fox D M, Schaffer D C 1991 Tax administration as health policy: Hospitals, the IRS, and the courts. *Journal of Health Politics, Policy, and Law* **16**: 251–79
- Frank R G, Salkever D S, Mullan F 1990 Hospital ownership and the care of uninsured and Medicaid patients: Findings from the National Hospital Discharge Survey 1979–1984. *Health Policy* **14**: 1–11
- Gray B H 1992 *The Profit Motive and Patient Care: The Changing Accountability of Doctors and Hospitals*. Harvard University Press, Cambridge, MA
- Gray B H 1997 Conversions of HMOs and hospitals: What's at stake? *Health Affairs* **16**: 29–43
- Gray B H 1999 The changing face of health care. In: Clotfelter C T, Ehrlich T (eds.) *Philanthropy and the Nonprofit Sector in a Changing America*. Indiana University Press, Bloomington, IN
- Gray B H, Carrino G, Collins S, Gusmano M 1999 *The Empirical Literature Comparing For-Profit and Nonprofit Hospitals, Managed Care Organizations, and Nursing Homes: Updating the Institute of Medicine Study*. The New York Academy of Medicine, New York
- Hansmann H 1980 The role of nonprofit enterprise. *The Yale Law Journal* **89**: 835–901
- Hartz A J 1989 Hospital characteristics and mortality rates. *The New England Journal of Medicine* **321**(25): 1720–5
- Hawes C, Phillips C D 1986 The changing structure of the nursing home industry and the impact of ownership on quality, cost, and access. In: Institute of Medicine (ed.) *For-Profit Enterprise in Health Care*. National Academy Press, Washington, DC
- Himmelstein D U, Woolhandler S, Hellander I, Wolfe S M 1999 Quality of care in investor-owned vs not-for-profit HMOs. *Journal of the American Medical Association* **282**: 159–63
- Institute of Medicine 1986 *For-Profit Enterprise in Health Care*. National Academy Press, Washington, DC
- Isaacs S L, Beatrice D F, Carr W 1997 Health care conversion foundations: A status report. *Health Affairs* **16**: 228–36
- Kuhn E M, Hartz A J, Krakauer H, Bailey R C, Rimm A A 1994 The relationship of hospital ownership and teaching status to 30- and 180-day adjusted mortality rates. *Medical Care* **32**(11): 1098–108
- LeBlanc A J, Hurley R E 1995 Adoption of HIV-related services among urban US hospitals: 1988 and 1991. *Medical Care* **33**(9): 881–91
- Lewin L S, Eckels T J, Miller L B 1988 Special report. Setting the record straight: The provision of uncompensated care by not-for-profit hospitals. *The New England Journal of Medicine* **318**(18): 1212–16
- Mark T L 1999 Analysis of the rationale for, and consequences of, nonprofit and for-profit ownership conversions. *Health Services Research* **34**(1): 83–101
- Marmor T R, Schlesinger M, Smithey R W 1987 Nonprofit organizations and health care. In: Powell W W (ed.) *The Nonprofit Sector: A Research Handbook*. Yale University Press, New Haven, CT
- Norton E C, Staiger D O 1994 How hospital ownership affects access to care for the uninsured. *RAND Journal of Economics* **25**: 171–85
- Powell W W (ed.) 1987 *The Nonprofit Sector: A Research Handbook*. Yale University Press, New Haven, CT
- Powell W W, Clemens E S (eds.) 1998 *Private Action and the Public Good*. Yale University Press, New Haven, CT
- Rodwin M A, Okamoto A 2000 Physicians' conflicts of interest in Japan and the United States: Lessons for the United States. *Journal of Health Politics, Policy, and Law* **25**: 343–75
- Salamon L M, Anheier H K, List R, Toepler S, Sokolowski S W 1999 *Global Civil Society: Dimensions of the Nonprofit Sector*. Johns Hopkins Center for Civil Society Studies, Baltimore, MD
- Schlesinger M 1998 Mismeasuring the consequences of ownership: External influences and the comparative performance of public, for-profit, and private nonprofit organizations. In: Powell W W, Clemens E S (eds.) *Private Action and the Public Good*. Yale University Press, New Haven, CT
- Schlesinger M, Bentkover J, Blumenthal D, Musacchio R, Willer J 1987 Privatization of health care and physician's perceptions of access to hospital services. *The Milbank Quarterly* **65**(1): 25–58
- Shortell S M, Morrison E M, Hughes S L, Friedman B S, Vitek J L 1987 Diversification of health care services: The effects of ownership, environment, and strategy. *Advances in Health Economics and Health Services Research* **7**: 3–40, 111–14
- Simon J G 1987 The tax treatment of nonprofit organizations: A review of federal and state policies. In: Powell W W (ed.) *The Nonprofit Sector: A Research Handbook*. Yale University Press, New Haven, CT
- Sloan F A 1998 Commercialism in nonprofit hospitals. In: Weisbrod B A (ed.) *To Profit or Not to Profit*. Cambridge University Press, Cambridge, UK

- Sloan F A, Picone G A, Taylor D H, Chou S Y 2000 Hospital ownership and cost and quality of care: Is there a dime's worth of difference? *Journal of Health Economics* forthcoming
- Starr P 1982 *The Social Transformation of American Medicine*. Basic Books, New York
- Stevens R 1989 *In Sickness and in Wealth: American Hospitals in the Twentieth Century*. Basic Books, New York
- Stocker K, Waitzkin H, Iriart C 1999 The exportation of managed care to Latin America. *The New England Journal of Medicine* 340(14): 1131–6
- Weisbrod B A 1990 *The Nonprofit Economy*. Harvard University Press, Cambridge, MA
- Young G J, Desai K R 1999 Nonprofit hospital conversions and community benefits: New evidence from three states. *Health Affairs* 18(5): 146–55

B. H. Gray

Health Care, Rationing of

The term 'rationing' is used with different meanings in health care discussions (Mechanic 1979). Most generally, it refers to processes of allocation when demand for care exceeds the available supply or the willingness of payers to increase it. In its more narrow application, it refers to how a scarce medical resource, such as hearts, kidneys, or livers, is allocated among competing recipients, only some of whom can receive it. Medical care has always been rationed in both senses, but the concept itself is associated with much affect and denials that rationing occurs.

To the extent that medical care is treated substantially as a commodity, as in the USA, then the market distributes services. Individuals with more comprehensive insurance or the ability to pay out-of-pocket command a greater quantity of and more sophisticated services. But even in market systems, some life and death services in scarce supply transcend the market and are governed by different rationing schemes. In publicly funded medical systems, where most of the population has health care entitlement, whether through national health insurance or a national health service, the availability of services depends on decisions about what services will be financed and their comprehensiveness and intensity (Klein et al. 1996). Some such systems, for example the British National Health Service, allow competing private systems of care. In contrast, Canada prohibits private provision of services covered by its health insurance program.

Rationing processes and decisions are implemented at different levels of government, by sick funds, large insurance plans, provider organizations such as hospitals, clinics, and health maintenance organizations, and at the point of service delivery between clinician and patient. Disagreements are common on where in the hierarchy of health care organization varying types

of rationing decisions should be made (macro, meso, or micro) and on appropriate processes for making such decisions (Mechanic 1997). At each level, rationing may be more or less explicit.

1. Explicit vs. Implicit Rationing

Explicit rationing refers to the detailed specification of how resources will be allocated by a rationing authority. Explicit rationing includes specifying eligibility rules, the specific services to be covered, the amount of coverage for recipients, intervals for receiving services, copayment requirements, the proportion of the budget to be allocated to varying services, the geographic allocation of services, the specific form and location of many services, and appliances and drugs that may or may not be allowed.

Implicit rationing, in contrast, involves setting budgetary limits but allowing the processes of allocation to be worked out at lower organizational levels (Mechanic 1979). Implicit rationing includes prospective reimbursements or capitation payments. In each case the authority sets the financial limits but then allows the budgeted or capitated entity to establish the processes for decision-making. Total responsibility is sometimes delegated to the lower unit but more typically there are some explicit guidelines and constraints over discretion. All medical systems combine explicit and implicit processes but the balance may differ considerably from one nation to another and within varying national contexts.

Klein et al. (1996) describe a range of strategies for implicitly limiting resources allocated to individuals. These include denial as beneficiaries or for the services claimed; selection of beneficiaries or services on the basis of need or likely success, or demographic characteristics such as age; deflection of care to other sectors or providers; deterrence through bureaucratic barriers such as waiting lists and other delays; diluting services by reducing intensity or restricting specialty care; and premature closure or early discharge.

2. Concerns About Rationing

Despite general acceptance of the inevitability of rationing, major issues of controversy include who should determine rationing strategies, the openness of decision processes, and whether decisions should be made by managers or clinicians (New 1997). Common criticisms are that the implicit rationing processes used are not open, often favor the privileged, and result in unequal care for persons of lesser status, education, and sophistication, and with less skill in managing bureaucratic systems (Churchill 1987). Critics also argue that implicit rationing gives doctors free reign to act on their prejudices. The lack of openness makes it difficult to assess the fairness of decision-making or to

allow aggrieved individuals to successfully appeal exclusions or dilution of care. Implicit rationing is also seen as undermining democratic processes through which communities can bring their preferences to bear in the allocation of important public resources.

A significant difficulty with explicit rationing, in contrast, is that it mobilizes interest groups who seek to affect decisions about coverage. Bringing medical decisions into the public arena energizes a political process and instability in the medical care system (Mechanic 1995). While a particular treatment may be of little value in general, there may be instances, difficult to specify beforehand, where it is appropriate. Explicit rationing can not, in most instances, accommodate the complex contingencies associated with the use of a particular treatment for an individual patient.

Questions about transparency apply at all levels of decision-making. The criteria used when a national government divides a health budget among regions can have significant consequences. Funds can be allocated politically, on the basis of historical funding, and on the basis of algorithms that take account of population characteristics and dispersion, and varying concepts of health status and need. The availability of facilities and specialists differs widely among population areas and accounts for large variations in expenditures among varying population groups. Nations can seek to bring greater equity among regions, types of services, and even individuals by how they decide to distribute resources.

Within organizations, decisions are made to allocate funding among primary and specialty services, among specialty units, and between medical and related health services. These decisions can have large consequences for the likelihood that a person with a particular need can receive responsive care. At the clinical level, decisions about waiting time, access to specialized services, availability of specialized treatments, or home or hospital treatment can profoundly affect patterns of care. Debate persists as to the extent such decisions should be governed by public discussion and decision-making or left to the discretion of clinicians based on their own professional judgment and experience (Symposium *The Law and Policy of Health Care Rationing* 1992).

Concerns about fairness and facilitating public discussion of priorities would support making practices explicit but there are significant competing concerns. Despite increasing interest in knowledge-based medical practice, much remains uncertain. There is no consensus, in most areas, on standards that apply to a diverse patient population with varying needs and inclinations and unique medical histories. Medical knowledge and practices are rapidly changing and it is uncertain whether explicit standards, once established, can be modified to keep pace with new knowledge. Moreover, medical care is an iterative process with assessment and disease management

developing over time. As a complex process of discovery and negotiation, clinical practice is not easily subjected to definitive technical rules.

The notion of explicit decision-making at the level of direct patient care implies that individuals with comparable conditions have similar needs and preferences. But treatment must be flexible and take account of varying social contexts, cultural inclinations, and personal preferences. Treatments of comparable medical effectiveness may have vastly different implications for work and family life, or for comfort under varying circumstances. To treat medical care simply as a technical process, devoid of meanings, preferences, and differential consequences, contributes to dehumanizing medicine, and is unacceptable to many professionals and patients.

3. Rationing of Specific Resources

Rationing becomes especially salient when problems result, not simply because of the unwillingness to provide more resources or allocate them fairly over some population (which may result in increased illness and death in a statistical sense), but when there is an inability to provide a life-saving resource to all identified individuals who seek it. Such situations involving the allocation of hearts, kidneys, livers, and lungs have been called 'tragic choices' because there are no fair or generally accepted rules for making the decisions as to who receives the resource and who dies (Calabresi and Bobbitt 1978). A variety of criteria have been applied from one occasion to another, such as allocation by age, by community contribution, by the largest opportunity to benefit, by family circumstances, by ability to cooperate in the treatment process, and many more, but all raise serious ethical conflicts (Elster 1992). Age is commonly suggested as a suitable criterion (Callahan 1990) and has been used in making tragic choices (Aaron and Schwartz 1984) but has been strongly contested. The most satisfactory criteria are medical ones based on likely technical success. But even these criteria do not adequately discriminate among potential beneficiaries, and social judgments and biases inevitably intrude (Halper 1989). For example, judgments about who can best cooperate in treatment typically give preference to patients with more education, higher intelligence, attractive personality characteristics, and strong social attachments.

Several analysts have concluded that because the equity issues cannot be resolved intellectually, the fairest mode of allocation is a lottery where persons with comparable medical qualifications have equal chances to be selected. There is little professional or public support for such selections because most people believe that consensual judgments can be made that would improve on any random process. Many rationally constructed strategies of allocation fail also

because of public perceptions. The public places greater value on rescuing identified lives than on saving statistical lives (Hadorn 1991).

4. Specific Rationing Policies/Practices

As evidence of inequalities in care increases, motivation to distribute limited health resources more equitably has led several countries to address rationing and seek explicit allocation principles and rules. The Oregon rationing program under Medicaid (the USA's federal-state partnership to provide health care to poor and disabled subgroups) was followed with great interest in Europe and represents perhaps the most extreme example of explicit rationing (Leichter 1999). This effort sought to increase enrollees under its Medicaid program by selecting to rank treatments consistent with judgments of efficacy and other community values. Through an elaborate process of evaluating condition-treatment pairs, a list was developed that prioritized these pairs. The basic idea was to establish a funding line, given available funds, that excluded from coverage the lowest ranking condition-treatment pairs. The formal evaluation and ranking process and the use of community meetings to involve consumers and professionals in defining the values for performing the rankings excited the imagination of many outside observers. The process was seen as a rational and fair way to make difficult public decisions about what would and would not be paid for. The Oregon program, however, was not quite what it appeared to be. It was a very public and explicit format for designing the benefit structure. The Oregon Medicaid program is predominantly a managed care program in which health maintenance organizations (HMOs) are capitated to provide all funded services. Once capitated, HMOs are free to provide whatever services they deem necessary and need not conform to the list. Thus, the list was more an instrument for bargaining over rates than an explicit set of rules for rationing health care. Early implementation was aided by increased state funding and a favorable economy, which allowed the establishment of acceptable capitation rates. The political sustainability of such an approach under more difficult times is yet to be tested.

The British National Health Service (NHS) exemplifies implicit rationing. There are no explicit guaranteed benefits in the NHS but the system seeks to provide all necessary care within the resources available. Care is rationed by the availability of professionals and treatment facilities, by professional decisions about intensity of care and referrals, and by waiting lists for various types of services (Klein et al. 1996, Aaron and Schwartz 1984). The extraordinary professional discretion within the NHS has led to efforts for greater equity and accountability in resource use. Under recent reforms, health authorities were given pur-

chasing responsibilities and could limit the treatments covered. The political difficulty of doing so explicitly in any but marginal aspects of medical care became evident. One exception was the effort to ration Viagra explicitly, a popular but expensive drug for treatment of impotence.

5. Ethical and Political Issues with Rationing Policies/Practices

A major ethical issue concerns the appropriateness of having physicians, acting as agents of their patients, also responsible for allocation decisions. Traditionally, doctors have been seen as sole agents of their patients, trying to get them what they need. Some believe it is unethical to dilute such agency with allocation responsibilities at the bedside. Others believe that physicians must have a dual responsibility for both their patients and the 'commons.' With the development of HMOs and managed care strategies in the USA, financial incentives are increasingly used to encourage doctors to allocate resources efficiently. There are worries that such incentives create ethical conflicts between the doctor's financial interests and patient care. Such conflicts are also common in unmanaged care when physicians have financial interest in diagnostic and treatment facilities to which they make referrals or in other practices for which they receive financial incentives. Studies have consistently found that such arrangements increase the use of ancillary services (Rodwin 1993).

There is widespread agreement that it is appropriate and necessary for some governing authority to establish financial constraints on health care expenditures. All national systems have such constraints in place for at least public expenditures. Few believe that it is any longer possible, even in the richest of nations, to provide open-ended financing, allowing physicians to provide all they deem useful. A more-contested issue is the extent to which such an authority, once having established such limits, should specify how the budget should be divided among different types of facilities, modes of care, professional groups, types of patient needs, and treatments and procedures, and by what criteria. The extent of authority given to the nation, as compared with regions, health districts, states, or even smaller population entities, elicits tensions in many contexts. Issues also arise as to whether organizational rationing decisions are best made by management, speciality services, or professionals exercising their best individual judgment.

A possible model that blends explicit and implicit rationing is one emphasizing clinical advocacy within a structure of procedural justice. Such a model would explicitly constrain some clinician and patient decisions through clear rules that are reasonably justified, while also allowing doctors and patients to challenge decisions and argue for revision through dispute

resolution mechanisms. Such compromise models offer opportunity for achieving more consensus.

Public views of rationing processes are important because any rationing process that lacks political legitimacy becomes unstable. One of the attractions of the Oregon program was its openness and the participation of large numbers of people. In contrast, in many areas of the USA, the implementation of utilization review (a form of explicit rationing) by managed care organizations overruling physician and patient definition of need has resulted in a significant political backlash. Managed care organizations are finding that transferring risk to professional groups (implicit rationing) kindles less friction than imposing explicit decisions on professionals and patients.

It will take decades for rationing issues to be sorted out. Health care systems and facilities are experimenting with a range of strategies that may improve decisions and help achieve legitimacy for rationing. These include community representation and involvement in decision-making, surveys to elicit community preferences, focus on evidence-based services, use of treatment guidelines and pathways, involvement of patients in treatment decisions, and implementation of disease management approaches. Rationing and the ability to sustain whatever system is preferred depends on achieving a sufficient level of value consensus so that the difficult choices are widely perceived as fair and responsive to individual need.

Bibliography

- Aaron H J, Schwartz W B 1984 *The Painful Prescription*. The Brookings Institution, Washington, DC
- Calabresi G, Bobbitt P 1978 *Tragic Choices*. Norton, New York
- Callahan D 1990 *What Kind of Life*. Simon and Schuster, New York
- Churchill L R 1987 *Rationing Health Care in America: Perceptions and Principles of Justice*. University of Notre Dame Press, Notre Dame, IN
- Elster J 1992 *Local Justice: How Institutions Allocate Scarce Goods and Necessary Burdens*. Cambridge University Press, Cambridge, UK
- Hadorn D C 1991 Setting health care priorities in Oregon. *JAMA—Journal of the American Medical Association* **265**: 2218–25
- Halper T 1989 *The Misfortunes of Others: End-stage Renal Disease in the United Kingdom*. Cambridge University Press, Cambridge, UK
- Klein R, Day P, Redmayne S 1996 *Managing Scarcity: Priority Setting and Rationing in the National Health Service*. Open University Press, Buckingham, UK
- Leichter H M 1999 Oregon's bold experiment: whatever happened to rationing? *Journal of Health Politics Policy and Law* **24**: 147–60
- Mechanic D 1979 *Future Issues in Health Care: Social Policy and the Rationing of Medical Services*. Free Press, New York
- Mechanic D 1995 Dilemmas in rationing health care services: the case for implicit rationing. *British Medical Journal* **310**: 1655–9
- Mechanic D 1997 Muddling through elegantly: finding the proper balance in rationing. *Health Affairs* **16**: 83–92

New B (ed.) 1997 *Rationing: Talk and Action in Health Care*. BMJ Publishing Group, London

Rodwin M 1993 *Medicine, Money, and Morals*. Oxford University Press, New York

Symposium, The Law and Policy of Health Care Rationing 1992 *University of Pennsylvania Law Review* **140**: 1505–2095

D. Mechanic

Health Care Systems: Comparative

1. Introduction

Each country's health care system is contained within a singular configuration of culture, history, politics, and economic capacity, and from these derive a specific mode of financing, delivering, and even evaluating health care. Yet, despite the uniqueness of each health care system, all are experiencing similar exogenous pressures, albeit to different degrees, arising from unprecedented developments in science and technology, demographic distributions, and patterns of disease, to name a few. And, despite their differences, many countries are responding to these pressures in remarkably similar ways. As Brown (1998, p. 36) puts it, 'Everywhere one hears the weary refrain: "We can no longer afford the [health care] system we have."'

To some scholars these similarities in pressures and responses suggest an impending convergence in health care systems, defined as 'a certain macro process in which a narrowing of system options takes place' (Mechanic and Rochefort 1996, p. 242). Mechanic and Rochefort emphasize, however, that convergence does not imply a blurring of social, cultural, and ideological distinctions among nations. Indeed, each nation's response necessarily builds on 'preexisting social institutions and professional organizations' (1996, p. 243). A number of scholars (cited throughout this article) have identified a specific kind of convergence in health care: many nations are turning to some type of market reform in order to control expenditures. This trend entails an unusual paradox: many of the ideas on market reform emanate from the United States, and yet the United States remains, not just an 'exception' among advanced industrial societies in its commitment to a welfare state in general and to health care in particular (Ruggie 1996), it is also a pariah from the perspective of universal coverage. At present, over 44 million Americans—approximately 18 percent of the population—are without health insurance, and millions more are underinsured. Adding to the paradox of the convergence, it is not altogether clear that the new methods offer a long-term solution to the problem of rising health care costs (Light 1997). For instance, an increase in administrative expenses is frequently as-

sociated with the need to assess competitive bids and manage the varieties of insurers and providers in a competitive market. In addition, concern about the consequences of market-based reforms on equity and access may already be diffusing, if not halting, recent innovations.

These inconsistencies in processes and outcomes suggest that the march toward marketization and the convergence it implies may be more apparent than real (OECD 1994, Gray 1998). Jacobs (1998, p. 5), among others, observes that while marketization connotes a direction of change and implies a choice of policy instruments, empirically it can become a 'highly eclectic venture.' His study shows that reforms—which introduced provider competition, greater cost-consciousness, and more patient choice—in the United Kingdom, Sweden, and the Netherlands actually resulted in divergence. Contrasting goals among political actors embedded in different institutional systems led to outcomes that were more significant for their dissimilarities than their resemblances.

Focusing on data from the 29 OECD countries and experiences from selected cases, this article explores the convergence/divergence debate by examining some of the main changes occurring in health care systems at present and some of the key issues driving change. It argues that convergence may be occurring at the level of interest among governments in experimenting with market-based forms of organizing health care, but whether this development constitutes what Max Weber would call a 'master trend' (a dominant movement in social change), sufficient to coalesce other distinctions, is far from clear. For an equally important concurrence is the resort on the part of governments to various types of regulation to mitigate the impact of market mechanisms. Hence, what we are witnessing is an oxymoron of modern society first identified by Karl Polanyi decades ago: planned markets. In general, this thesis also captures the experience of developing countries beyond the OECD, many of which are trying to introduce market reforms, and not only in health care. However, since planned markets require a relatively strong state, developing countries may be hindered in their capacity to harness viable market-based innovations.

2. Financing and Spending: An Overview

Health care systems are most commonly categorized according to three main methods of financing: general taxation, compulsory contributions to social insurance, and voluntary private insurance or direct payment (see Table 1). Health care systems funded through taxation tend to achieve the highest degree of universal coverage and, in principle, equity in access and provision. Within each category, however, significant differences occur in the delivery of health care services and these may affect patterns of coverage.

For example, Britain, Canada, Denmark, and Sweden all raise funds for health care through general taxation. But Britain considers its system to be a national health service, a concept more commonly associated with socialist countries. The other countries have national health insurance systems. The main difference is that under a national health service, ownership and/or control of the factors of production is public. Ownership of the factors of production in national insurance systems varies, but in general most hospitals are in the public domain whereas physicians consider themselves to be independently employed. Another difference pertains to the degree of centralization. In Canada, the national insurance system is funded by a combination of federal and provincial taxes. Also, the Canadian system is administered by the provinces, the Danish and Swedish systems by counties. While administration of the British national health service is decentralized, overall priorities emanate from the central government insofar as local governments have no independent source of revenue.

National or social health insurance systems are much more frequent than national health service systems. In most national health insurance systems, unlike in Canada and Scandinavia, financing is based on employer and employee contributions to an insurance fund (Belgium, France, Germany, the Netherlands, Norway); these contributions are compulsory for all employers and many workers. The funds may be administered privately (this is most likely if the funds are closely tied to employer or employee organizations) or publicly. Governments sometimes contribute to the funds as well and invariably set up some contributory mechanism to provide for special groups (such as low income or unemployed individuals). Social insurance systems generally achieve their universality by being tied to social security systems. Ownership of the factors of production varies considerably, but in general there is more private control than in systems funded through taxation.

Voluntary private health insurance is purchased by either employers or individuals or, in some cases, by a combination of the two. While this system predominates in the United States, it occurs in many other countries as a supplement to the national system. Depending on the country, private insurance may cover either private services only or some portion of public services as well.

Most countries house some mix of these models. For instance, in New Zealand hospital and specialist services are funded through general taxation, whereas financial support for ambulatory services and pharmaceuticals derives from an earmarked social security fund. In the United States, the Veterans Administration resembles a national health service and is funded through general taxation; Medicare is a national insurance system for those aged 65 and over as well as for the disabled, and is based on compulsory employer and employee contributions; and Medicaid is an

Table 1
Financing methods

General taxation	Compulsory social insurance	Voluntary private insurance
Britain (national)	Belgium France	USA Elsewhere to supplement national system
Canada (federal and provincial)		
Denmark (primarily county)	Germany	
New Zealand (hospital and specialist, national)	Netherlands	
	New Zealand (ambulatory and pharmaceutical)	
Sweden (county)	Norway	
USA (VA—federal; Medicaid—state and federal)	USA (Medicare)	

insurance program for the poor funded through general taxation.

Some recent efforts to control rising costs have tinkered at the margins of basic financial arrangements, but none has significantly changed a country's distinctive modality. Many cost-control efforts have sought to increase the contribution of private payments to overall expenditures. Even though all advanced industrial societies except the United States have 'national' systems, there is a considerable range in the public/private mix (Ruggie 1996). The term privatization suggests a direction of change away from the state, but actual cases of privatization demonstrate that the process is complex and does not necessarily shed government responsibilities.

Most countries have encouraged an increase in the role of private insurance by, for instance, granting a tax deduction to employers or individuals (Australia, Britain) or reducing the amount of public provision of certain services, leaving the private sector to fill the gap (Canada). Most countries have also increased charges to patients each time they use certain services (including hospital care in Sweden and general practitioners in New Zealand), but these additions tend to be nominal and aimed at enhancing cost consciousness and deterring frivolous or unnecessary use. Efforts to shift financial responsibility invariably introduce not only consumerism and commercialism but also stratification and selectivity. For instance, private insurance companies may discriminate by health status and social characteristics (such as age and class), charging certain individuals more while offering a greater range of services to those who can pay higher premiums. As a result, new government regulations have been deployed to correct these inequities.

For their part, private insurance companies are also engaging in cost-shifting by raising the price of premiums, deductibles, copayments, and so on. Similar increases in recipient contributions to social insurance funds have occurred as well. Some social insurance systems have attempted to increase cost consciousness by offering consumers greater choice

among plans (Germany, the Netherlands, and Sweden). While this fundamental market principle—choice—has increased competition within health care systems, its impact on costs is dubious because, unlike the predictions of market models of rationality, patients are not inclined to choose less care.

It is difficult to assess the effect of these privatization measures on the overall burden of health care expenditures, because they rarely appear alone. Also, while privatization by definition connotes some form of marketization, what exactly occurs is too variable and vague to be analytically useful. To confuse the issue even more, marketization can occur without privatization (as in Britain). However, just as privatization may in the end not entail a decline in the role of government, so too marketization is sometimes coupled with greater government regulations. Far from being contradictory, this combination may well provide the key to achieving the new triad of goals in health care: efficiency, effectiveness, and equity.

3. *Factors of Production: State vs. Market*

There is general agreement among scholars in the field of health policy that market mechanisms alone are not appropriate in matters of health care. Not only are consumers driven by need and not want, but when in need consumers do not exercise rational choice and prefer to spend to their limits (Evans 1999). However, few consumers are free actors in health care: governments and other third-party insurers buffer individual decisions and indeed shape them. What exists in all health care systems, then, are degrees of marketization and regulation, with pendulum-like swings between them reflecting the inevitable difficulty of staying too long on one side. That we are currently witnessing a (re)turn to market-based mechanisms indicates certain shortcomings in government controls, to be sure. But we ought not expect the market to replace the state, simply to stimulate correctives. Markets in health care, and especially those introduced by governments,

Table 2
Physician reimbursement

Fee-for-service	Salary	Capitation
Belgium	Britain (hospital specialists)	Britain (GPs)
Canada	Norway (some municipalities and counties)	Denmark ($\frac{1}{2}$ GP income)
Denmark (specialists and $\frac{1}{2}$ GP income)	USA (some HMOs)	Netherlands (GPs and for lower-income patients)
France (ambulatory)	Elsewhere in public hospitals and health centers	
Germany		
Netherlands (for higher income patients)		
New Zealand (ambulatory)		
Norway (ambulatory)		
Sweden		
USA (some Medicaid and Medicare, and much private insurance)		

are bound to be 'quasi' at best and reflective of the dominant social values and norms implicit in their construction.

Market forces have been introduced in various areas of the delivery and provision of health care. However, either the introduction of market reforms has been conducted within the framework of government regulations, especially as these pertain to issues of equity and access, or government regulations have been reintroduced when the negative consequences of market mechanisms on equity and access have exceeded acceptable norms.

3.1 Physicians

Of the three main methods of reimbursing physicians—fee-for-service, salary, and capitation—fee-for-service is the most open-ended and has been the most difficult to control. Governments (as payers) have attempted either to limit overall increases in fee schedules (as in Canada) or to specify prospective rates for procedures (as in the American Medicare Fee Schedule) (Table 2). In countries with social insurance systems, efforts to exercise these controls are frequently encumbered by corporatist negotiations involving representatives of employers, workers, and physicians. Private insurance companies have found it very difficult to impose reimbursement controls or develop other disincentives to increased utilization of physician services under fee-for-service plans, perhaps because the main competitive edge of private insurance is the greater choice it offers consumers who are willing to pay for higher-priced physician services. Because fee-for-service reimbursement controls encourage physicians to increase the number of patient visits in order to maintain their level of income, governments in Canada and Germany have begun to monitor the volume of physician services.

Physicians tend to be salaried when they work in public hospitals, health centers, or certain types of health maintenance organization (HMO). While salaries are inherently easier to control, they may elicit reduced service provision, especially when specialists also have private fee-for-service practices outside the health care facility. Capitation payments (a specified annual or monthly amount for each patient) commonly accrue to primary care physicians or general practitioners. Although there are outer limits on the number of patients on a physician's list, in order to control costs there must be strong disincentives to refer patients to specialist physicians or services.

Greater control over physician reimbursement has been facilitated by the introduction of managed care. Managed care is also responsible for ushering in a shift away from fee-for-service reimbursement toward either modified fees (with prospective fee schedules of some sort) or increased capitation if not salaried reimbursement within a physician's portfolio. Managed care plans rely on primary care physicians to serve as gatekeepers to the system of care, referring patients only when they deem it necessary. Various incentives exist to reward physicians for their efforts, ranging from bonuses in American HMOs to managerial independence among fundholding GPs in Britain. Fundholding GPs are allotted a lump sum payment for their practices and allowed to make all expenditure decisions themselves. Developments such as GP fundholding also represent an attempt to increase provider cost consciousness by making providers purchasers of health care with significant budgetary discretion.

Managed care has also introduced increased forms of competition among providers. Some managed care organizations negotiate contracts with selected providers for their services, encouraging providers to devise attractive offers that promote the goal of 'value

for money.’ Some managed care organizations simply enlist providers who are willing to offer their services for a discount; physicians are motivated to participate in these plans because of the promise of increased volume. In many countries competition is occurring among providers within the public sector, but, while reimbursements levels are affected, the practice is oriented more toward enhancing cost consciousness than achieving any significant reduction in overall expenditure on physician services.

Physicians and patients in the United States have undergone a dramatic and relatively abrupt transition to managed care, but at the same time, this revolution has been stymied by a shortage of primary care physicians. Unlike other countries the government in the United States exercises only weak control over physician supply and distribution, based entirely on monetary incentives. It is expected that the still growing market for managed care will continue to influence physicians to choose primary care as a field of specialization. Meantime, however, experiments are underway (for example, in New York City) in the use of highly trained advanced nurse practitioners as primary care providers. These nurses perform all of the functions of primary care physicians with regard to prescription, referral, and hospital admitting authority, and they are paid at the same rate. They and their employers insist that they are not threatening the livelihood of physicians, simply filling a gap. But their appearance is part of a wider movement to demote the level of skill in certain jobs. In all countries, most types of health care worker are undergoing a deskilling as hospitals cut middle-range positions while hiring lower-skilled replacements, some of whom work only part-time.

Other controls on physicians aim to standardize what in some cases has been wildly divergent practice patterns (Wennberg 1999). Clinical or practice guidelines, also known as medical audits, contain specifications for how physicians should go about treating certain ailments. In most countries these guidelines are developed by the medical profession and monitoring is based on peer reviews, but in other countries (such as the United States) government officials or other interested parties have a hand in the process as well. As protocols for treatment options, clinical guidelines have a dual purpose, for they are also central to efforts to improve quality assurance. Accordingly, in some places they have been extended to include primary care. Finally, in a number of countries, physicians’ expenditures on pharmaceuticals are coming under greater control. Britain and Germany are holding physicians directly responsible for spending over a specified limit.

These and related developments have led to the suggestion that the dominance and authority of the medical profession is declining (Mechanic and Rochefort 1996, Raffel 1997). Certainly, managed care has removed much clinical decision making from

the sole domain of the physician and muddled it with administrative rules and budgetary constraints. Further proof derives from the British experience. Having once enjoyed considerable clinical autonomy (especially compared to the United States), physicians in Britain are complaining about the introduction of a planned market with its greater managerial control over medical decision making (Harrison 1995). Alternatively, managed care has enhanced the role of primary care physicians in relation to the secondary sector. Even in countries that already had well-developed primary care services, the incorporation of managed care has strengthened the vertical integration of sectors and placed primary care physicians at the intersection.

There can be little doubt that the heyday of physician dominance over health care is past. But we must look beyond the medical bureaucracy to understand the forces behind this development. For example, better patient knowledge, triggered not only by successful health promotion programs but also by the media and increased access to the Internet, has greatly altered patient–physician relations. Medical authority is also being tested by negative reactions against increasingly more invasive technologies and procedures as patients are turning away from overly medicalized treatments and institutional settings. While we can note change, we ought not jump to conclusions about the endpoint. There is no reason why demedicalization cannot exist alongside a continuation of medical authority. Furthermore, demedicalization has generated greater accountability criteria on the practice of medicine. Accountability requirements have perhaps gone furthest in the United States, where some state and local governments are publishing physician (and hospital) ‘report cards,’ ostensibly to provide consumers with the information they need to make their health care choices. The status of physicians is undoubtedly coming under greater scrutiny, due as much to marketization and its competitively-based principles as to broader social changes in the culture of health care.

3.2 Hospitals

In most countries hospitals consume the greatest portion of health care budgets. By now, various forms of prospective payment systems to hospitals exist in most countries. Prospective payment systems frequently include additional incentives to contain costs, such as allowing hospitals that underspend to keep the difference and average out their costs. Some countries, such as Germany, have prospective *per diem* rates. Unless they are combined with disincentives to extensive lengths of stay, *per diem* rates do little to hold down hospital expenditures. Many more countries have already adopted or are shifting to prospective per case rates based on the American Diagnostic-Related Group (DRG) system. In this system each of several

hundred categories of diseases and disorders is assigned a numerical value based on service utilization that is then translated into a reimbursement level. Most hospitals responded to the incentives within the DRG system to contain costs by reducing lengths of stay (for example, in Sweden the decline was from an average of 21.3 days in 1985 to 7.8 days in 1995, OECD 1998). The number of hospital beds per population also decreased as a result. However, both of these indicators of reduced hospitalization have occurred in all OECD countries except Korea and Mexico, whether or not countries adopted the DRG system or similar cost containment methods, indicating that a more generalized decline may be occurring in the role of the hospital.

Another prospective payment system, used for years in Britain and Canada, and now being widely adopted elsewhere (France, the Netherlands, Sweden, Finland) is global budgeting on an annual basis. Spending decisions are usually decentralized to the local and even hospital level. In some countries there are strict controls on public sector capital investment in hospitals, sometimes with separate budgeting mechanisms.

A novel form of independent decision making by hospitals has occurred in Britain where hospitals are now self-governing trusts. The move was prompted by the expectation that hospital managers would become more cost conscious if they were responsible for all decisions—purchasing, servicing, delivering, administering, and even paying specialist physicians. Hospitals advertise their services and compete for contracts with, among others, fundholding GPs and local authorities. Similar developments have occurred in Italy and Denmark.

The introduction of cost control measures in hospitals has raised concern about declines in intensity of service and concomitant quality of care. Because of the great diversity among countries in the size of the hospital sector, goals regarding efficiency differ. Countries that are attempting to reduce total hospital output are faced with the challenge of closing hospitals or at least beds that are not being used efficiently (US, Britain). These countries nevertheless join others in striving to encourage hospitals to produce more so as to achieve certain goals, such as reducing waiting times for nonemergency treatments. The ability to regulate hospital costs goes beyond administration, however, for what happens in hospitals is highly dependent on technological advances. Yet, everywhere technology is undergoing contradictory developments.

3.3 Technology: From High to Low

While cause-effect relationships are always difficult to impute, it is safe to say that hospital cost containment has been made possible because of new technologies

that have both reduced the risks of certain surgeries and enabled an increase in outpatient surgeries. Paradoxically, rather than expanding megacenters of health care, as in the past, technological innovations may be contributing to the opposite effect. For instance, with the ability to perform more office-based procedures, physicians may be spending less time in traditional hospital settings.

At the same time, reactions against certain technological developments have engendered a number of nonmarket forms of cost containment and a search for less invasive and more personal forms of care, all of which signal new conceptions of what constitutes quality in health care. For example, there has been an exponential increase in the home health industry, which in many countries is contained within the private sector.

In some countries there is also renewed interest in supplementing both primary and secondary care with community care facilities. These provide a range of caregiving services either at home or in clinics, or day centers not only for those recovering from hospital treatment but also for the chronically ill, the mentally ill, and the elderly. Palliative care facilities and hospices are also proliferating to meet the demand for more personal and less technologized care for the dying.

Community care goes hand in hand with a renewed emphasis on preventive medicine, which was first heralded as a cost containment mechanism by the early health maintenance organizations. Insofar as community care, preventive medicine, and related alternative forms of health and healing entail a greater role for patients, both in self care as well as in determining the course of treatment, they also represent a demedicalization of sorts. Nevertheless, as with other evolving relations in health care, we may be witnessing less a replacement of allopathic medicine than a complement to standard modalities.

4. Managing Markets in Health Care

An American economist, Alain Enthoven, first introduced the idea of 'managed competition' in health care in 1978. For the American context emphasis was put on the term 'managed,' which carried a dual meaning. First, Enthoven was as much an advocate for the development of managed care as for the competition that he felt should occur among managed care organizations based on issues of quality and cost. As his formulations have been imported into other countries, this emphasis has shifted somewhat, but not entirely. Countries with strong primary health care sectors—whether these have been based on general practitioners or community care clinics, or both—have always had the foundations for managed care. For these countries competition among primary, secondary, and tertiary

care providers has been more important for enhancing cost consciousness than reducing health care expenditures.

Enthoven also suggested that competition itself should be managed, that is to say regulated, by public agencies who would act as sponsors for consumers. Each sponsor was to be: 'an active, intelligent collective purchasing agent on the demand side ... that creates, develops, administers and enforces the rules of competition in a never-ending effort to root out market failures and to perfect the market' (1994, p. 1416). Countries with systems of financing based on social insurance funds already had the foundations of such sponsorship. The incorporation of managed competition in these countries (Germany, the Netherlands, New Zealand) has primarily enhanced patient choice of fund, although some selective contracting by the funds has also occurred. The consequent increased competition among insurers has raised some predictable problems. In Germany the federal government has introduced risk adjustment schemes to compensate funds that have been negatively affected by the infusion of new patients. Some observers suggest that the equalization and standardization that is resulting may lead Germany toward a single national sickness fund—an odd consequence of managed competition (Freeman 1998). In other words, as new methods of competition are introduced into health care systems, their potentially unsettling consequences are eventually, if not immediately, controlled by new regulatory mechanisms. Even in the United States, proposals for a national Patients' Bill of Rights promise to reestablish fundamental criteria of justice over the exigencies of competition. Everywhere the result has been to tame market externalities without completely debilitating market mechanisms.

5. Policies, Politics, and People

Given the pressure to expand medical services and the reality of other demands on strained government budgets, countries have been struggling with the dilemma of trying to maintain full and equitable access to high quality health care while at the same time controlling cost increases if not actually cutting expenditure. During the 1980s and 1990s, most countries experimented with some form of market mechanism in order to balance the equation. Interestingly, the effort to contain costs has occurred regardless of level of expenditure. Among those OECD countries actively engaged in health care reform, expenditures on health as a percentage of GDP range from 6.4 in Denmark and 6.9 in Britain to 10.4 in Germany and 14.2 in the United States (OECD 1998). The rate of growth in expenditures has slowed in a number of countries. Many scholars attribute this consequence to the diffusion of a single idea—that the market is best suited to achieve efficiency. Because

marketization has occurred regardless of the degree of state vs. market penetration across diverse national settings, some scholars have suggested that a convergence in health care systems is occurring. To some extent countries have accepted new policy instruments cautiously and rationally, learning from each others' successes and mistakes. But to a larger extent as countries have borrowed ideas and instruments, and adapted these to their own situations, the inevitable tensions between market and state, state and society, professions and politics, values and rationalities have given full play to the rich diversities among health care systems. The current swing toward market solutions to economic problems in health care is limited in its penetration of health care systems, for few countries tolerate for long the negative consequences of market forces on equality and access. Indeed, 'planned markets appear as an attempt to find more effective administrative instruments, rather than as a conscious or intended surrender of political control over the main objectives of health policy to market forces' (Saltman and Otter 1995, p. 10).

In the end we must ask, however, whether we are witnessing actual reform of health care systems or something less tangible. If markets are constantly subject to government control, perhaps the underlying goal of reform is cost consciousness and the efficiency it purportedly entails. But beyond the buzzwords, a more profound eventuality is in the offing. If actors are expected to become more aware of the economic consequences of their decisions to utilize health care services, they must also become more attuned to their willingness to forego medical options, for themselves and for others. That health care must be rationed is probably the most difficult realization encountered by governments and citizens. It rests on the premise that health care is a scarce or at least finite resource, a premise that is hotly debated among medical professionals and ethicists. But as the discussion above has indicated, implicit forms of rationing abound—inability to pay; primary care practitioner 'gatekeeping'—limited number of professionals; limited number of technologies; queues for treatment—to name but a few. It is unclear how fully these practices are understood to be forms of rationing. Few governments have confronted the issue head on and developed explicit protocols.

The state of Oregon in the United States has one of the only plans specifically denying coverage for certain treatments. But the plan applies only to a population unable to purchase private insurance—Medicaid (poor) recipients. No matter what type of rationing exists, someone has to make the decision about denying access. How a country goes about this task reveals as much about prevailing values and norms as do decisions about financing. The Oregon plan, which on the surface appears to be a Leviathan in a land of individual freedom, was actually the product of considerable grass roots participation. In Britain,

government control over the macrolevel of expenditure and priority setting is coupled with confidential case-by-case rationing decisions made by physicians, a process that relies on an unusual level of trust. The persistence of these cultural features of health care systems explains the incomplete successes of market reforms.

Ultimately, however, the story of market-based reform, and the answer to the questions raised about convergence or divergence, rest on an understanding of the role of political institutions in health care systems. For reform to be effective, governments must have the political will to control the demands and behaviors of all relevant actors. But at the same time, for health care to be adequately and justly distributed, governments must marshal citizen involvement in decision making processes in order to construct and legitimate collective goals in health care. This is a universal point, applicable to newly industrializing countries or countries with struggling economies no less than to those discussed in this article.

See also: Health Care, Rationing of; Health Policy; National Health Care and Insurance Systems; Public Psychiatry and Public Mental Health Systems

Bibliography

- Brown L D 1998 Exceptionalism as the rule? US health policy innovation and cross-national learning. *Journal of Health Politics, Policy and Law* 23(1): 35–51
- Enthoven A 1994 On the ideal market structure for third-party purchasing of health care. *Social Science and Medicine* 39(10): 1413–24
- Evans R 1999 Health reform: What ‘business’ is it of business? In: Drache D, Sullivan T (eds.) *Market Limits in Health Reform: Public Success, Private Failure*. Routledge, London
- Freeman R 1998 The German model: the State and the market in health care reform. In: Ranade W (ed.) *Markets and Health Care: A Comparative Analysis*. Longman, London
- Gray G 1998 Access to medical care under strain: New pressures in Canada and Australia. *Journal of Health Politics, Policy and Law* 23(6): 905–47
- Harrison S 1995 Clinical autonomy and planned markets: The British case. In: Saltman R B, Otter C von (eds.) *Implementing Planned Markets in Health Care: Balancing Social and Economic Responsibility*. Open University Press, Buckingham, UK
- Jacobs A 1998 Seeing difference: market health reform in Europe. *Journal of Health Politics, Policy and Law* 23(1): 1–33
- Light D W 1997 Lessons for the United States: Britain’s Experience with Managed Competition. In: Wilkerson J D, Devers K J, Given R S (eds.) *Competitive Managed Care: The Emerging Health Care System*. Jossey-Bass, San Francisco, CA
- Mechanic D, Rochefort D A 1996 Comparative Medical Systems. *Annual Review of Sociology* 22: 239–70

OECD 1994 *The Reform of Health Care Systems: A Review of Seventeen OECD Countries*. Organization for Economic Cooperation and Development, Paris

OECD 1998 *OECD in Figures*. Organization for Economic Cooperation and Development, Paris

Raffel M W 1997 Dominant issues: Convergence, decentralization, competition, health services. In: Raffel M W (ed.) *Health Care and Reform in Industrialized Countries*. Pennsylvania State University Press, PA, pp. 291–301

Ruggie M 1996 *Realignments in the Welfare State: Health Policy in the United States, Britain, and Canada*. Columbia University Press, New York

Saltman R B, Otter C von 1995 Introduction. In: Saltman R B, Otter C von (eds.) *Implementing Planned Markets in Health Care: Balancing Social and Economic Responsibility*. Open University Press, Buckingham, UK

Wennberg J 1999 Understanding geographic variations in health care delivery. *New England Journal of Medicine* 340(1): 52–3

M. Ruggie

Health Care Technology

1. Definition

According to the US Office of Technology Assessment, healthcare technology consists of ‘the drugs, devices, and medical and surgical procedures used in medical care, and the organizational and supportive systems within which such care is provided,’ (US Congress 1982) including diagnostic, preventive, therapeutic, rehabilitative, organizational, information, educational, and supportive technologies.

Social scientists routinely distinguish ‘advanced’ medical technologies from the more mundane healthcare technologies. The distinction between advanced and normal technology often implies that technology has become increasingly central in modern medicine. But in a broader historical perspective, technology has always been central to healthcare, indeed social life. Features of life once popularly regarded as advanced technology, such as patient records or needles, have now been absorbed into routine.

2. The Importance of Healthcare Technology for Social Scientists

The broad definition of healthcare technology encompasses any tool from a mundane shake-down thermometer to the latest version of the computerized patient record, presenting the healthcare field as layered with technologies. Social scientists have been mainly interested in the upper layers of the medical geology: they have written extensively about advanced

life-supporting technologies, organ transplantation innovations, reproductive advances, information technologies, and genetic breakthroughs that pose profound ethical and sociological questions about equity. The introduction of new biomedical technologies not only maps medical advances but also reflects changes in cultural understandings of health, illness, life, death, the body, and even the family (e.g., when reproductive technologies offer an alternative to sexual procreation), religion, politics, crime, nature, race, and gender.

A good example of how medical technologies influence cultural notions is the institution of brain death criteria by the Harvard Ad Hoc Committee on Death in 1968. In an archival study of the committee's correspondence, Mita Giacomini confirms the general suspicion that a significant factor tailoring the redefinition of death were the interests of the transplant community (Giacomini 1997). In 1967, after the first heart transplant in South Africa, the general public had become sensitized to the potential conflict of interest of transplant surgeons to declare patients dead before their heart had stopped beating. Up until that time, physicians and lay people had been unable to come up with criteria for death beyond the century old criterium of the heartbeat cessation. The tacit consensus seemed to be to view death as a bedside physician prerogative. The transplantation community needed more flexibility to receive timely access to comatose bodies and was actively involved in the creation of the new criteria. For example, clinical criteria were adapted to facilitate organ procurement. Even the term 'brain dead' rather than 'irreversible coma' fit with the transplant community's needs because only 'dead' people could serve as cadaver organ donors. In addition to the interests of the transplant community, the redefinition of death was influenced by the emergence of the electroencephalograph and the application of resuscitative and life support technologies.

While most sociological writings have had analytical purposes, some scholars have entered the field of health care technologies with interventionist intentions. The purpose of such socially engaged studies by feminists and other patient advocates is to influence health care policy. In their pioneering study of dying in hospitals, Barney Glaser and Anselm Strauss discussed how technological advances at the end of life created complicated lingering dying trajectories, posing special problems of information management for nurses and physicians (Glaser and Strauss 1965). Glaser and Strauss painted a poignant picture of dying patients and their relatives engaged in a painful game of extracting information from healthcare providers reluctant to reveal the terminal condition. As part of a general death awareness social movement which swept the United States in the 1960s, this study contributed to a revision of the medical school curriculum to include doctor-patient communication courses. Simi-

lar studies discussed how the rapid introduction of pharmaceutical, surgical, and technological innovations during the post-Second World War decades featured in the socialization process of individual physicians.

3. Different Conceptualizations of Healthcare Technology

Social scientists have viewed healthcare technology in at least three different ways, each of which comes with a number of assumptions about the role of technology, human actors, and the social world. At stake in the different conceptualizations is the autonomy of the technology and its power to shape social relationships. These different approaches greatly impact the critical potential of the analysis.

3.1 Social Science with Technology Perspective

In the first conceptualization, technology is not the central focus of the analysis but more a catalyzing variable in a network of social relationships. Healthcare technology remains unproblematic, assumed in the same way that most sociologists take the physical layout of hospitals for granted. What matters is how humans (as professionals, relatives, patients) cope with the new possibilities generated by an influx of instruments in technology-intensive medical settings. The primary goal of social scientists working in this approach is to analyze well-established topics of interests to observers of health and illness, that is, professionalization, social equality, social control, or medicalization, often relying upon the canon of a broad social constructivist perspective.

A good example of this approach is Renée Anspach's study of decision-making about life and death in a neonatal intensive care unit (Anspach 1993). She situates her ethnography in a technology-intensive environment and states that most of the dilemmas about keeping premature and sick infants alive result from the availability of the technology and expectations of staff and relatives. Yet, she does not name or describe what technology she is talking about, discuss its origins, or elaborate on the staff's work to keep the technology functioning and interpretive. Instead, Anspach writes about the neonatal intensive care unit as a complex social organization where different groups of people rely on diverse sources of information and weigh varying priorities to reach a decision about the continuation of aggressive treatment. Physicians, for example, rely more on technology-derived data than nurses who depend on closer, interactive contact with infants, leading to conflicts in life-and-death decisions. The healthcare technology confirms the power differences between an emerging and established profession.

Even scholars who make technology more central often rely on this socially determined conceptualization of medical technology. Susan Bell's interactive diffusion model of the artificial hormone DES (diethylstilbestrol) applies a social constructivist social network approach to healthcare technology (Bell 1989). Bell shows how the artificial hormone originated from within scientific and health communities with distinctive ways of defining problems and offering solutions. A broader political-economic context (US research funding post-Second World War) structured the relationships of these communities and reflected the way the technology diffused among physicians and their women patients. In this conceptualization, the focus of the social scientist's analysis remains with the social networks advancing the new technology. Bell explores how social communities gain or lose power because of their support for DES. Yet, how the meaning of the hormone itself changes as a result of those interactions remains unexplored.

The main assumption underlying this approach to healthcare technology is that the tools and instruments of medicine are shaped by social forces. Technology is a given (Anspach 1993), or a contingently diffused product following a trajectory set by social circumstances (Bell 1989). Dominant social groups invest values, power, and preferences in the healthcare technology, and it is these norms and paradigms that dominate the analysis. This approach offers interesting insights into the *social* aspects of healthcare technology but often glosses over the *technological* characteristics of medical technology. In other words, what makes health-care technology such a unique case, is often left unexamined.

3.2 Humanistic Social Science Argues Against Technology

The second conceptualization of healthcare technology originates from a humanist concern about the dangers of technology and is primarily intended as a critique of the healthcare field. In contrast with the first approach, the spotlight of the analysis is firmly centered on healthcare technology. Healthcare technology is singled out, its effects are scrutinized, compared to a technology-free past, and usually found wanting. Healthcare technology symbolizes here the power and arrogance of the medical professions, the hegemony of the medico-industrial complex, and the excesses of Western modernity. Technology becomes a modernist instrument by which some in society enforce power over others.

Authors working in this tradition emphasize how the presence of medical technology turns caregivers into technicians and leads to depersonalized and dehumanized care. Technology results in distancing and alienation in patient-care-provider interactions, a loss of dignity and integrity, and a colonization of the

body. This kind of analysis receives its persuasive power from a historical or cross-cultural comparison. The technology-saturated present is compared with idealized tropes of the 'natural' (the natural body, birth, or death) of past epochs and of others, unpolluted by modern healthcare technology.

This antitechnology analysis has been fueled by the Luddite legacy and the threat of a nuclear apocalypse, and is popular with both conservative scholars lamenting the moral, spiritual, and existential crisis in the secular West (Aries 1977) and Marxist scholars who analyze healthcare technology as an integral part of the capitalist profit structure (Waitzkin 1979). These politically diverse analyses of healthcare technology have in common that they believe that technology is a powerful force shaping social relationships. This belief has been referred to as *technological determinism*. Technological determinism comes in different strengths (Winner 1980): strong technological determinists believe that technology develops as the sole result of an internal dynamic and molds society to fit its patterns. In his book *Medical Nemesis*, for example, libertarian social critic Ivan Illich views the exportation of Western medical technologies as part of an international conspiracy to spread illness across the world and establish Western medical hegemony. According to Illich, international collaboration to eradicate virus-carrying mosquitoes resulted in short-term pyrrhic victories, in the long-term technological inference led to more virulent strains of tropical diseases and an aggravation of malnutrition (Illich 1976, Chap. 7). In a weaker version, medical technology is viewed as a political phenomenon in itself. Technology as politics occurs when a particular technological device or system becomes a way of settling an issue in a particular community or when technologies require particular kinds of political relationships. For example, physician-generated demand for expensive diagnostic technologies such as MRI scanners led to the consolidation and integration of large medical centers at the expense of many local community hospitals that did not have the patient load to justify the large capital investment. In both versions, the focus of the analysis is on what has been lost after the introduction of medical technologies. The purpose of these critiques is to warn of over-technologization and to recapture a mystified technology impoverished past.

This conceptualization of healthcare technology often works from an analysis of extremes and is therefore limited in the analytical insights it brings in the everyday workings of healthcare technology. Yet, because of its strong rhetorical packaging, this approach to technology has often lead to more social mobilization and actual social change than the other two conceptualizations. One can detect technological determinist tenets in the hospice movement, rallying for palliative care for terminally ill people and in the 'natural' birth movement, pointing out the physical

dangers of an overmedicalized birth. In both social movements, antitechnology sentiments united activists and led to concrete social change.

3.3 Technological Practice

A third conceptualization of healthcare technology finds a middle ground between the view that an interplay of social structure and culture shapes the use of technology and the belief that technology determines society. Drawing from the interdisciplinary field of science and technology studies, researchers have turned to what specifically is gained and lost in medical encounters when new instruments and drugs are introduced. This approach led to a number of detailed studies of the practice of healthcare technology in which the technology takes center stage. Healthcare technology is here not an *a priori* established entity but both the technological elements and the social characteristics emerge during the design process and while the technology is used. Social and technological characteristics are continuously reconfigured in interaction, to the point that it is difficult to tell where technology ends and the social begins.

As a prime example of this approach, sociologist Marc Berg addressed whether medicine has become more rational and scientific since the Second World War because of its reliance on formal tools such as standardized protocols, expert systems, clinical decision analysis, and automated decision systems (Berg 1997). Berg neither supported the Luddite camp which argues that such technologies primarily lead to deskilling nor bolstered the protechnology argument that the formal tools necessarily reduce the staff's workload, but instead showed how technologies have shifted medicine's understanding of rationality. Rationality was not a pregiven quality but achieved in tandem with the production and implementation of formal tools. Berg analyzed how the designers of the automated clinical decision analysis systems and protocols went beyond solving technical conundrums. The designers of formal tools studied the work activities and places in which their tools needed to function and made assumptions about users, environment, input, and expected output. For example, the designers of an automated cardiac decision-making system aimed for a computer that calculated a diagnosis and suggested treatment options after a healthcare provider entered diagnostic data about chest pain. The designers prescribed medical practice with guidelines (the algorithms of the diagnostic system), prerequisites (which data is mandatory or optional), and a reshuffling of existing professional relationships to make their devices work (not only physicians but nurses could enter data). In the end, the automated decision-making system provided a diagnosis and the tool worked, but not in the way the designers intended it because the tool's requirements became just one

element in the management of patients' medical conditions. For example, nurses quickly learned to enter the 'right' data based on the availability of beds in the intensive care unit. Medical practice and tools emerged simultaneously to constitute new forms of knowledges, new challenges, and problems to keep both the tools and the users on track, and ultimately new opportunities for diagnosing and curing.

The most radical conceptual innovation of the science and technology scholars was to extend agency from humans to things. According to the French sociologist Bruno Latour, this ontological extension of human agency occurs through the *delegation* of agency from humans to nonhumans (Latour 1987). Machines and instruments, such as thermostats and tape recorders, perform as agents because they substitute for the actions of people. As Stefan Hirschauer has shown, during surgery, pumps and monitors become a patient's 'externalized organs,' taking over vital functions (Hirschauer 1991). Because a malfunctioning ventilator might mean the end of life, there is no easy distinction between the human and the technical realm. Technology acts similarly to people. This does not mean that humans and nonhumans are equal. Andrew Pickering, for example, draws the line at intentionality; intentions cannot be attributed to nonhumans (Pickering 1995). Some critics have charged that analyzing technology as actors is disempowering, because ultimately humans need to be held accountable. But other researchers have linked the premise of technological practice to a more critical and interventionist agenda (Timmermans 1999).

The key characteristic of this approach is that changes in the social and the technical realm are investigated symmetrically as they enfold over time. Instead of assuming the predominance of social or technological forces in advance, the social scientist studies how technological innovations advance new medical practices and understandings about health and illness.

4. A Selection of Research Topics in Healthcare Technology

4.1 Organ Transplantation Technology

Like ethicists and physicians, social scientists have been particularly puzzled by the Janus face of biomedical technology: the instruments of medicine have the potential to postpone mortality and enhance the quality of life, yet, also possess the possibility of destroying the world as we know it. Nowhere is the Janus face of biotechnology better researched than in the area of the biomedical organ transplant technologies. The transplant of kidneys, hearts, lungs, heart and lungs, liver, and pancreas benefitted from the development of immunosuppressive drugs that counteracted the body's tendency to attack the foreign

organ and, in the case of US kidney transplants, from Medicare funding in 1972. Since the technology moved down a continuum from experimental to therapeutic, its utilization has prolonged and saved lives that would otherwise almost definitely be lost, but it also spawned many unintended, complex ethical, financial, legal, and social problems.

Renée Fox and Judith Swazey point out in a classic study that more than just organs are exchanged during organ transplantation (Fox and Swazey 1978). The technological procedure poses important questions about the corporeality of personal identity, the relationship of a body to its parts, and the boundaries that define social ties. Organ procurers present the donation of cadaver organs or, in the case of kidneys, live organs as the ultimate 'gift of life,' invoking Judeo-Christian values of altruism, generosity, humanitarianism, and freedom of choice. For the recipient, the donated organ might not only be under continuous threat of immunological rejection but might also carry a heavy symbolic load. Because a donated organ can not be reciprocated with another gift, the organ exchange locks two people and their relatives in an unequal relationship, creating what Fox and Swazey term 'the tyranny of the gift' (Fox and Swazey 1978, pp. 20–32). Relatives of donors might insist on developing kin relationships with recipients because a piece of their loved one 'lives on' in another person. For recipients, the reconstruction of the body requires an active reshaping of the self (Sharp 1995).

These social complications of organ donation are exacerbated by the 'scarcity of organs.' In the US, organ donation seems to have leveled off at 5,000 while the demand for organs has topped more than 33,000 and keeps rising (Sharp 1995). With a much higher demand than supply, who is more likely to receive organs? According to the American Medical Association, medical need should prevail. But does medical need mean that the person who is the sickest should get priority or the person who has the best chance of long-term survival? According to social scientists, the chance to receive an organ is positively correlated with socioeconomic status and patient's race (Kutner 1987). African Americans are less likely to be donors, organ procurers, or recipients of organs. In addition, people with history of mental, emotional, or family problems are less likely to receive organs, raising ethical questions about disability when patients with Down's Syndrome were initially refused a position on the organ donation waiting list. Economists have tried to move away from the gift giving ideology towards a market ideology that considers organs as any other commodity to be bought and sold. Several proposals circulating in medical circles propose to pay a financial benefit to the estate of the donor to motivate donation. In other countries and some US states, the governments have tried to enhance organ donation with a system of presumed consent: everyone is presumed to be an organ donor unless documented otherwise.

Anthropologist Margaret Lock's study of organ transplants and brain-death in Japan and the US (Lock 1989) offers an important counterargument against the idea of the 'technological imperative.' The technological imperative implies an inherent force in healthcare technology; once a medical technique is available there is virtually no objection possible for not using it. The concept, introduced by economist Victor Fuchs, has caught on in other studies to explain the quest for more and better technologies and the quick adoption of questionable technologies (Rothman 1997). Lock explained that any technological imperative is not inherent to technology but to the broader culture. Although Japanese physicians and scientists possess all the technical know-how and skills to implement organ transplants, Japanese people have until recently chosen not to take the cadaver organ transplant route but instead focus on the development of artificial organs. The reason for Japan's reluctance to jump the organ transplant bandwagon is that brain death is not accepted as a sufficient criterion of death for religious reasons and because Japanese families nurse their terminally ill and would be reluctant to accept that a breathing body with a heartbeat is 'dead.' In addition, Japanese have a different concept of the self and feel reluctant to just emulate every technological innovation from the West.

4.2 Reproductive Technologies

Another area flush with research attention is the field of reproductive technologies. Different infertility treatments, biotechnologies, and fetal surgeries have opened the categories of parent, child, fetus, human, family, and even population to some renegotiation. Key questions include control over the human body, the manipulation of eggs, sperm, and embryos, and the threat these technologies pose for the identity and legal status of the 'products' created. Reproductive technologies have been largely analyzed from feminist inspired perspectives but the emphasis of the analysis has shifted over time. Initially, feminist scholars looked at reproductive technology as a tool to liberate women because it could sever the link between sexuality and reproduction. A technological fix would erase some of the differences between men and women and create equal sexuality (Firestone 1970).

While reproductive possibilities seem to be more about fulfilling than rejecting the traditional female role, some feminist scholars gaze at new technologies as the next installment of patriarchal exploitation of women's bodies and reproductive processes. Reproductive technologies are viewed as attacking women, a modern incarnation of eugenics. The technologies become 'technical delegates,' tools used to compensate for the perceived deficiencies of its users. Women's bodies are opened up, manipulated, and controlled to be ultimately bypassed. The consequence is an alien-

ation, commodification, and depersonalization of women's bodies (Corea et al. 1985).

More recently, feminists have looked at the same reproductive technologies from a postmodern perspective (Casper and Clarke 1998), giving more agency to women themselves. The postmodern strategies concentrate upon the design and transformation of reproductive bodies. Poststructuralist feminists have rejected the notion that the 'real female body' is passively 'acted upon,' instead preferring to view it as being both inscribed and constituted through reproductive discursive practices and processes. Reproductive technologies generate different kinds of subjectivity. For example, Charis Cussins analyzed how undergoing infertility treatments figures in women's self making and at the same time transforming their body conceptions (Cussins 1998). She located the fertility treatment in women's biography. Before entering the infertility clinic, many women have already a history of trying not to get pregnant using anticonceptiva, followed by a period of increased awareness of what could have gone wrong while trying to conceive. Cussins showed that the objectification, naturalization, and bureaucratization of the women's bodies during pelvic exams, ultrasound, explanatory surgery, and embryology laboratory is associated with particular forms of agency. If the activities in the infertility clinic lead to pregnancy, the patient exercises agency in her active participation of the different forms of objectification. If the treatment cycle has been unsuccessful, women reported their lack of agency or insufficient attention to the specificity of the patient. The consequence of the process in which women render themselves compatible with instruments, drugs, and material surroundings is what Cussins called an 'ontological choreography' with the potential of transforming the women's long-term self-perceptions as potential mothers.

Most feminist writing, both past and recent work, is ultimately aimed at creating awareness of the reproductive technologies cost and benefits for women, a women centered reappropriation of the technologies, or to offer alternatives for social change (see the book *The New Our Bodies, Ourselves* (Boston Women's Health Book Collective 1992).

4.3 Technology and Healthcare Costs

With the explosion of healthcare costs worldwide, social scientists have also investigated the financial role of medical technologies in the healthcare system. Although there seems to be a general consensus that technology has contributed to rising healthcare raised costs, exact figures of how much cost has increased or mechanisms of cost raising are not well documented. Researchers at the US Office of Technology Assessment, for example, estimated that new technologies accounts for anywhere between one-third to three-quarters of the cost of hospital care—a range of

several billions of dollars (Banta et al. 1981). Others have estimated that healthcare technology takes up 10 to 40 percent of the increase in healthcare expenditures over time, again a very wide range (Neumann and Weinstein 1991).

Problems with calculating the cost of technology include the broad definition of healthcare technology and the scope of health expenditures. Should medical technology be limited to 'big ticket' items or include the entire technical geology? When calculating health expenditures, some researchers take only the capital cost, while other scholars combine the capital cost with more expensive operating costs and increased use of auxiliary services. Other problems include measuring cost saving and balancing health outcomes of life-saving technologies against long-term institutional expenditures. In some studies, all increases not explained by other factors are attributed to technology, while others focus on the cost of treating certain illnesses or the cost of particular technologies (Neumann and Weinstein 1991).

Relying on total healthcare costs, Frans Rutten and Gouke Bonsel note that labor intensive, low-technological interventions such as mental healthcare are two to four times as expensive as the cost of high cost technology, such as kidney dialysis and transplantation, megavolt therapy, CT tomography, cardiac surgery, *in vitro* fertilization, and picture archiving. They also found that even when one isolates technology, it is often the specialized and highly salaried personnel that take up the major part of the total cost. They conclude that 'high cost technology is expensive per item, but that its contribution to total healthcare expenditure is limited' (Rutten and Bonsel 1992). Economist Victor Fuchs calculated that technology contributed 0.6 percentage points of the 8 percent annual rate of increase in health expenditures from 1937 to 1967 (Fuchs 1972).

Pulitzer-prize winner sociologist Paul Starr has also offered a caveat to the reasoning that technology is to blame for financial healthcare woes. Big-ticket technologies such as new imaging machines, organ transplantation, intensive care units, and renal dialysis have undoubtedly brought higher costs 'But, only about a third of the higher levels of spending (in the United States), compared with Canada, reflects the expense of hospital care—and of that, only a portion is due to greater use of technology' (Starr 1994, p. 22). Starr instead suggested that we adopt a systemic view to explain the rising healthcare costs in the US. He located expanding costs in the historical development of a number of incentives for private decision makers to expand and intensify medical services.

The factors singled out to explain the importance of technology in the healthcare field relate to technological characteristics and the wider healthcare field. In the innovation literature, healthcare technologies are considered 'dynamic,' that is, they develop as they are diffused. This means that technological

innovations are regularly introduced before information about effectiveness and cost is established. Social observers also point to the billion-dollar medical industry, the public's seemingly unsatiable appetite for more technology, and varying incentive structures. Countries with stronger centralized government regulation have been able to curtail costs via planning and technology allocation. Countries following more of a market model create a situation where hospitals invest in technologies to remain competitive. The consequence of these incentives is, paradoxically, an under-use of services.

Consider the case of early detection of breast cancer through the use of mammography [in the US]. With fully utilized mammography machines, a screening mammography examination should cost no more than \$55 ... But because machines are typically used far beneath capacity, prices run double that amount. With prices so high, many women cannot afford a mammogram ... In other words, *because we have too many mammography machines, we have too little breast cancer screening* (Starr 1994, p. 25, emphasis in original)

5. Future Trends

The electronic patient record, telemedicine, artificial organs, and genetics will continue to raise the specter of healthcare technologies in the coming decades. The challenge for social scientists remains to keep up with the multitude of micro and macro changes healthcare technology poses for societies without falling into simplistic forms of technological determinism or social essentialism.

See also: Biomedical Sciences and Technology; History and Sociology; Health Care Delivery Services; Health Care: Legal Issues; History of Technology; Innovation and Technological Change; Economics of; Reproductive Medicine; Ethical Aspects; Technological Determinism; Technological Innovation; Technology and Social Control; Technology, Anthropology of; Technology, Social Construction of

Bibliography

- Anspach R 1993 *Deciding Who Lives: Fateful Choices in the Intensive Care Nursery*. University of California Press, Berkeley, CA
- Aries P 1977 *The Hour of Our Death*. Allen Lane, London
- Banta H D, Behney C J 1981 *Toward Rational Technology in Medicine: Considerations for Health Policy*. Springer, New York
- Bell S E 1989 Technology in medicine: Development, diffusion, and health policy. In: Freeman H E, Levin S (eds.) *Handbook of Medical Sociology*, 4th edn. Prentice Hall, New York, pp. 185–204
- Berg M 1997 *Rationalizing Medical Work: Decision Support Techniques and Medical Practices*. MIT Press, Cambridge, MA

- Casper M J, Clarke A E 1998 Making the pap smear into the 'right tool' for the job: Cervical cancer screening in the USA, circa 1940–95. *Social Studies of Science* **28**(2): 255–90
- Boston Women's Health Book Collective 1992 *The New Our Bodies, Ourselves*. Simon and Schuster, New York
- Corea G, Duelli R, Klein R 1985 *Man-made Women: How New Reproductive Technologies Affect Women*. Hutchinson, London
- Cussins C 1998 Ontological choreography: Agency for women in an infertility clinic. In: Berg M, Mol A (eds.) *Differences in Medicine: Unraveling Practices, Techniques and Bodies*. Duke University Press, Durham, NC, pp. 166–201
- Firestone, S 1970 *The Dialectic of Sex: The Case for Feminist Revolution*. William Morrow, New York
- Fox R C, Swazey J P 1978 *The Courage to Fail: A Social View of Organ Transplants and Dialysis*. University of Chicago Press, Chicago
- Fuchs V R (ed.) 1972 *Essays in the Economics of Health and Medical Care*. Columbia University Press, New York
- Giacomini M 1997 A change of heart and a change of mind? Technology and the redefinition of death in 1968. *Social Science and Medicine* **44**(10): 1465–82
- Glaser B G, Strauss A L 1965 *Awareness of Dying*. Aldine, Chicago
- Hirschauer S 1991 The manufacture of bodies in surgery. *Social Studies of Science* **21**: 279–319
- Illich I 1976 *Medical Nemesis*. Random House, London
- Kutner N G 1987 Issues in the application of high cost medical technology: The case of organ transplantation. *Journal of Health and Social Behavior* **28**(March): 23–36
- Latour B 1987 *Science in Action: How to Follow Scientists and Engineers through Society*. Harvard University Press, Cambridge, MA
- Lock M 1989 Reaching consensus about death: Heart transplants and cultural identity in Japan. *Society-Society* **13**(1): 15–26
- Neumann P J, Weinstein M C 1991 The diffusion of new technologies: Costs and benefits to health care. In: Gelijns A C, Halm E A (eds.) *The Changing Economics of Medical Technology*. National Academy Press, Washington, DC
- Pickering A 1995 *The Mangle of Practice: Time, Agency, and Science*. University of Chicago Press, Chicago
- Rothman D J 1997 *Beginnings Count: The Technological Imperative in American Health Care*. Oxford University Press, Oxford, UK
- Rutten F F H, Bonsel G J 1992 High cost technology in health care: A benefit or a burden? *Social Science and Medicine* **35**(4): 567–77
- Sharp L A 1995 Organ transplantation as a transformative experience: Anthropological insights into the restructuring of the self. *Medical Anthropology Quarterly* **9**(3): 357–89
- Starr P 1994 *The Logic of Health Care Reform: Why and How the President's Plan Will Work*. Whittle, New York
- Timmermans S 1999 *Sudden Death and the Myth of CPR*. Temple University Press, Philadelphia
- US Congress, Office of Technology Assessment 1982 *Strategies for Medical Technology Assessment*. US Government Printing Office, Washington, DC
- Waitzkin H 1979 A Marxian interpretation of the growth and development of coronary care technology. *American Journal of Public Health* **69**(12): 1260–8
- Winner L 1980 Do artifacts have politics? *Daedalus* **109**: 121–6

S. Timmermans

Health Economics

An individual writing an entry summarizing the field of health economics in 1960 would have had difficulty finding much to say. By contrast, an individual writing at the turn of the twenty-first century faces an embarrassment of riches: the field has exploded since the 1960s. There are now several textbooks, as well as a two-volume *Handbook of Health Economics* (Culyer and Newhouse 2000) and two international journals: the *Journal of Health Economics* and *Health Economics*. This article will thus necessarily be selective.

At least four reasons can be identified for the growth in health economics. The first is the enormous expansion of the health-care sector in almost every developed country during the second half of the twentieth century. Between 1960 and 1997 real health-care spending per person in the G-7 countries grew at annual rates between 3.7 percent (the United Kingdom) and 7.4 percent (Japan) (Anderson and Poullier 1999 give figures in nominal US dollars, and I have deflated by the US Gross Domestic Product, GDP, deflator). As a result of this growth, the G-7 countries in 1997 spent in real terms between 4 and 14 times more on health care per person than they spent in 1960. Because their economies did not grow nearly as rapidly, all the countries saw a rise in the share of GDP going to health care. By 1997 that share ranged from 6.7 percent in the UK to 13.6 percent in the USA.

Second, developments in economic theory, especially the growth of information economics, have aided the development of health economics. Third, although this is not really taken up below, the development of computational power has greatly facilitated empirical work by permitting analysis of large administrative databases. Finally, in several countries, especially the USA, the field has been generously supported when compared with most other subfields of economics, although not when compared with most natural or biomedical sciences.

1. Uncertainty, Medical Insurance, and Medical Care Services

The seminal article in health economics is generally considered to be by Arrow (1963). In explaining the institutions surrounding the delivery of medical care, Arrow emphasized two related features: the pervasiveness of uncertainty in medical care, and the asymmetric information between the physician and the patient. Uncertainty gave rise to a demand for insurance, but because insurance markets were seriously incomplete, nonmarket institutions evolved to compensate. In particular, medical ethics dictated that physicians were not to act in a narrow, profit-maximizing fashion, and the patient trusted that the physician was acting in the patient's interests. Much of what was to follow in

health economics over the next four decades flowed out of these notions.

The subsequent development of information economics in the 1970s stimulated developments in health economics; indeed, medical care was an example used in several of the early seminal articles in information economics (for example, Rothschild and Stiglitz 1976). These articles showed that if buyers in competitive insurance markets knew more about their risk than sellers, a market equilibrium would not exist and hence certain markets would not exist. One necessary condition for lack of equilibrium in the Rothschild–Stiglitz model was that the proportion of bad risks be relatively small, which is descriptively accurate about health insurance. Typically the highest five percent of spenders account for 50 percent or more of resource use (Berk and Monheit 1992). Relaxing the assumption of symmetric information between buyer and seller thus gave a theoretical rationale for Arrow's observation that insurance markets were often incomplete. It also gave a policy rationale for the public sector interventions in health insurance markets—or the direct provision of services by the public sector—that had by then occurred in every developed country.

Zeckhauser (1970) pointed out that health insurance—and risk-spreading arrangements more generally—inevitably created a second-best situation: although insurance offered protection against risk, it did so at the expense of increasing moral hazard. The traditional meaning of moral hazard in the insurance literature was failure of insured individuals to take preventive measures. In the case of medical care, that insured individuals might forgo prevention seemed of less concern because their health, including pain and suffering, was not insured. But medical care insurance typically reimbursed a certain proportion of spending rather than reimburse a lump sum contingent upon a certain event's occurring. Hence, moral hazard came to mean the additional consumption stimulated by the subsidy implied by insurance. Although the additional consumption was often construed as creating a dead-weight loss (Pauly 1968), this is not necessarily the case. One could want resources transferred to certain states of the world precisely so that one could consume medical care that would otherwise be unaffordable in those states (DeMeza 1983). As a result, the conventional test of a welfare loss from health insurance is whether a consumer, to insure a given state of the world, would purchase an actuarially fair insurance policy with an appropriate lump-sum reimbursement (Goddeeris 1984). Any willingness to pay on the part of others (i.e., an externality) should also be added.

There is, however, a literature in health economics that rejects the conventional test of a welfare loss from insurance for various reasons. Two prominent reasons given in this literature are that the observed demand curve does not reflect the preferences of an informed consumer (about which more below) or that income should not affect the delivery of medical services

(Williams and Cookson 2000). The applicability of the conventional test thus remains a controversial issue.

The understanding of how the health insurance contract itself affects health care and even health has progressed over several years. Following Zeckhauser, many early US studies focused attention on the extent of the moral hazard introduced by health insurance, or more concretely the demand response of the consumer to more or less generous insurance coverage. Indeed, the RAND Health Insurance Experiment, one of the largest social science projects to date, was in effect a controlled trial that was in part aimed at quantifying the consumer's demand response to varying levels of coverage (Newhouse and the Insurance Experiment Group 1993). (The results of this experiment are covered below.)

In addition to specifying the price to the consumer or the degree of subsidy, the health insurance contract also specifies the terms at which the insurer would reimburse the supplier. (Terms must also be specified when the government is a direct provider.) In many cases the supplier agrees to accept the insurer's terms as payment in full and does not bill the consumer an additional amount; indeed, in some instances, such as in Canada, billing additional amounts to the consumer is illegal.

Numerous studies have showed that the amount paid to the supplier, especially the physician, affects the quantity of services, holding constant the price paid by the consumer or patient (for example, Rice 1983). These studies have given rise to a lengthy debate among health economists about whether supplier-induced demand rendered standard demand theory, which postulated an exogenous set of consumer preferences, inapplicable to medical care.

The dominant empirical evidence purporting to show that supplier-induced demand violated standard theory was that the volume of health-care services often rose when physician fees were reduced. This was sometimes interpreted as physicians' delivering more services in order to maintain their income, and was referred to as the target-income hypothesis. But it subsequently was shown that this behavior could be accommodated by a standard neoclassical model, provided the physician's utility function was assumed to include total income, work effort or leisure, and the degree of inducement (the latter was assumed to be negatively valued) (McGuire and Pauly 1991). Under these conditions, if the cut in fees implied a large enough cut in income, the physician would increase inducement and services delivered would rise, but if the reduction in fees had only a modest effect on income, services would fall. (In the former case an income effect would dominate a substitution effect, and conversely in the latter case.) This modification, although potentially leaving the construct of a demand curve intact from the point of view of positive economics, left the applicability of conventional welfare economics on shaky ground.

The role of the physician as an agent for the patient gave rise to a literature that emphasized altering financial incentives facing physicians to improve the trade-off between risk spreading and moral hazard. Specifically, it was proposed to put the physician (or other supplier such as a hospital) at some risk for the quantity of services used by a group of patients (Ellis and McGuire 1986). Fee-for-service reimbursement of the physician, with fees set in an administrative transaction between the insurer and the physician, was thought to contain economic rents in many cases, with a resulting incentive for the physician to overprescribe services and procedures. For if fees were set at marginal cost there would be no incentive to overprescribe, and no concern about supplier-induced demand (Pauly 1980). Thus, the use of two-part pricing, wherein the physician (or other supplier) received a lump sum for each enrollee—a capitation—as well as a diminished fee, could in principle reduce the moral hazard from rents in fees without increasing the financial risk borne by the patient. (But some cost sharing on the demand side would still presumably be useful to address the patient's incentive to seek care in the first place.) Of course, if the consumer were well informed and could monitor the services delivered, the basis of payment to the physician, capitation, or fee-for-service, would not matter, but this was a strong assumption. Indeed, what the consumer often wanted from the physician was diagnostic information. The principle of a two-part payment could be extended to other kinds of services, for example paying a lump sum for each hospital admission, with perhaps additional payments for very expensive cases, rather than a varying amount for each admission, depending on the exact services delivered. This latter example shows that the fee portion of the payment could in principle be paid in nonlinear fashion.

The importance of pricing in affecting supplier behavior has blossomed in the USA with the development of managed care (Glied 2000). Health-maintenance organizations, which provide all necessary medical services in return for a monthly premium, had existed since the 1930s. Until the 1980s, however, they had a small market share (under 10 percent). But as medical capabilities and the resulting costs of health care rose, the scale of moral hazard and potential welfare loss of the then dominant insurance form rose concomitantly. Probably as a result, starting in the mid-1980s managed care grew to dominate the US market, in part because it was seen as a set of techniques to reduce the moral hazard without increasing the financial risk on the consumer.

Managed care consists of two broad techniques. The first is financial incentives to physicians, usually groups of physicians, to reduce services. A common form of such incentives is capitation, but there are also various kinds of risk-sharing arrangements with payment according to whether utilization is over or under a target. The second technique is command-and-

control-type regulation, such as utilization review or prior authorization of hospitalization, in order to reduce the delivery of low-valued services. Baumgardner (1991) showed how, as the technical capabilities of medicine increased, command-and-control techniques could become more valuable in addressing moral hazard.

But capitation had a downside as well. Because in practice the same capitation amount was paid for patients whose expected costs varied, there was an incentive to select good risks and avoid bad risks. This issue has been the subject of much analysis (Van de Ven and Ellis 2000). Some of the analysis stressed situations in which regulation prohibited charging different risks different amounts even when information was symmetric. Other analysis stressed asymmetric information and the attraction of sicker individuals to more generous benefits even if insurance plans made no active effort to discriminate among risks. The latter point of view gave rise to a literature on risk adjustment, meaning that an external party (government, an employer, a coalition of employers) would alter premiums facing consumers choosing among plans in accordance with the risk mix of their enrollees (Newhouse 1996, Van de Ven and Ellis 2000). After risk adjustment, plans that had a disproportionate number of bad risks would not be more expensive to the consumer than other plans.

The Rothschild–Stiglitz model of selection in insurance markets, which assumes asymmetric information, has been influential, but makes two predictions that do not correspond with experience. First, it predicts no pooling equilibrium (meaning heterogeneous risks purchase the same insurance plan). Second, it predicts that high-risk individuals can obtain the insurance they want, whereas low-risk individuals cannot. Real-world insurance markets obviously exhibit some pooling; furthermore, it is typically high-risk rather than low-risk individuals who have difficulty obtaining as much insurance as they want in competitive insurance markets. Newhouse (1996) showed that if it is costly for insurers to discriminate among varying risks or persons with varying expected costs, there will be some pooling of risks (there will be more pooling with higher costs of discriminating). Moreover, under these conditions it is high-risk rather than low-risk individuals who will not be able to buy all the insurance they wish. Although these predictions appear to accord better with reality, it is not clear that the magnitude of costs to insurers to discriminate among different risks is large enough to carry the burden assigned to them. Theoretical understanding of insurance markets remains imperfect.

If insurers (or health plans) compete for individual members, as is observed in the USA and in certain other countries (e.g., The Netherlands, Israel), competition will take place on risk type if premiums do not approximate expected cost at the individual level. In other words, if paid a uniform premium, it is in a

plan's financial interest to enroll relatively healthy persons and not to enroll the sick. A similar phenomenon can occur among physicians to the degree that they are reimbursed a fixed capitated amount per patient and are at risk for spending above that amount (e.g., primary care groups in the UK National Health Service). Selection based on expected costliness of patients implies that premium spirals may occur, in which certain plans are left with poorer and poorer risks, and ultimately may go out of existence (Cutler and Reber 1998). It also implies that prices quoted by health plans reflect the average enrollee risk in addition to any product differences that would affect cost for a given risk. In turn this means that prices facing an individual enrollee do not generally reflect the marginal cost of providing services to that enrollee, and that the allocation of persons to plans differs from that of a perfect market.

Remedying this situation by charging premiums geared to individual risk means an individual cannot avoid the financial risk of a deterioration in health if premiums are adjusted periodically, say annually, to reflect risk in the next period. For example, persons newly diagnosed with diabetes mellitus would face an increased premium and would thus bear the financial risk of contracting diabetes. In principle, the risk would be avoided if insurance policies paid a lump sum to an individual diagnosed with a disease at the time of diagnosis (Cochrane 1995). Such a feature is not observed, however, probably for the reason that health insurance tends not to be written as a fixed indemnity policy in the first place, namely that the large number of states of the world (i.e., the severity and combinations of various diseases) makes it impractical to enumerate the states. Furthermore, such insurance would fail to protect against the risk that a better but more expensive treatment will become available in the future, a common occurrence in medicine.

A monopoly public insurance program or the direct public provision of services effectively insures against the possibility of becoming a bad risk and also avoids the selection problem inherent in competitive arrangements, but potentially gives up whatever benefits in more favorable prices or quality may be available from competition. Moreover, it cannot readily cater to diverse preferences or willingness to pay for alternative treatments of varying expense. Also, even a monopoly insurance plan may leave competition among providers with which it contracts (e.g., physicians) and thereby introduce selection issues.

One middle ground between a monopoly insurer and competitive individual insurance markets is competition to insure groups of individuals who have formed for purposes other than obtaining health insurance (Diamond 1992). Provision of insurance through the place of employment, as is common in several countries, is one method of achieving this. Employer-provided insurance, however, breaks down

for small firms or the self-employed, as well as for individuals not attached to the labor market.

Another middle ground is so-called carve-out arrangements, whereby an employer or a public health insurance plan asks for competitive bids from suppliers of particular services such as mental-health services. A contract for the given service is then awarded to a single supplier, and the contract can be periodically rebid. For the service that is carved out, there is a single supplier to an entire group, but for services that are not carved out there can be competitive suppliers within the group. Carve-outs eliminate selection for the service that is carved out, but at the price of potential coordination problems among various suppliers. For example, it may be unclear whether a visit by a depressed patient who is complaining of fatigue should be charged to the supplier of carved-out mental-health services or the health plan that is responsible for all other health services.

Currently there is much interest in the consequences of different methods of reimbursing suppliers, especially physicians. The interest stems in part from the wide variety of methods used around the world, including fee-for-service, capitation (a fixed payment per period of time for 'necessary' services), and salary. Similarly, hospitals may be paid a fee for each detailed service, an all-inclusive fee per day, or an all-inclusive fee per admission. Several countries have changed their method of reimbursement (e.g., greater use of capitation in the USA). The consequences of different methods of paying providers will almost certainly be one of the main topics of research over the next several years.

2. Health

In 1972 Grossman developed a model that construed the demand for medical care services as derived from the demand for health, which he treated partly as an investment good and partly as a consumption good; for an exposition of the model and subsequent developments, see Grossman (2000). As an investment good, health care was assumed to increase the stock of health and thereby the flow of healthy days and productivity; as a consumption good, health care was assumed to improve well-being. Health, of course, was not traded in the market, but factors such as age and schooling altered the optimal level of health by altering its shadow price. Grossman's model, by emphasizing health rather than medical care as what was desired, was consistent with a stream of empirical work that sought to measure clinical outcomes of health care and did not simply take spending on medical services as equivalent to output. Starting from a different position, Sen (1992) has also emphasized the importance of health as a good in its own right.

The importance of health has given rise to two types of literatures in health economics, one on cost-effective

interventions to change health and the other on the determinants of health. One strand of literature evaluates various clinical interventions according to conventional economic criteria. Valuing health outcomes in dollar terms, however, is controversial; as a result, alternative methods for measuring benefits that are independent of willingness to pay have evolved. Two prominent methods are Quality Adjusted Life Years (QALYs) and Disability Adjusted Life Years (DALYs). These methods attempt to attach utility weights to various states of the world in order to compare them. In effect, they are measures of life expectancy adjusted for quality of life. For explications of methods and issues refer to Garber (2000).

It was natural for economists to study how additional medical services affected health outcomes. The RAND Health Insurance Experiment showed that free medical care induced participants to consume some 30–40 percent more services relative to a plan with a large deductible, but the additional services did little or nothing to affect the average person's health (Newhouse and the Insurance Experiment Group 1993). Such a finding was consistent with the prediction of a standard economic model, namely that ever-greater subsidies would lead to the consumption of ever-lower valued services. That additional services could have a small marginal effect on health, of course, is consistent with their having a large average effect; reducing medical services to zero could have catastrophic consequences.

Nonetheless, closer analysis of the experimental data suggested that the allocation of services by person did not fit the standard prediction. Fully insured consumers were more likely than those with cost sharing to receive services of substantial clinical value, which should have left them on average in better health, but they also were more likely to receive services of negative value (e.g., antibiotics for viral conditions). On average, the additional positive and negative services consumed by those with free care appeared roughly to offset, although for the poor who were chronically ill the additional services when care was free appeared to have a net benefit. In this group, of course, the prior odds that a given service would be beneficial rather than harmful was considerably greater.

The apparent misallocation of services to persons is consistent with one of the best established facts in health economics: there is enormous variation across areas—both small and large—in the delivery of various medical services, even conditioning on diagnosis (Phelps 2000). For example, rates of specific procedures per Medicare beneficiary in the USA generally varied by more than a factor of two from high to low across 13 large areas in 1981 (Chassin et al. 1986). And areas that had high rates for one particular procedure did not necessarily have high rates for another. Moreover, the percentage of 'appropriate' services in the various areas was roughly constant, rather than

declining in the areas with higher rates, and there were seemingly high rates of zero or negatively valued services in all areas, one-sixth to one-third for several procedures (Chassin et al. 1987). In other words, allocation of services to persons did not appear to be following the principle of maximizing health benefits in the population. Further studies have revealed high rates of zero or negatively valued services in other health systems as well (Brook 1993), which implies that the efficiency of health-care delivery worldwide may be well within the frontier.

In addition to the effect of medical services on health, economists have studied other determinants of health, in particular the consequences of various personal behaviors. The use of tobacco has probably been the most studied (Chaloupka and Warner 2000). Abuse of alcohol has also been studied intensively (Cook and Moore 2000). For both tobacco and alcohol, economists have shown that higher taxes reduce use substantially. In the USA, however, cigarette taxes are already at a level that roughly offsets the average and probably the marginal externality of smoking. By contrast, alcohol taxes are well below the average and probably the marginal externality created by excessive drinking (Manning et al. 1991).

Recently there have been a number of studies of the potential effect of income inequality on health outcomes, although most of these have been carried out by noneconomists. Many studies purport to show that increased inequality, controlling for the level of income, damages health, but there are issues as to which way causality runs and whether any inferences can be drawn from the aggregate data that many studies have employed. A useful review and critique is found in Smith (1999).

3. Policy Considerations and the Future

Health economists in several countries have been active in policy debates over financing medical services and measures to promote health. In some European countries, especially the UK, an important thrust has been to break apart the government role as both buyer and seller of services in order to promote competition on price among sellers. Similarly in the USA the spread of managed competition has been an effort to promote price competition among sellers of medical services (Enthoven 1988). The essence of managed competition is that a consumer should bear much or all of the incremental cost of choosing one health plan rather than another, which in turn motivates the health plan to seek economical suppliers of care.

The competition needs to be managed, however, so that it does not degenerate into a competition for good risks. As mentioned above, the 'manager' could be an employer, a coalition of employers, or government; in fact, all three are observed. Current techniques for

management or risk adjustment, however, are imperfect, explaining something less than half of the systematic or predictable variance across individuals in annual spending (Newhouse 1996, Van de Ven and Ellis 2000).

Although institutional arrangements in all countries attempt to control moral hazard, the efforts are imperfect for several reasons. First, medical knowledge is continually changing, so that what was thought to be the treatment of choice today may be discredited tomorrow. And sometimes the opposite happens as well. Because the spread of knowledge is imperfect, physicians may unwittingly be delivering a service that is known to be no longer effective or which is dominated by a newer service. Second, any efforts the physician makes to elicit the patient's preferences among alternative treatments may be imperfect. In extreme cases they may be futile, as in the case of a demented patient. As a result, the physician may give the patient a service a more informed patient would not have wanted or vice versa. Third, there is often lack of consensus about how health-care services should be rationed and in particular what role, if any, income should play (Garber 2000, Williams and Cookson 2000). Indeed, the advent of managed care in the USA and its effort at controlling moral hazard has led to a demand for 'patient protection.' In particular, under managed care arrangements patients have no ready way at the point of service to express the intensity of their preference or demand for a service. As a result, it is certainly possible that rationing would occur among highly valued services. Fourth, the degree to which others are willing to pay for an individual's treatment is variable across societies and across individuals. Improving methods for rationing services is likely to remain on many policy agendas for a long time.

One of the features of the late twentieth and early twenty-first century has been the aging of most developed countries. Several factors are contributing to this trend including increased life expectancy among those over 65, the post-World War II baby boom, and the ensuing baby bust. The result will be an increased demand for medical care (those over 65 typically spend from three to five times as much on medical care as those under 65) and an increased demand for personal care services as individuals become unable to live independently without assistance. Much of this demand will be financed from public sources, but the decline in fertility in developed countries means there will be more retired elderly per worker in most countries, and hence a greater tax burden for workers.

Historically, as noted above, real per capita medical care costs have gone up four percent or five percent per year in many countries over several decades. It is now generally accepted that this increase is mainly due to technological change or the increased capabilities of medicine (Newhouse 1992, Fuchs 1996). If this rate of increase in cost continues, medical care, much of

which is publicly financed, will consume an ever greater share of GDP. There is, of course, nothing inherently wrong with this, provided the benefits justify the costs, although the presence of insurance raises a question about whether, at the margin, the benefits do justify the costs. To the degree that the growth in costs is attributable to new capabilities, the issue is willingness to pay for those capabilities, both for oneself and others. Given the similar rates of increase in medical costs across countries with dissimilar institutions, there is some presumption that much of the increase is desired.

Continued biomedical research, especially the revolution in molecular biology and the mapping of the human genome, promise further technological change in medicine. Unfortunately, the implications of the change for cost are highly uncertain. Consequently, the future of health care and what of the increased menu of valuable services various societies will wish to purchase remains highly uncertain.

See also: Health Care Delivery Services; Health Care: Legal Issues; Health Care Markets: Theory and Practice; Health Care Organizations: For-profit and Nonprofit; Health Insurance: Economic and Risk Aspects; Health Policy; Health Promotion in Schools; Health Surveys; Medical Geography; Medical Sociology; Occupational Health and Safety, Regulation of; Welfare State; Well-being and Burnout in the Workplace, Psychology of

Bibliography

- Anderson G F, Poullier J-P 1999 Health spending, access, and outcomes: Trends in industrialized countries. *Health Affairs* **18**: 178-92
- Arrow K J 1963 Uncertainty and the welfare economics of medical care. *American Economic Review* **53**: 941-73
- Baumgardner J 1991 The interaction between forms of insurance contract and types of technical change in medical care. *RAND Journal of Economics* **22**: 36-53
- Berk M L, Monheit A C 1992 The concentration of health expenditures: An update. *Health Affairs* **11**: 145-9
- Brook R H 1993 Maintaining hospital quality: The need for international co-operation. *Journal of the American Medical Association* **270**: 985-7
- Chaloupka F, Warner K E 2000 The economics of smoking. In: Culyer J, Newhouse J P (eds.) *Handbook of Health Economics*. North-Holland, Amsterdam, pp. 1539-627
- Chassin M R, Brook R H, Park R E, Keesey J, Fink A, Kosecoff J, Kahn K, Merrick N, Solomon D H 1986 Variations in the use of medical and surgical services by the medicare population. *New England Journal of Medicine* **314**: 285-90
- Chassin M R, Kosecoff J, Winslow C M, Kahn K L, Merrick N J, Keesey J, Fink A, Solomon D H, Brook R H 1987 Does inappropriate use explain geographic variation in the use of health care services? A study of three procedures. *Journal of the American Medical Association* **258**: 2533-7

- Cochrane J H 1995 Time-consistent health insurance. *Journal of Political Economy* : 445-73
- Cook P, Moore M 2000 Alcohol. In: Culyer J, Newhouse J P (eds.) *Handbook of Health Economics*. North-Holland, Amsterdam, pp. 1629-73
- Culyer J, Newhouse J P (eds.) 2000 *Handbook of Health Economics*. North-Holland, Amsterdam
- Cutler D M, Reber S J 1998 Paying for health insurance: The tradeoff between competition and adverse selection. *Quarterly Journal of Economics* **113**: 433-66
- DeMeza D 1983 Health insurance and the demand for medical care. *Journal of Health Economics* **2**: 47-54
- Diamond P 1992 Organizing the health insurance market. *Econometrica* **60**: 1233-54
- Ellis R P, McGuire T G 1986 Provider behavior under prospective reimbursement. *Journal of Health Economics* **5**: 129-51
- Enthoven A C 1988 *Theory and Practice of Managed Care Competition in Health Care Finance*. North-Holland, Amsterdam
- Fuchs V R 1996 Economics, values, and health care reform. *American Economic Review* **86**: 1-24
- Garber A 2000 Recent developments in CBA/CEA. In: Culyer J, Newhouse J P (eds.) *Handbook of Health Economics*. North-Holland, Amsterdam, pp. 181-221
- Glied S A 2000 Managed care. In: Culyer J, Newhouse J P (eds.) *Handbook of Health Economics*. North-Holland, Amsterdam, pp. 707-53
- Goddeeris J H 1984 Medical insurance, technological change, and welfare. *Economic Inquiry* **22**: 56-7
- Grossman M 2000 The human capital model. In: Culyer J, Newhouse J P (eds.) *Handbook of Health Economics*. North-Holland, Amsterdam, pp. 347-408
- Manning W G, Keeler E B, Newhouse J P, Wasserman J, Sloss E 1991 *The Cost of Poor Health Habits*. Harvard University Press, Cambridge, MA
- McGuire T G, Pauly M V 1991 Physician response to fee changes with multiple payers. *Journal of Health Economics* **10**: 385-410
- Newhouse J P 1992 Medical care costs: How much welfare loss? *Journal of Economic Perspectives* **6**: 3-21
- Newhouse J P 1996 Reimbursing health plans and health providers: efficiency in production versus selection. *Journal of Economic Literature* **34**: 1236-63
- Newhouse J P, Insurance Experiment Group 1993 *Free for All? Lessons from the RAND Health Insurance Experiment*. Harvard University Press, Cambridge, MA
- Pauly M V 1968 The economics of moral hazard: Comment. *American Economic Review* **58**: 531-7
- Pauly M V 1980 *Doctors and Their Workshops*. University of Chicago Press, Chicago
- Phelps C E 2000 Information diffusion and best practice adoption. In: Culyer J, Newhouse J P (eds.) *Handbook of Health Economics*. North-Holland, Amsterdam, pp. 223-64
- Rice T 1983 The impact of changing Medicare reimbursement rates on physician-induced demand. *Medical Care* **21**: 803-15
- Rothschild M, Stiglitz J 1976 Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics* **90**: 629-50
- Sen A K 1992 *Inequality Re-examined*. Harvard University Press, Cambridge, MA
- Smith J P 1999 Healthy bodies and thick wallets: The dual relation between health and economic status. *Journal of Economic Perspectives* **13**: 145-66
- Van de Ven W P M M, Ellis R P 2000 Risk adjustment in competitive health plan markets. In: Culyer J, Newhouse J P

- (eds.) *Handbook of Health Economics*. North-Holland, Amsterdam, pp. 755–845
- Williams A, Cookson R 2000 Equity in health. In: Culyer J, Newhouse J P (eds.) *Handbook of Health Economics*. North-Holland, Amsterdam, pp. 1863–1910
- Zeckhauser R J 1970 Medical insurance: A Case study of the tradeoff between risk spreading and appropriate incentives. *Journal of Economic Theory* 2: 10–26

J. P. Newhouse

Health Education and Health Promotion

Health education is any combination of learning experiences designed to facilitate voluntary actions conducive to health (Green and Kreuter 1999). Health promotion is the combination of educational and environmental supports for actions and conditions of living conducive to health (Green and Kreuter 1999), thereby including health education. The two definitions represent a historical development in approach from more individual to more ecological where the role of the environment has acquired increased relevance in understanding and changing conditions for health. For this reason, we will use the term health promotion in the remainder of this article. Health promotion can be characterized by four other main developments: the need for planning, the importance of evaluation, the use of social and behavioral science theories, and the systematic application of evidence and theories in the development of health promotion interventions. Finally, we will describe recent developments in information technology (IT) and their effect on health promotion.

1. Health, Environment, and Behavior

In the ecological approach to health promotion, health is viewed as a function of individuals and the environments in which individuals are embedded, including family, social networks, organizations, and community and public policies. The relation between social economic status and health is a clear example of environmental influences on health. At the same time it also constitutes a major challenge. The central concern of health promotion is health behavior. However, health behavior refers not only to the individual's behavior but also to the behavior or actions of groups and organizations. Stress at work may be related to individual coping behavior, but also to managers' decision-making behavior (organization). Richard et al. (1996) describe the various environmental levels as embedded systems. They indicate that individuals exist within groups, which are in turn embedded within organizations and higher order systems. The individual is influenced by, and can influence directly or through groups and organiz-

ations, the higher order systems. The picture that emerges is a complex web of causation as well as a rich context for interventions. In the stress example, the individual as well as the manager will both be targets for health promotion interventions. Moreover, at the society level, the intervention may target politicians' decision-making related to a healthier organization of labor. We see managers and politicians as agents in the environment who serve as targets for health promotion interventions (Bartholomew et al. 2000).

2. Health Promotion Planning

Health promotion is a planned activity. The most widely used health promotion planning framework is Green and Kreuter's (1999) PRECEDE/PROCEED model (see Fig. 1).

The model begins on the right with the assessment of quality of life and health problems, changes in which should be the proposed outcomes of a health promotion intervention. It then guides the planner to assess the behavioral and environmental causes. In the behavioral assessment typically one asks what the individuals at risk are doing that increases their risk of death from the health problem. In the environmental assessment we ask what factors in the environment are related to the health problem directly or to its behavioral causes. In the previous subsection, it was explained how the environmental causes can be viewed as behaviors or actions at various environmental levels: groups, organizations, communities, and society.

In the next phase of PRECEDE/PROCEED, the determinants of the behavioral and environmental factors are assessed. Green and Kreuter (1999) describe determinants affecting behavior as predisposing, reinforcing, and enabling. Predisposing factors relate to the motivation of an individual or a group to act. Notice that the target may also be an agent in one of the environmental levels, such as a manager or a politician. These factors mostly fall into the psychological domain and include the cognitive and affective dimensions of knowing, feeling, believing, valuing. Reinforcing factors are those consequences of action that determine whether the actor receives positive (or negative) feedback and is supported socially after it occurs. Reinforcing factors thus include social support, peer influences, and advice and feedback by healthcare providers. Notice that the perception of social support would be a predisposing factor. Reinforcing factors also include physical consequences of the behavior such as well-being or pain. Enabling factors, often conditions of the environment, facilitate the performance of an action by individuals, groups, or organizations. Factors included are the availability, accessibility, and affordability of healthcare and community resources. Enabling factors also include new skills that a person, organization, or community needs to learn to carry out a behavioral or environmental

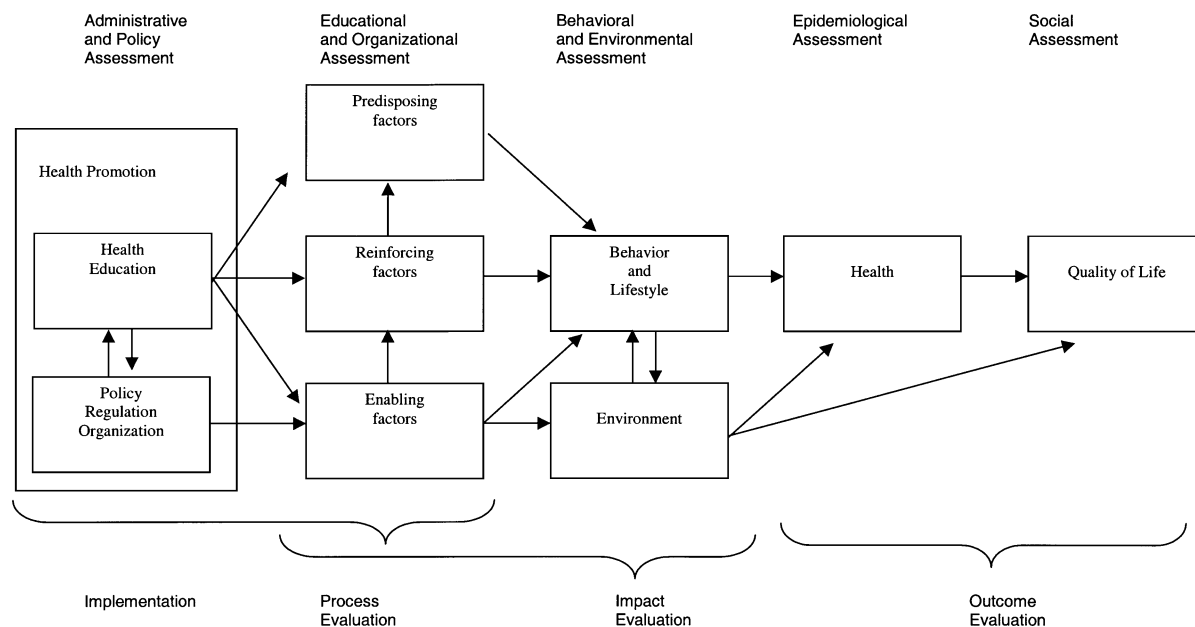


Figure 1
PRECEDE/PROCEED model (Green and Kreuter 1999)

change. Notice that self-efficacy expectations are a predisposing factor, the actual skills an enabling factor.

PRECEDE/PROCEED suggests that program planners start with the assessment of quality of life and health problems. In practice, that is a rare situation. Often health educators start somewhere within the model, depending on their task description or the specific moment in the ongoing health promotion planning process. For example, a health educator may be hired to develop a smoking prevention program for schools. We then enter the model at the behavior box. However, we first want to go back, to the right, and (re-)assess the quality of life and health problems as well as the behavioral and environmental conditions before we go on to the assessment of determinants of the behavior and environmental conditions and the development of the program.

The next phases of PRECEDE/PROCEED involve the administrative and policy assessment. The planner will now develop the plans into health education and other health promotion interventions, such as policy, regulation, or organization. The assessments include identifying barriers to overcome in implementing the program and policies that can be used to support the program. We then proceed to program implementation. Implementation is an essential element in the planning process. Without implementation, the intervention will not have any impact on determinants, behaviors, or health. The final phases of PRECEDE/PROCEED describe the evaluation levels: process,

impact, and outcome. Thinking about implementation and evaluation has to start as early as possible, not after the intervention development. In the Intervention Mapping protocol, anticipating implementation and evaluation are steps in the development process.

3. Program Evaluation

Even the best planned health promotion intervention may not have much effect. It is important to always evaluate the effect of an intervention, its impact and process. When the program has failed, we need to know why it failed. Rossi et al. (1999) suggest four levels of evaluation:

(a) *Program conceptualization and design* refers to the planning process and the use of theory and evidence in the development of the program.

(b) *Program monitoring* provides feedback on the actual implementation of the program and its reception by the target population.

(c) *Impact evaluation* is defined by Rossi et al. as the extent to which a program causes change in the desired direction. They use the word impact to make clear that the desired effect of a program may be defined at different levels, and often not at the health and quality of life level. It is useful to agree on the operationalizations of desired effects before the actual implementation of the program.

(d) *Program efficiency* evaluates the program in terms of costs and effects, such as cost-benefits and cost-effectiveness analyses.

Rossi et al. are very outspoken on when to evaluate and when not to evaluate: 'Clearly, it would be a waste of time, effort and resources to estimate the impact of a program that lacks measurable goals or that has not been properly implemented.'

Effect evaluation relies heavily on quasi-experimental designs (Rossi et al. 1999). The basic principle in effect evaluation is the random assignment of participants to the program group and a control group. Only then can we give a straightforward answer to the question if an intervention was successful. In health education practice the random assignment of individuals is often impossible. For instance, students from secondary schools cannot be assigned randomly to a school program or a control program, because most school programs are school-wide or at least class-wide. In that case, we ask various schools if they are willing to participate in the program, and we then assign schools randomly to the program condition or the control condition. A variety of quasi-experimental designs allow us to compare two or more groups that are as similar as possible. We have to expect that the groups are not completely equivalent on a number of relevant characteristics, and we cannot even assume that we know all the relevant characteristics. Statistically, we can control for most of these differences and levels, but only when we measure them before the program starts (see Bartholomew et al. 2000).

4. Social and Behavioral Science Theories

A health promotion program is most likely to benefit participants and the community when it is guided by social and behavioral science theories of health behavior and health behavior change (Glanz et al. 1997). Theory-driven health promotion programs require an understanding of the components of the theory as well as the operational or practical forms of these theories. Finding and applying relevant theories is a professional skill that health educators have to master (Bartholomew et al. 2000).

Notice that we assume that all theories are potentially applicable to all levels and also to adoption and implementation. The Theory of Planned Behavior, for example, is often applied to individual health behavior (Godin and Kok 1996) but also to politicians' behavior (Flynn et al. 1998) and to implementers' behavior (Paulussen et al. 1995).

The behavioral science theories that try to explain behavior and behavior change have two types of roots: health and health promotion in particular or behavior and behavior change in general (the last mainly being social psychological theories). Health and health promotion oriented theories are often related to perceptions of health risks, for example, the Health Belief Model (Sheeran and Abraham 1996, Strecher and Rosenstock 1997) or Protection Motivation Theory (Boer and Seydel 1996). Some other theories were developed in a health setting, but have evolved into a

general theory such as the Transtheoretical Model of Stages of Change (Prochaska et al. 1997) or Relapse Prevention Theory (Marlatt and Gordon 1985). Most general social psychological models were developed for a broad range of behaviors, but are easily applicable to health behavior and change. Some examples are Learning Theories (Westen 1996), Information Processing Theory (Hamilton and Ghatala 1994), Theory of Planned Behavior (Connor and Sparks 1996, Montano et al. 1997), Social Cognitive Theory (Bandura 1997), Goal Setting (Strecher et al. 1995), Attribution Theory (Weiner 1986), Self-regulatory theories (Clark and Zimmerman 1990), Social Networks and Social Support theories (Heaney and Israel 1997), Persuasive-communication Model (McGuire 1985), and Diffusion of Innovations Theory (Rogers 1995).

At the organizational level, health educators draw from various Organizational Change theories (Goodman et al. 1997). Typically we use constructs from these theories for considering policy development within organizations and for promoting program adoption and implementation.

At the community and society level, theoretical input can be found in the classic models of community organization and community development and the current perspectives being used for health promotion (Minkler 1997). Issues of power, participation, and goals, for instance, have received much discussion among health educators in recent years. In selecting the constructs to use for community change, planners must clarify their assumptions and values about the nature of the change process and select and implement strategies congruent with these. Finally, there are theories of policy making that have been used primarily at the national, state, and governmental level, such as Policy Window Theory (Kingdon 1995).

Theories are very important tools for professionals in health education and promotion. On the one hand, theories have become available to health promotion practice through textbooks such as Glanz et al. (1997) or Connor and Norman (1996). On the other, the application of theory has long been a challenge, for researchers as well as practitioners. Students of health promotion learn of theories and learn how to apply theories to well-selected practical problems. However, in real life the order is reversed: the problem is given and the practitioner has to find theories that may be helpful for better understanding or changing behavior (Kok et al. 1996). Recently, a protocol was published that describes a process for developing theory-based and evidence-based health education programs: Intervention Mapping (Bartholomew et al. 2000).

5. Intervention Mapping

Intervention Mapping distinguishes five steps: define proximal program objectives, select theoretical methods and practical strategies, design the program,

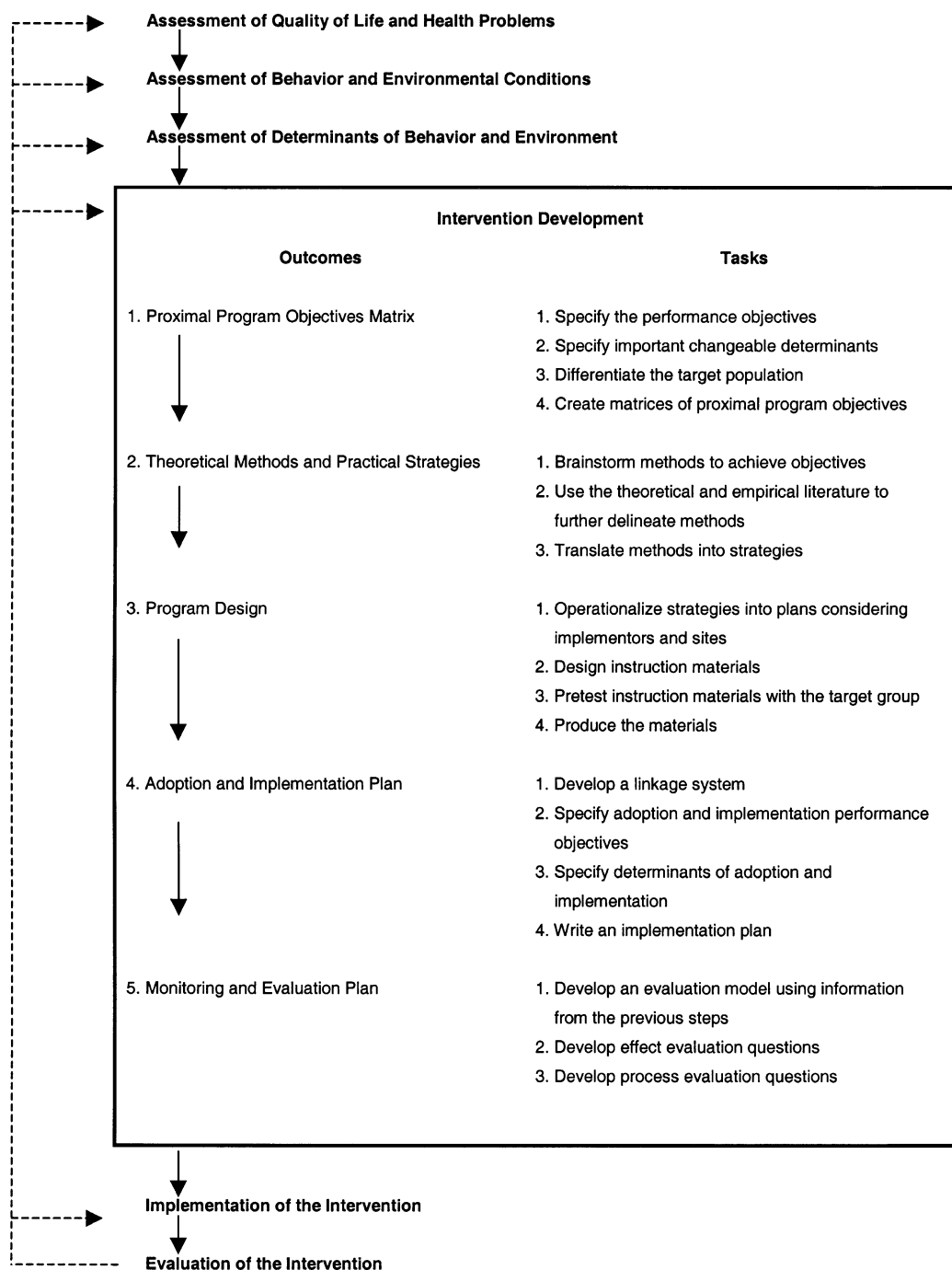


Figure 2
Intervention Mapping (Bartholomew et al. 2000)

anticipate adoption and implementation, and anticipate process and effect evaluation (see Fig. 2). Bartholomew et al. (2000) see planning as an iterative

process: two steps forward and one step back. Bartholomew et al. (2000) describe three core processes for Intervention Mapping, tools for the professional

health promoter: searching the literature for empirical findings, accessing and using theory, and collecting and using new data. From the literature search, a provisional list of answers is developed, which is often not adequate in finding solutions for the problem. The planner must go further to search for theory, using three approaches: issue, concept, and general theories approach. The issue approach searches the literature for theoretical perspectives on the issue. The concept approach begins with the concepts in the provisional answers, linking these concepts to theoretical constructs and theories that may be useful. The general theory approach considers general theories that may be applicable. Finally, it is important to identify gaps in the information obtained and collect new data to fill these gaps.

The first step in Intervention Mapping is the development of proximal program objectives, a crossing of performance objectives, determinants, and target groups. For instance, one proximal program objective for an HIV-prevention program in schools would be: 'adolescents (target population) express their confidence (determinant) in successfully negotiating with the partner about condom use (performance objective).' Performance objectives are the specific behaviors that we want the target group (or the environmental agents) to 'do' as a result of the program. For example, in the case of HIV prevention: buy condoms, have them with you, negotiate with the partner, use them correctly, and keep using them (Bartholomew et al. 2000). Determinants of behavior can be personal or external. Personal determinants are, for instance, outcome expectations, social influences, and self-efficacy expectations. External determinants are, for instance, social norms and support, and barriers. Target groups can be subgroups of the total group, for instance, men/women, people in different stages of change. Proximal program objectives are numerous and should be ordered by determinant. So we end this step with a series of lists, for instance, all proximal program objectives which have to do with skills training.

Intervention Mapping Step 2 is the selection of theoretical methods and practical strategies. The theoretical method is the technique derived from theory and research to realize the proximal program objective; the strategy is the practical application of that method. For instance, the method for self-efficacy improvement could be modeling and the strategy could be peer modeling by video. An important task in this step is to identify the conditions or parameters that limit the effectiveness of theoretical models, such as identification, observability, and reinforcement as necessary conditions for the effectiveness of learning by modeling.

Intervention Mapping Step 3 is the actual designing of the program, organizing the strategies into a deliverable program, and producing and pretesting the materials. In the example of HIV-prevention in

schools, the program comprised five lessons, an interactive video, a brochure for students, and a workbook for teachers (Bartholomew et al. 2000).

A solid diffusion process is vital to ensure program success. So, in Intervention Mapping Step 4, a plan is developed for systematic implementation of the program. The first thing to do, actually at the start of intervention development, is the development of a linkage system, linking program developers with program users. Then, an intervention is developed to promote adoption and implementation of the program by the intended program users. It may be clear that the anticipation of implementation is a relevant process from the very beginning of planning, not only at the end.

Finally, Intervention Mapping Step 5 focuses on process and effect evaluation. Again, this process is relevant from the start, not only at the end. For instance, 'adolescents express their confidence in successfully negotiating with the partner about condom use' is an objective, but is also a measure of that objective, that can be asked in pre- and postinterviews with experimental and control group subjects.

6. Recent Developments in IT

Mass media health promotion is able to reach large numbers of people, but a major weakness of mass media programs is their limited effect in terms of impact and outcome. However, more effective individualized health promotion is costly and time-consuming. Computer technology may help break several important barriers in health promotion related to individualization and program delivery. During the 1990s, health promotion researchers have been testing simple computer expert systems that enable the tailoring of communications to match certain characteristics that have been measured in individuals in the target group (Velicer et al. 1993, Brug et al. 1998, Dijkstra and De Vries 1999).

Velicer et al. define an expert system as a collection of facts and rules about something and a way of making inferences from the facts and the rules. The most common types of expert systems in health education are computer programs that are tailored to specific characteristics of the receiver. Thus, the expert system comprises one or more databases of messages based on theoretical constructs that vary as they apply to different characteristics of individuals, and on algorithms for matching the messages to individuals. The message channel could in general be anything that facilitates delivery of the message to the target: a report, a letter, computer-assisted instruction or any other channel. In the work of Velicer et al. (1993), messages were based on the Transtheoretical Model of Stages of Change and included processes of change tailored to the stage of the individual in regard to quitting smoking. Feedback included current status

and stage of change, current use of change processes, suggested strategies, and high-risk situations. Feedback was compared against a normative database as well as against participants' progress. All examples of computer-tailored health promotion in the literature are based on a similar configuration comprising (Brug et al. 1998):

(a) A theoretical framework and specification of relevant hypothesized determinants of the health behavior.

(b) Use of the determinant model to create a data collection tool and a series of messages that are tailored to these determinants.

(c) Several databases including, at least, a determinants file and a delivery message file.

(d) A set of decision rules and a tailoring program that executes these decisions.

(e) The actual communication messages.

(f) Delivery vehicles, such as a printed letter.

It is to be expected that computer tailoring will become more popular in the future and will also be applied using the Internet.

7. Conclusion

Health education and health promotion are planned activities. PRECEDE/PROCEED provides a very useful framework for the systematic planning of health promotion and Intervention Mapping provides a very useful protocol for evidence-based intervention development. Both models acknowledge the role of the individual as well as the relevant ecological levels in the environment.

See also: Health Behavior: Psychosocial Theories; Health Behaviors; Health Interventions: Community-based; Health Promotion in Schools

Bibliography

- Bandura A 1997 *Self-Efficacy: The Exercise of Control*. Freeman, New York
- Bartholomew L K, Parcel G, Kok G, Gottlieb N 2000 *Intervention Mapping: Designing Theory and Evidence-based Health Promotion Programs*. Mayfield, Mountain View, CA
- Boer H, Seydel E R 1996 Protection motivation theory. In: Connor M, Norman P (eds.) *Predicting Health Behaviour: Research and Practice with Social Cognitive Models*. Open University Press, Buckingham, UK
- Brug J, Glanz K, Van Assema P, Kok G, Van Breukelen G P 1998 The impact of computer-tailored feedback and iterative feedback on fat, fruit, and vegetable intake. *Health Education and Behavior* 25: 517–31
- Clark N M, Zimmerman B J 1990 A social cognitive view of self-regulated learning about health. *Health Education Research* 3: 371–9

- Connor M, Norman P (eds.) 1996 *Predicting Health Behaviour: Research and Practice with Social Cognitive Models*. Open University Press, Buckingham, UK
- Connor M, Sparks P 1996 The theory of planned behavior and health behaviors. In: Connor M, Norman P (eds.) *Predicting Health Behaviour: Research and Practice with Social Cognitive Models*. Open University Press, Buckingham, UK, pp. 121–62
- Dijkstra A, De Vries H 1999 The development of computer-generated tailored interventions. *Patient Education and Counseling* 36: 193–203
- Flynn B S, Goldstein A O, Solomon L J, Bauman K E, Gottlieb N H, Cohen J E, Munger M C, Dana G S 1998 Predictors of state legislators' intentions to vote for cigarette tax increases. *Prevention Medicine* 2: 157–65
- Glanz K, Lewis F M, Rimer B K 1997 *Health Behaviour and Health Education: Theory Research and Practice*, 2nd edn. Jossey Bass, San Francisco, CA
- Godin G, Kok G 1996 The theory of planned behavior: A review of its applications to health related problems. *American Journal of Health Promotion* 11: 87–98
- Goodman R M, Steckler A, Kegler M C 1997 Mobilizing organizations for health enhancement: Theories of organizational change. In: Glanz K, Lewis F M, Rimer B K (eds.) *Health Behavior and Health Education: Theory, Research and Practice*, 2nd edn. Jossey Bass, San Francisco, CA, pp. 287–312
- Green L W, Kreuter M W 1999 *Health Promotion Planning: An Educational and Ecological Approach*, 3rd edn. Mayfield, Mountain View, CA
- Hamilton R, Ghatala E 1994 *Learning and Instruction*. McGraw Hill, New York
- Heaney C A, Israel A 1997 Social networks and social support. In: Glanz K, Lewis F M, Rimer B K (eds.) *Health Behavior and Health Education: Theory, Research and Practice*, 2nd edn. Jossey Bass, San Francisco, CA, pp. 179–205
- Kingdon J 1995 *Agendas, Alternatives and Public Policies*, 2nd edn. Harper-Collins, New York
- Kok G, Schaalma H, De Vries H, Parcel G, Paulussen T H 1996 Social psychology and health education. In: Stroebe W, Hewstone M (eds.) *European Review of Social Psychology*. Wiley, Chichester, UK, Vol. 7, pp. 241–82
- Marlatt G A, Gordon J R 1985 *Relapse Prevention: Maintenance Strategies in the Treatment of Addictive Behaviours*. Guilford Press, New York
- McGuire W J 1985 Attitudes and attitude change. In: Lindzey G, Aronson E (eds.) *The Handbook of Social Psychology*, Vol. 2. Random House, New York, pp. 233–346
- Minkler M (ed.) 1997 *Community Organizing and Community Building for Health*. Rutgers University Press, New Brunswick, NJ, pp. 244–58
- Montano D E, Kasprzyk D, Taplin S H 1997 The theory of reasoned action and the theory of planned behavior. In: Glanz K, Lewis F M, Rimer B K (eds.) *Health Behavior and Health Education: Theory, Research and Practice*, 2nd edn. Jossey Bass, San Francisco, CA, pp. 85–112
- Paulussen T, Kok G, Schaalma H P, Parcel G S 1995 Diffusion of AIDS curricula among Dutch secondary school teachers. *Health Education Quarterly* 22: 227–43
- Prochaska J O, Redding C A, Evers K E 1997 The trans-theoretical model and stages of change. In: Glanz K, Lewis F M, Rimer B K (eds.) *Health Behavior and Health Education: Theory, Research and Practice*, 2nd edn. Jossey Bass, San Francisco, CA, pp. 60–84

- Richard L, Potvin L, Kishchuk N, Prlic H, Green L W 1996 Assessment of the integration of the ecological approach in health promotion programs. *American Journal of Health Promotion* **10**: 270–81
- Rogers E M 1995 *Diffusion of Innovations*, 4th edn. Free Press, New York
- Rossi P H, Freeman H E, Lipsey M W 1999 *Evaluation: A Systematic Approach*, 6th edn. Sage, Newbury Park, CA
- Sheeran P, Abraham C 1996 The health belief model. In: Connor M, Norman P (eds.) *Predicting Health Behaviour: Research and Practice with Social Cognition Models*. Open University Press, Buckingham, UK, pp. 23–61
- Strecher V J, Rosenstock I M 1997 The health belief model. In: Glanz K, Lewis F M, Rimer B K (eds.) *Health Behavior and Health Education: Theory, Research, and Practice*. Jossey Bass, San Francisco, CA, pp. 41–59
- Strecher V J, Seijts G H, Kok G, Latham G P, Glasgow R, DeVellis B, Meertens, R M, Bulger D W 1995 Goal setting as a strategy for health behavior change. *Health Education Quarterly* **22**: 190–200
- Velicer W F, Prochaska J O, Bellis J B, DiClemente C C, Rossi J S, Fava J L, Steiger J H 1993 An expert system intervention for smoking cessation. *Addictive Behaviors* **18**: 269–90
- Weiner B 1986 *An Attributional Theory of Motivation and Emotion*. Springer, New York
- Westen D 1996 *Psychology: Mind, Brain, and Culture*. Wiley, New York

G. Kok

Health, Global Budgets for

Introducing global budgets for health, i.e., setting prospective limits to expenditure on health care, is by definition the most effective means of controlling spending. If expenditure levels are fixed for the coming year or years, and if the limit is made to stick, then tautologically the rise in spending on health will be contained within the planned total. Not surprisingly, therefore, an increasing number of governments adopted this strategy in the 1980s, as a reaction to the largely unplanned and unanticipated rise in health care spending throughout the 1960s and much of the 1970s: what came to be known as the health care cost explosion. This article examines, first, the feasibility, form and effectiveness of this strategy in different health care systems; second, the conceptual and practical issues involved in fixing the global budget and then allocating it to populations or health care sectors.

1. Strategies and Mechanisms

In 1960, total health expenditure (privately as well as publicly financed) in the world's 18 richest countries—

those belonging to the Organisation for Economic Co-operation and Development—accounted for 4.2 percent of their combined gross domestic product (GDP). Ten years later, the figure was 5.8 percent, rising to 7.2 percent in 1980 (OECD 1987). Both the rate of increase, and the proportion of GDP devoted to health, differed among countries. Between 1960 and 1980, spending in the United States increased from 5.3 to 9.5 percent of GDP, rising to 10.7 percent by the mid-1980s, whereas the figures for the UK were 3.9 and 5.6 percent respectively, leveling out thereafter. From this flowed a widespread perception that spending on health was out of control, imposing damaging fiscal burdens on the economy and a consequent interest in mechanisms for halting, or at least slowing down, the seemingly inexorable rise.

The problem of ever-escalating costs could not be solved by invoking the usual market mechanisms and restraining demand by raising prices. Not only is health care financed predominantly through insurance schemes of various types; i.e., third parties rather than the consumers pay most of the costs. Equally, there is only limited scope for using co-payments or user charges to limit demand: the long-term effects on population health of deterring use may outweigh any short-term fiscal gains. Controlling the supply of health care thus seemed a more feasible and effective option, all the more so because of evidence that much of the demand for health care was generated by the system itself; i.e., was provider-driven. While it is consumers who make the initial contact with the health care system, it is the professionals who decide how they should be treated. And it is these decisions—about the number of tests to be carried out, the drugs to be prescribed and so on—which drive spending. Enter global budgeting.

Here the UK's National Health Service (NHS) was the paradigmatic case. Outside the communist bloc countries, the UK was the only health care system funded overwhelmingly by central government out of general taxation. In this it differed both from the tax-funded Scandinavian model, where both control and fiscal responsibility was divided between central and local government, and the social insurance based West European model, where responsibility was divided between central government and an often large number of insurance funds. In the UK, therefore, the Treasury controlled the fiscal tap, for both current and capital spending, to the health care system. Almost from the launch of the NHS in 1948, the system operated what in effect was a global budget; i.e., the Treasury set the limits of total spending for the next fiscal year. In turn, the UK's record suggested two conclusions. First, it appeared to demonstrate that global budgeting could restrain the rate of increase in health care expenditure: the UK continued, and continues, to be a restrained spender relative to other countries with a comparable national income. Second, it seemed to show that there was a direct relationship

between fiscal centralization and the successful operation of a global budget.

Both conclusions were borne out, if to a differing extent, as an ever higher proportion of countries adopted variations on the global budgeting theme in the 1980s and early 1990s as part of their strategies for cost containment. The rate of increase in health care expenditure, as a proportion of national incomes, did indeed slow down (Abel-Smith 1992, OECD 1992, 1994), and much of this could be attributed to the introduction of global budgeting. In all this, the USA was the most conspicuous exception; it did not introduce global budgeting and its spending on health care was almost 14 percent of GDP by the end of the 1990s (twice the UK figure). However, success did not necessarily correlate with fiscal centralization. Sweden not only halted, but reversed, the upward trend, even though its model of health care divides responsibility between central and local government: no doubt because the former carries by far the heavier fiscal clout. And the Netherlands, perhaps the most pluralistic variant on the European social insurance model, introduced global budgeting covering both the public and the private insurers (Maarse 1996) with the same set of fee schedules applying to both sectors.

This suggests that there are various mechanisms for achieving global budgeting—i.e., control of the supply side of health care—although the UK's single-payer model offers the most direct leverage. Without the simplicity of the UK model, these mechanisms vary with the nature of the reimbursement system used in different sectors of health care services and the political determination of governments to make their policies stick. In the case of the hospital sector, it proved relatively easy to impose global budgets (Glaser 1993). In the case of service payments for general practitioners or office-based specialists, controlling spending turned out to be less easy. Freezing the fee schedule tends to multiply activity as physicians safeguard their incomes by doing more. Here the strategy adopted was to impose a limit on the total amount to be paid out in fees, with retrospective downward adjustments of the payments to individual physicians if the total exceeds the target figure: Germany providing one such example, leaving responsibility for administering the pool to the medical profession (Schwartz and Busse 1996). And while the USA is a conspicuous failure in macro-terms, it has generated innovative models for microglobal budgeting; i.e., various forms of capitation payments to providers.

Systems also vary in the extent to which global budgets are underpinned by, or complementary to, other supply side measures. Some systems, such as that in the UK, control the production and distribution of the medical labor force. Most systems seek to control the number and distribution of hospitals and to exercise birth control over the diffusion of expensive new technologies. In the case of health care systems based on the social insurance model, govern-

ments can further buttress policy by regulating—in some instances freezing—the level of contributions charged by insurance funds, so controlling the flow of money into the system.

Global budgets have their pathologies as well as their advantages. By removing the incentive to maximize income by maximizing activity (possibly a desirable policy goal, since it also lessens the risk of overtreatment), global budgets for hospitals may also undermine the incentives to maximize efficiency (so leading to the undesirable policy outcome of queues for treatment). Hence Britain's brief experiment with a mimic market in the 1990s. This separated the purchaser and provider roles. Previously, health authorities had been responsible for managing hospitals and other services directly for the population within their geographical boundaries. Under the new system, they became responsible only for buying the necessary care from semi-autonomous agencies. Both before and after the changes, however, health authorities worked within budgets fixed by central government. The aim (never fully achieved) was to introduce an element of competition between providers—and thus provide incentives to greater efficiency—within the framework of a global budget for the NHS as a whole.

Some other issues also require noting. Global budgets for item-of-service payments raise awkward questions about the distribution of income between physicians: between those who maximize their activity and those within a fixed total who do not. Above all, global budgets tend to be associated with rationing: if the supply is limited, waiting lists may be the result. And while the phenomenon of rationing is far from exclusive to health care systems with global budgeting, it does raise questions about the adequacy of the budget level set.

2. Setting the Budget

The adoption of global budgets represents the pragmatic reaction of governments searching for policy instruments to restrain the growth of collective spending on health. A global budget for the health care system of a country is thus the byproduct of a government's decisions about priorities between expenditure on health and other programs competing for funds, between maintaining public services and cutting taxation. Its level reflects as much calculations of political feasibility—assumptions about how much spending can be restrained without provoking a political backlash from professional providers and the public—as calculations about what the appropriate level of spending on health should be. As such, it is an entirely atheoretical exercise. For no formula for determining a global budget, independent of the actual level of spending in different countries, has yet been devised: there are no agreed benchmarks against which the adequacy or otherwise of a given level of spending

can be assessed. Nor is it likely that such benchmarks can be devised, given the heterogeneity of the services produced under the 'health' label, the scope for substitution between them, and the productive factors involved and variations in productivity.

Global budgets are therefore best seen as history modified by politics, and often rationalized by appeals to demography. The inherited level of spending on health inevitably provides the baseline for setting global budgets. This is then modified in the light of factors either extrinsic to the health care system (usually the prospective capacity of the economy to produce extra wealth), or intrinsic to it (usually demography and the cost of technological innovation). The first strategy puts the emphasis on what a country can afford. The second puts the emphasis on what a country needs. Neither is as objective as it appears (Smee 1996). 'Affordability' is an elastic concept. And so is need. Even the most commonly used indicator of 'need,' the demographic structure of any given population, is problematic. For sure, people need more health care as they age. But projecting the cost of demographic change into the future assumes that patterns of both morbidity and service use will remain the same as in the past: a highly questionable assumption. And many of the same problems arise when it comes to apportioning global budgets, whether geographically or by health care sectors.

3. Dividing up the Budget

The issue of distributing global budgets geographically arises only in the case of health care systems—on the UK and Scandinavian models—where public agencies are responsible for delivering health care to a defined population. In these cases, the equity principle demands that resources should be distributed according to need. In the past, need was often equated with use: if use was high so, it was assumed, was need. But given increasing evidence that demand (or use) was strongly influenced by supply, emphasis has switched to devising other indicators. The demographic composition of the population provides, as noted above, the core of the various formulas that have been devised to allocate resources. Standardized Mortality Ratios are another. So is morbidity data, where available. Debate continues as to whether social deprivation is in itself a need indicator, over and above its contribution to the mortality and morbidity figures. In all this, need is, of course, a relative concept. That is, the various distributive formulas that have been developed purport to measure the needs of population A as against those of population B, not to measure the absolute level of provision that either requires.

In the case of the West European social insurance model of health care provision, the question of geographical distribution clearly does not arise. Re-

sponsibility for funding health care is divided between funds whose members—as often defined by occupation as by place of residence—do not necessarily live within the same geographical areas. However, some of the same distributional issues do arise. One frequent element in social insurance systems is a central pool of money which is then distributed to individual funds according to the risks that they carry. Designed to weaken the incentives of funds to discriminate against potential members who may—because they are older or sicker—require more expensive health care, risk compensation in turn means devising a formula for assessing relative need. This raises the same conceptual issues as devising a formula for distributing global budgets geographically and leads to many of the same solutions, notably heavy weighting for demographic factors.

Global budgets are, in most systems, also broken down by sectors for accounting and control purposes. Indeed, a national global budget may be the product of a step-by-step process of imposing expenditure limits on specific sectors; primary care, acute hospital care and so on. However, in this case there have not been—and probably cannot be, given the absence of a conceptual basis for such an exercise—any attempts to devise a formula for allocating funds between sectors. The sectoral distribution of global budgets, even more than the determination of their total, remains the product of history modified by incremental policy decisions.

See also: Health Care Delivery Services; Health Care: Legal Issues; Health Care Markets: Theory and Practice; Health Care, Rationing of; Health Care Systems: Comparative; Health Economics; Health Policy; National Health Care and Insurance Systems; Pecuniary Issues in Medical Services; Ethical Aspects; Public Health

Bibliography

- Abel-Smith B 1992 Cost containment and new priorities in the European Community. *The Milbank Quarterly* 70(3): 393–416
- Glaser W A 1993 How expenditure caps and expenditure targets really work. *The Milbank Quarterly* 71(1): 97–128
- Maarse H A M 1996 Fixed budgets in the inpatient sector: The case of the Netherlands. In: Schwartz F W, Glennerster H, Saltman R B (eds.) *Fixing Health Budgets: Experience from Europe and North America*. Wiley, Chichester, UK
- OECD 1987 *Financing and Delivering Health Care: A Comparative Analysis of OECD Countries*. OECD, Paris
- OECD 1992 *The Reform of Health Care: A Comparative Analysis of Seven OECD Countries*. OECD, Paris
- OECD 1994 *The Reform of Health Care Systems: A Review of Seventeen OECD Countries*. OECD, Paris
- Schwartz F W, Busse R 1996 Fixed budgets in the ambulatory care sector: The German experience. In: Schwartz F W, Glennerster H, Saltman R B (eds.) *Fixing Health Budgets:*

- Experience from Europe and North America*. Wiley, Chichester, UK
- Schwartz F W, Glennerster H, Saltman R B (eds.) 1996 *Fixing Health Budgets: Experience from Europe and North America*. Wiley, Chichester, UK
- Smee C H 1996 Setting regional allocations and national budgets in the UK. In: Schwartz F W, Glennerster H, Saltman R B (eds.) *Fixing Health Budgets: Experience from Europe and North America*. Wiley, Chichester, UK

R. Klein

Health in Developing Countries: Cultural Concerns

Human health and wellbeing worldwide are profoundly affected by complex interactions between economic, social, and political forces as well as by behavioral patterns and the application of technology (Wilkinson 1996). When considering health in countries around the world, use of the terms 'developed' and 'developing' countries implies a linearity of progress towards a condition that has been achieved by some countries but not others. It is clear that such a simplistic approach equating development with economic status in terms of average per capita gross national product (GNP) or the level of industrialization is both inaccurate and unhelpful. So-called developed countries may lack what many would consider very important aspects of development; for example, a sensitivity to the impact of consumption on ecology, or concern for the adverse effects of opulent and aggressive lifestyles on billions of people in other countries, or more obviously in relation to health, the achievement of universal access to health care. Similarly, some so-called developing countries are more developed than may be apparent from their economic status; for example in Kerala, India, the degree of social solidarity and health status achieved exceeds that expected from economic status alone—revealing the impact of human capital resources such as female literacy and empowerment.

A more sophisticated model of development (such as that used in the United Nations Human Development Report 1991) emphasizing achievements in health, education, female literacy, gender equality, and political liberty, more accurately reflects the real human development required. However, it should be acknowledged that even such development indices are incomplete, as they do not measure some important aspects of human flourishing. When comparing development and health in countries around the world, it is also necessary to appreciate that consumption patterns and the ways of life in industrialized countries

cannot become the norm for the whole world. For example, while the world's population has grown fivefold in the past 150 years, human consumption has increased 30-fold. The USA, with 5 percent of the world's population, accounts for 24 percent of annual global energy consumption compared to China with 20 percent of the world's population using 10 percent. Within health care, even a nation that spends \$4,000 per person per year on health care cannot afford to provide access for all citizens to all that modern medicine can offer (Iglehart 1999). Modern consumption patterns and health care practices are thus neither universalizable nor sustainable. Such considerations reveal the inadequacy of the terms 'developed' and 'developing' countries.

1. A Polarized World

While the twentieth century is considered by many to be the most successful in world history, with spectacular achievements in science, technology, and health care, the above considerations make it necessary to focus on the down side of progress—as reflected in widening disparities in economic and health status, the extent of expenditure on weapons of mass destruction, and the social construction of human rights abuses. Improvements in human health globally in the twenty-first century and achieving real development of all people throughout the world will be dependent on acknowledging such adverse features of modern life, and on developing new paradigms of thinking that will promote widespread human flourishing.

1.1 Economic Disparities

At the beginning of the twentieth century the wealthiest 20 percent of the world's population were nine times richer than the poorest 20 percent. This ratio has grown progressively—to 30 times by 1960, 60 times by 1990, and to over 70 times by 1997. The extent of absolute poverty has also increased. Today almost half the world's population lives on less than US\$2 per person per day and more than a billion people live on less than US\$1 per day. Tens of millions of people, including many children, die every year of starvation—in a world with enough food to feed all.

1.2 Disparities in Health and Disease

At the end of the twentieth century patterns of diseases and of longevity also differ markedly across the world. Of the approximately 52 million people who die each year, 17.5 million die of infectious and parasitic diseases (16 million of these—many in their youth—in

Table 1

Causes of death in 1997 by main causes and by level of development (The World Health Report 1997)

	Developed world (12 million deaths)	Developing world (45 million deaths)	World
Infectious and parasitic diseases	1 percent	43 percent	33 percent
Perinatal and maternal causes	1 percent	10 percent	8 percent
Cancers	21 percent	9 percent	12 percent
Diseases of the circulatory system	46 percent	24 percent	29 percent
Diseases of the respiratory system	8 percent	5 percent	6 percent
Other and unknown causes	23 percent	9 percent	12 percent

the developing world), 10 million die of diseases of the circulatory system (4.5 million of these in the developing world), and 6 million die of malignant diseases (3.5 million of these in the developing world) (see Table 1).

Among the poorest quintile of people in the world, 55 percent die of communicable diseases, as compared with 5 percent of the richest. The World Health Organization (WHO) estimated that in 1998, 11 million children and adults of working age died of six infectious diseases that could have been prevented at the cost of US\$20 per life saved. Poor countries bear over 80 percent of the global burden of disease in disability adjusted life years (DALYs). This burden is likely to increase as the epidemiological transition progresses, with added disability and suffering from noncommunicable causes of disability such as vascular disease, malignant neoplasms (especially of the lung associated with smoking), neuropsychiatric disease, accidents, and trauma.

Life expectancy at birth ranges from well over 70 years in highly industrialized countries to below 50 years in many poor countries. Wide disparities in life expectancy are also observed between rich and poor in rich countries. In sub-Saharan Africa gains in longevity achieved during the first half of the twentieth century are rapidly being reversed by the HIV/AIDS pandemic.

1.3 Mental Health

Improvements in physical health over the past 50 years have not been matched by improvements in mental health—even in developed countries. Mental illness accounts for 8 percent of the global burden of disease (measured in DALYs), and a further 34 percent of the global burden of disease is accounted for by behavior related problems such as violence, diarrhoeal diseases, AIDS and other sexually transmitted diseases, and

motor vehicle accidents and other nonintentional injuries (Desjarlais et al. 1995). The inadequacy of the knowledge base about mental disease (compiled almost exclusively in North America and Europe) to explain psychiatric diseases in other parts of the world poses formidable challenges for progress.

1.4 Militarization and the Social Construction of Human Rights Abuses

Industrialized countries spend on average 5.3 percent of GNP on the military but only 0.3 percent on economic aid to developing countries (Sivard 1996). The arms trade and its link to economic aid has contributed significantly to the escalating number of wars, conflict, and widespread torture since the Second World War which have rendered over 25 million people to refugee status and displacement—with profound disruptive effects on life and health. Global military expenditure in 1990 amounted to US\$1 trillion (Sivard 1996). The annual budget of the WHO is equivalent to 2.5 hours of global military expenditure. While money is always available to make war, it is less readily available to promote peace. Military expenditure has decreased in recent years, but released resources have not been channeled towards effective human development processes.

2. Relationship Between Economics and Health

It is now well established that there is a definite relationship between wealth/poverty and health/disease, although this relationship is complex and not linear. For example, one of the wealthiest nations in the world (the USA) has worse health statistics (infant mortality and longevity) than some other highly industrialized countries, and a particularly poor state

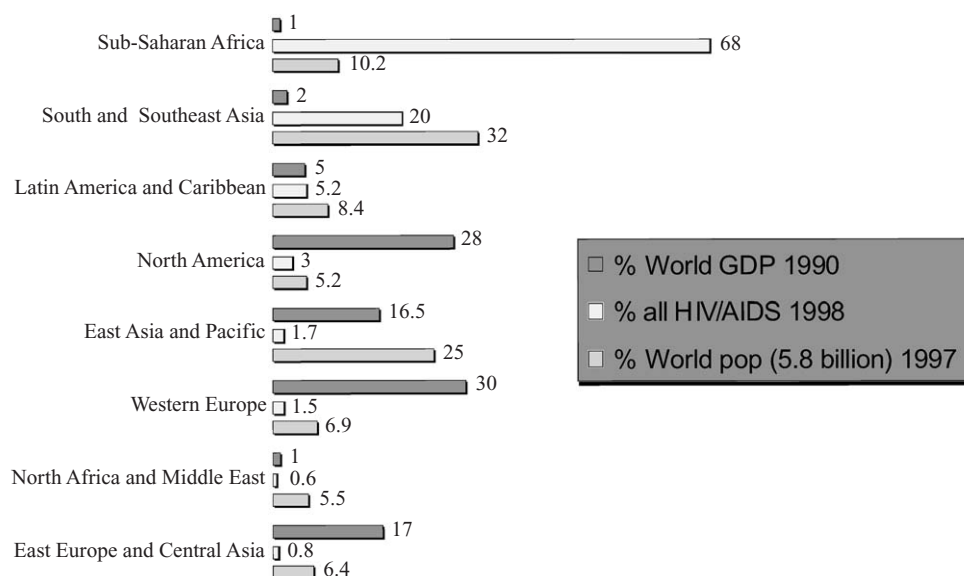


Figure 1

Distribution of economic resources (top bar), the burden of HIV/AIDS (middle bar) and world population (lower bar) in the world

Table 2

Health and other statistics in the 40 poorest countries in the world (WHO 1995)

Annual per capita GNP (US\$) ^a	168	282	404	570
Life expectancy at birth (years) ^a	47	51	56	60
Infant mortality rate per 1,000 live births ^a	126	112	90	70
Under 5 year mortality rate	158	169	132	109
Population access to safe water ^a	37%	53%	53%	63%
Population access to sanitation ^a	29%	26%	39%	48%
Population per physician (1970)	33,925	39,381	23,009	15,956
Adult literacy rate (1992)	42%	44%	46%	63%
Average annual population growth ^a	2.7%	2.8%	2.6%	2.3%

^a 1991 data.

(Kerala) in India has achieved lower infant mortality rates and greater longevity than many wealthier nations (Wilkinson 1996). Both absolute and relative wealth affect health. Among the developed countries it is not the richest societies that have the best health, but rather those with the smallest income differentials between rich and poor (Wilkinson 1996). Despite the nonlinearity of the relationship between wealth and health, its existence underscores the need to see health and disease as intimately linked to social and economic conditions (Fig. 1).

Poverty (defined as lack of economic resources, lack of education, lack of access to basic life resources such as food water and sanitation, and lack of control over

the reproductive process) directly accounts for almost one-third of the global burden of disease. Poverty leads to poor health, which in turn aggravates poverty. Of the 40 poorest countries in the world, 11 have annual per capita GNPs ranging from US\$80 (Mozambique) to US\$210 (Sierra Leone), 10 from US\$220 (Bangladesh) to US\$340 (Nigeria), 9 from US\$370 (China) to US\$450 (Nicaragua), and 10 from US\$500 (Sri Lanka) to US\$600 (Zimbabwe) (WHO 1995). The relationship between income and some health indicators for these groups are shown in Table 2. Despite the difficulty in making accurate intercountry comparisons, it is clear that improvements in economic status are associated with better health statistics, and that

economic growth is necessary to improve the living conditions conducive to good health, independent of benefits that may arise directly from health care.

3. Health Care Expenditure

Since the 1960s, major advances in medicine and technology have been associated with escalating expenditure on their application in health care—most of this in highly industrialized countries. Annual per capita expenditure on health care ranges from \$4,000 in the US to less than \$5 in the poorest countries in Africa. Half the world's population lives in countries that cannot afford annual per capita health expenditures of more than \$5–10, and many people do not have access to even basic drugs. The WHO has estimated that in 1990 the annual cost of providing a basic package of public health and essential clinical services in a low-income country was at least \$12. In most such countries health care expenditure is typically less than \$6. Even though some parasitic diseases (for example, onchocerciasis, schistosomiasis, and lymphatic filariasis) could be controlled by mass treatment campaigns using inexpensive drugs, the infrastructure required to provide such coverage is inadequate. Effective treatment of diseases such as tuberculosis and leprosy, and HIV/AIDS requires more complex infrastructures than can be afforded on current health care budgets in poor countries.

In the 1990s, 89 percent of annual world expenditure on health care was spent on 16 percent of the world's population who bear 7 percent of the global burden of disease (in DALYs) and less than 10 percent of medical research expenditure was on health problems of the developing world. These are examples of global injustice that we should surely find intolerable if we have a real commitment to universal human rights and human dignity. Vaccine development programs and drug donation programs for poor countries, admirably promoted and supported by pharmaceutical companies, are necessary but insufficient responses to such injustices. Modest increments in economic status combined with a provision of a basic health care package have the potential to improve health in poor countries significantly. However, it should be noted that health is not merely a resource issue even in rich countries—as exemplified by comparative health statistics around the world (Wilkinson 1996).

4. The Global Political Economy and Health

The world at the end of the twentieth century is thus characterized by widening disparities in economic and health status (within and between countries), and by suffering, conflict, and alienation associated with pervasive social forces (Wilkinson 1996, Sivard 1996,

Benatar 1998). Erosion of the economy of many poor countries, under the impact of globalization and the 'debt trade,' has obstructed the introduction of effective forms of modern medicine into many poor countries and thus prevented achievement of widespread access to even basic health care for billions of people (Falk 1999). Third world debt requires special mention. The ways in which such debt have been created, the fact that third world debt is a small component of total world debt and can never be repaid, and that in sub-Saharan Africa four times as much is spent on debt repayment each year than on health and education combined, makes insistence on repayment unconscionable (Pettifor 1996).

The emergence of new diseases such as AIDS, that afflict predominantly those marginalized by poverty (80 percent of HIV-positive persons live in the poorest countries in the world), has been attributed to the social and environmental niches created by the nature of the global political economy and its ideology (Lee and Zwi 1996). Failure to appreciate such associations will result not only in failure to control such diseases as HIV/AIDS (as with tuberculosis in the past, Benatar 1995), but more importantly, will probably favor the emergence of new infectious diseases in the future. The changing spatial dimensions of human relations resulting from urbanization, migration, travel, ethnic conflict, and displacement of people and refugees, facilitates the spread of such diseases from which no one should feel immune.

5. Access to Health Care, and the Nature of Health Care Systems

Today many countries consider access to at least a minimum package of health care as a basic human right that nation-states should be committed to providing for their citizens. Some form of socialized and equitable health care is provided in all Western European nations and in Canada.

Medical care, delivered as a marketable commodity—albeit with considerable state assistance for the poor and the aged—is the example that has been set by one of the wealthiest nations in the world. However, this results in many citizens having inadequate access to health care. In considering how to overcome this, while recognizing that private/public alliances cannot be avoided and indeed may be beneficial, Buchanan has provided cogent arguments for the nine elements he considers essential for a just health care system. He argues further that achieving a just health care system as a cooperative between the public and private sectors requires an organization, such as the state, to act as a coordinator and guarantor of equitable access to care and fair distribution of costs. Such a balance between private and public medicine with appropriate coordination could over-

come the deficiencies of either system when one dominates (Buchanan 1995; see also Benatar 1996).

The effects of intensive privatization, aided and abetted by the structural adjustment programs promoted by the International Monetary Fund (IMF) and the World Bank, have been much more marked in poor countries, with severe adverse effects on health. Acknowledging the need for, and the right to, universal access to a basic health care package and achieving this goal pose challenges for the future. The WHO's renewal strategy for health-for-all places emphasis on equity, solidarity, and appropriate technical, political, and economic strategies that could promote health and sustainable health care systems as central to development. Here too an appropriate balance needs to be struck between public and privately funded medicine. How this balance can be struck and how health can be coordinated within an effective strategic alliance between many parties remain hotly debated subjects.

6. Traditional Healers

Despite the pervasiveness of Western influences, traditional healers have survived and play an increasing role in some countries. For example, while there are about 25,000 Western trained medical practitioners in South Africa, and Western medicine has a powerful foothold in this middle-income country, over 200,000 traditional healers carry out a thriving trade. Increasing contact and collaboration between Western medicine and traditional healers is being encouraged and the potential impact of their contributions to health, in particular through development of a primary health care approach, in many countries in Africa in the future should not be underestimated (Van Rensburg and Benatar 2000).

7. Prospects for the Future

As we stand at the threshold of a new era in which exponential increments in genetic information will be applied to medical practice, we need to contemplate whether such new power will be used for the benefit of all humankind or whether it will be reserved for the privileged. Acquiring a deeper understanding of the forces promoting health and diseases in developing countries needs to be coupled to a commitment to address these if health and human flourishing are to be achieved on a wider scale (Benatar 1999).

See also: African Studies: Health; Economics: Overview; Health Economics; Health Policy; Infant and Child Mortality in the Less Developed World; Poverty and Gender in Developing Nations; Reproductive Rights in Developing Nations; Third World

Bibliography

- Benatar S R 1995 Prospects for global health: Lessons from tuberculosis. *Thorax* **50**: 487–9
- Benatar S R 1996 Editorial on Buchanan (1996). *British Medical Journal* **313**: 1567–8
- Benatar S R 1998 Global disparities in health and human rights. *American Journal of Public Health* **88**: 295–300
- Benatar S R 1999 A perspective from Africa on human rights and genetic engineering. In: Burley J (ed.) *The Genetic Revolution and Human Rights*. Oxford Amnesty Lectures. Oxford University Press Oxford, UK, pp. 159–89
- Buchanan A 1995 Privatization and just health care. *Bioethics* **9**: 220–39
- Desjarlais R, Eisenberg L, Good B, Kleinman A 1995 *World Mental Health: Problems and Priorities in Low-income Countries*. Oxford University Press, New York
- Falk R 1999 *Predatory Globalization: A Critique*. Polity Press, Cambridge, UK
- Iglehart J 1999 The American health care system: expenditures. *New England Journal of Medicine* **340**: 70–6
- Lee K, Zwi A B 1996 A global political economy approach to AIDS. *New Political Economy* **1**: 355–73
- Pettifor A 1996 *Debt, the Most Potent Form of Slavery*. Debt Crisis Network, Christian Aid, London
- Sivard R L 1996 *World Military and Social Expenditures*, 16th edn. WMSE, Leesburg, VA
- Van Rensburg H C J, Benatar S R 2001 History of medicine in Africa. In: Last J, Lock S (eds.) *The Oxford Illustrated Companion to Medicine*. Oxford University Press, Oxford
- Wilkinson R G 1996 *Unhealthy Societies: The Afflictions of Inequality*. Routledge, London
- World Health Organization 1995 Poverty and health in developing countries. No. 16 in the WHO series *Macroeconomics, Health and Development*. World Health Organization, Geneva

S. R. Benatar

Health Insurance: Economic and Risk Aspects

Over the course of a year, individuals are at risk for a variety of unfortunate and unpredictable events, e.g., property damage due to fire or hurricane, property loss due to burglary, or physical and mental pain and suffering due to adverse health events. In most cases, affected individuals can purchase goods or services to replace the loss or to alleviate the adverse effects of an untoward event. In some cases, it is possible to reduce the probability that an adverse event will occur. In many instances, through the purchase of insurance, it is also possible to reduce financial loss associated with an adverse event. In general, people dislike risk and are willing to trade a small amount of money, in the form of an insurance premium, for protection against a potentially large loss in income. (Depending on what kind of health insurance system, particularly whether it is a private or public system, the payment can be

either a private insurance premium paid out of pocket or through a reduction in wages or a public premium in the form of taxes. For simplicity, we will refer to the payment as a premium. Similarly, the bearer of risk under insurance can be a private company or a public governmental body. We will refer to the bearer of risk as the insurer.)

This entry outlines the basic economic models of health insurance, emphasizing the relationship between risk and insurance demand. It also discusses the distortions in the market for health care services created by health insurance.

1. Risk and the Demand For Insurance

The relationship between risk and insurance demand is derived from the relationship between utility and wealth (for additional discussion see Arrow 1963). Utility is a function of wealth, with more wealth leading to higher levels of utility or satisfaction. Although utility rises with wealth, it is generally assumed to do so at a decreasing rate. This property is called diminishing marginal utility. Each incremental increase in wealth provides a smaller incremental increase in utility. It therefore follows that the gain in utility associated with any incremental gain in wealth is less than the loss in utility associated with an equivalent loss of wealth.

The property of diminishing marginal utility implies individuals are *risk averse*. That is, individuals would prefer to have any level of wealth with certainty than a gamble providing the same level of wealth on average. For example, a risk averse individual would prefer \$100 dollars with certainty as opposed to a gamble with a 50 percent chance of winning \$200 and a 50 percent chance of winning \$0. Risk aversion is an inherent property of a concave utility function. Because risk averse individuals prefer certainty, the premium they are willing to pay for an insurance policy that removes risk exceeds the actuarially fair premium (AFP) of the insurance, which is the amount that the insurer would have to pay out, on average, for this policy. The gap between the premium individuals are willing to pay and the AFP is termed the *risk premium* (Phelps 1997).

The risk premium allows insurers to cover expenses above the medical payout, e.g., claims processing. Several factors determine the size of the risk premium an individual will be willing to pay. First, risk aversion is likely to vary with wealth. For very different reasons, both individuals with very high and those with very low levels of wealth are less willing to pay a risk premium (Feldstein 1999). The marginal disutility of an incremental decrease in wealth falls at high levels of wealth, reducing any utility gain to avoiding risk. In contrast, the marginal disutility becomes very large at very low levels of wealth, making the opportunity cost of purchasing insurance too high.

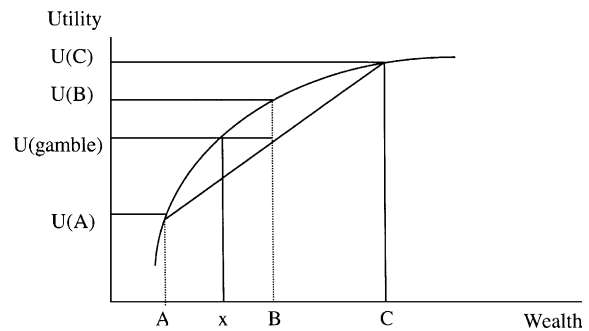


Figure 1
Expected utility and risk

Second, the probability of a loss will influence the size of the risk premium. As the probability of the loss approaches 1, the willingness to pay for insurance rises, but more slowly than the increase in the AFP for that individual. Thus the risk premium falls. In the extreme case, when the probability of a loss equals 1, the risk premium goes to zero. In this case, there is no risk and individuals would not be willing to pay any risk premium. Similarly as the probability of a loss goes to zero, both the willingness to pay and AFP fall, but the willingness to pay falls faster and eventually the risk premium equals zero. Individuals would not pay for insurance if the probability of a loss equaled zero.

Third, the magnitude of the loss affects the risk premium individuals are willing to pay. A greater loss represents an increase in the variance of income. An individual will be willing to pay a higher risk premium for a higher cost illness.

The utility curve depicted in Fig. 1 can be used to analyze the demand for insurance described above. Throughout the analysis we assume individuals know the probability they will suffer a loss (or benefit from a gain). Because greater wealth leads to higher utility, the utility at point C, $U(C)$, is greater than the utility at points A or B. Diminishing marginal utility implies that if point B is equidistant from points A and C, the utility gained from moving from B to C is less than the utility lost from moving from B to A.

Consider an individual at wealth B evaluating a gamble with outcomes A and C, each of which has a 50 percent probability of occurring (e.g., tossing a fair coin). If point B is equidistant from points A and C, the expected wealth resulting from the gamble, i.e., where *on average* the individual could expect to be if tossing the coin, is the initial level of wealth, B. The expected utility of such a gamble, $EU(\text{gamble})$, is the probability weighted sum of the two outcomes A and C and, because each outcome has a 50 percent chance of occurring, can be determined by finding the midpoint on the chord connecting A and C. Because the utility curve is concave, even though the initial level of

wealth is the same as the expected wealth of the gamble, the utility of B, $U(B)$, will exceed the expected utility of the gamble, $EU(\text{gamble})$. This utility gain of avoiding risk is key to the demand for insurance.

Models of health insurance fit exactly into this model. For simplicity, assume a world in which there is only one type of adverse health event and spending in the unhealthy state is unaffected by the presence of insurance. Individuals are assumed to start with wealth level C and remain there if healthy during the year. If they suffer an illness shock, they will then spend $C - A$ on health care services. Assuming the probability of a loss is 50 percent, the expected loss is $C - B$. Expected wealth, under uncertainty, is B. Without insurance the individual would have expected utility of $EU(\text{gamble})$.

Now imagine individuals could purchase an insurance contract that would pay the costs of medical care in the event of an illness. With a 50 percent chance of each individual incurring the loss, the expected cost to the insurer of each enrollee is $C - B$. The AFP therefore is $C - B$. If individuals pay the AFP, their wealth level is B, regardless of whether the illness occurs. In other words, the risk of financial loss has been eliminated. The individual has turned a potentially large loss in income into a smaller known loss of income (in the form of the AFP). In this situation, the utility would be $U(B)$ because the income level B is achieved with certainty. As is clear from the figure, the individual prefers to purchase insurance relative to self-insuring. The amount of the utility gain is equal to the vertical distance between $U(B)$ and $EU(\text{gamble})$. The greater the concavity of the utility curve, the greater the risk aversion and the greater the utility gain from insurance.

The utility curve depicted in Fig. 1 also can be used to assess the amount that an individual would be willing to pay for insurance. Utility of a wealth level of x, with certainty, is equivalent to the expected utility of remaining uninsured (facing the risk of having a loss that leaves them at wealth level A). Thus an individual would be indifferent between being uninsured or paying a premium of $C - x$ for insurance. This premium is their maximum willingness to pay for insurance (WTP), with $C - B$ the AFP and $B - x$ the risk premium.

2. Moral Hazard

Up until now, the magnitude of the loss associated with an illness was not affected by the presence of insurance coverage for the desired medical care services. Substantial theoretical and empirical work suggests that this is not the case (Pauly 1968, Manning et al. 1987, Newhouse 1992). Specifically, because health insurance lowers the price of medical care services faced by the consumer at the time of purchase, individuals will often consume more care than they would if they paid the full price. The providers of this

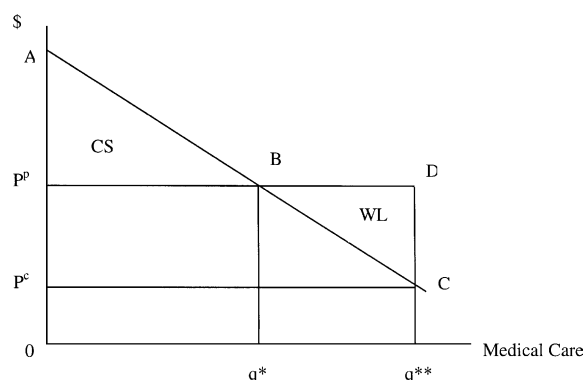


Figure 2
Welfare loss and moral hazard

care are paid the full price, however, receiving out of pocket payments from the patient and reimbursement from the insurer.

The consequences of this behavior, known as moral hazard, depend in part on the difference between the cost of the care and the value individuals place on that care. Because individuals do not demand this same level of care without insurance, standard economic models assume that the value they place on the care is less than the full cost of that care. If this is true, insurance causes a loss in consumer welfare because of the insurance-induced increase in consumption, whose costs are passed back to the consumer in the form of higher insurance premiums. The welfare loss is thus incurred by the enrollee at the time the insurance premium is paid, not when care is received. This welfare loss accrues even if, as is the case in most complex insurance markets, the individual is not purchasing health insurance directly and may in fact not be aware of the welfare loss. In aggregate, adjusting for transfers between individuals, the amount paid through premiums exceeds the value of care financed by the insurance.

This analysis is depicted graphically in Fig. 2. The demand curve tells us how much medical care will be demanded for any given price. For example, assume the price of care, as paid to the provider of care, is denoted P^p . In this case, uninsured individuals would purchase q^* units of care. If insurance lowered the individual's out of pocket price to P^c , insured individuals would then consume q^{**} units of care.

The demand curve is taken as the measure of value because it also represents the maximum amount an individual would be willing to pay for any given quantity of care. For example the maximum amount an individual would pay for q^* is P^p . The overall value placed on all care up to any quantity is the area under the demand curve up to that point. For an uninsured individual, the value of care consumed would be the area under the demand curve up to q^* (ABq^*0). For

an insured individual, it would be represented by ACq^{**0} .

The total cost of care, paid by the individual and insurer, is the price paid to the provider times the quantity of services provided. In the case of an uninsured individual, it would be P^0 times q^* (P^0Bq^{*0}). Notice that the value of care to the consumer exceeds the costs of the care by the portion of the area under the demand curve not included in the box (P^0AB). This net value to the consumer is termed consumer surplus (CS).

If the individual purchases insurance and then increases consumption, the total cost of care increases to P^0Dq^{**0} . Notice the incremental cost of the care purchased, BDq^{**q^*} , exceeds the incremental value to the consumer, BCq^{**q^*} . The amount by which the incremental cost exceeds the incremental value represents the welfare loss (WL) of insurance and is represented by the triangle BDC.

At the time that care is consumed, the individual only pays a price of P^c , so the out of pocket cost is P^cCq^{**0} . The actuarial cost of the care to the insurer, however, is P^0DCP^c . It is this cost, not the out of pocket cost, that determines the AFP. As noted above, the welfare loss is thus incurred at the time of insurance purchase. If the demand curve were inelastic, that is, if the level of consumption were not very responsive to changes in price, the welfare loss would be relatively small. Consumers, if aware of the welfare loss, would prefer to consume less in exchange for an appropriately lower premium, particularly in cases in which the welfare loss is large (elastic demand).

There are three important assumptions that would affect the size of the welfare loss. First, the analysis assumes that individuals other than the patient receiving the care do not receive positive utility from that consumption. If this positive externality exists, the welfare loss is reduced. Second, the analysis assumes that the price paid to providers is equal to marginal cost (as would be the case if it were set in a competitive market). If the price were set in an imperfectly competitive market, so that it exceeded marginal cost, the welfare loss from a societal perspective would be diminished. Third, the preceding analysis assumes that income effects associated with paying the premium and the transfer of income from healthy to sick individuals have only small effects on the consumption of medical care. Nyman and Maude-Griffin (2001) suggests that if these effects are large, the welfare loss is considerably reduced.

Many different approaches might be taken to minimize the welfare loss associated with insurance. For example, increasing consumer cost-sharing for medical care services would reduce utilization but has the disadvantage of increasing the size of the financial risk faced by the consumer. Increasing provider cost-sharing through a variety of financial incentive systems would reduce utilization, but has the potential disadvantage of too little care being prescribed. Private

regulatory mechanisms, such as utilization review, can be implemented to constrain both where and what kinds of covered treatments an individual can receive. Managed care plans in the United States have been using these and other approaches to restrain use. In some cases, plans select providers whose preferred practice style leads to lower levels of utilization, particularly of higher cost alternatives. Some plans also educate both enrollees and providers to keep utilization below that which would otherwise occur. The extent to which various strategies are used will depend on the extent to which plans are in competition with other plans and the preferences of purchasers.

Nationalized systems can use similar strategies, limiting the resources available to provide care or providing financial and nonfinancial incentives to both providers and consumers to keep utilization at lower levels. Nationalized systems may also use their bargaining power to obtain favorable prices for medical care services.

3. Insurable Hazards

Given the analysis above, the theory of insurance would make several predictions regarding for which types of risks individuals would likely seek insurance coverage. First, they would seek insurance for services to treat events for which the probability of occurring is neither zero nor one. In other words, individuals would not want to buy insurance for protection against events for which they are not at risk, either because they know they will seek these services or because they know that they are at minimal risk of needing those services. Theory would predict that events with intermediate probabilities, that is, services with high-risk premiums, would generate the most demand. Thus the model would predict that most individuals would not purchase insurance to cover preventive services. The use of preventive services is not stochastic; that is, individuals know the probability of using these services in advance. Second, services that are costly would also have high risk premiums and would be attractive to insure. Third, individuals would not seek coverage for services for which demand is very elastic because such insurance would generate substantial welfare loss.

4. Adverse Selection

The preceding discussion was based on the utility curve for a single individual. To move the analysis to the market level, one must consider the behavior of many individuals, each with their own utility curve, initial wealth, and risk. Some individuals will be relatively low risk with low AFPs and others will be higher risk with high AFPs. Most models assume that individuals have more information about the risk that

they face than do the insurers. If an insurer were to base the premium it charged to a defined group of individuals on the estimated average AFP of all these individuals plus some mark-up, there would be some whose WTP was below the premium and others whose WTP was above the premium. If insurance purchase were voluntary, the individuals whose WTP was below the premium would opt not to purchase coverage, leaving a smaller pool of higher risk buyers. The average AFP for those remaining in the pool would rise. Of course as the AFP rose, the insurer would need to raise the premium in the next cycle. Some of the individuals who had remained in the first cycle would find the new premium above their WTP and would also opt out of buying coverage. This nonrandom selection out of the pool would result in an increase in the average AFP for the decreasing number of ever-higher risk individuals still interested in insurance.

Several papers outline scenarios in which insurance markets may collapse due to adverse selection (Akerlof 1970, Rothschild and Stiglitz 1976, Cutler and Reber 1998). In their seminal work, Rothschild and Stiglitz (1976) demonstrate that the set of policies that will be offered in the market is constrained by the ability of insurers to enter the market with new policies. This situation arises because insurers cannot perfectly ascertain the characteristics of the individuals they enroll. Certain policies cannot be offered in equilibrium because new insurers could enter the market and attract the lowest cost enrollees, leaving incumbent insurers with the higher cost enrollees, whose premiums do not cover their expected losses. In this environment, if an equilibrium does exist, it will be a separating equilibrium characterized by high risk individuals purchasing their most preferred actuarially fair policy and low risk individuals purchasing an insurance policy inferior to their most preferred actuarially fair policy, which will not be offered. Equilibrium is more likely to exist when there is a relatively large number of high-risk individuals and when the variation in risks is great.

Subsequent work indicates that the constraint on the set of policies offered depends on the equilibrium concept employed. Specifically, alternative equilibrium concepts can give rise to pooling equilibria (Wilson 1977) or even equilibria in which insurers may use profits on one product to subsidize losses on another (Spence 1978). All of the work in this area emphasizes the point that free entry into a market with asymmetric information limits the set of policies available because of adverse selection. Feldman et al. (1998) demonstrate that compulsory partial coverage may improve consumer welfare relative to a competitive equilibrium or compulsory comprehensive coverage.

Nationalized systems of insurance can reduce or eliminate adverse selection, depending on the degree to which coverage is compulsory and competition occurs within the system. Typically this would benefit

those with the highest expected expenditure, at the expense of those with less expected expenditures. The benefits of reducing adverse selection must be weighed against reductions in the ability of individuals to select their most preferred benefit design.

5. Empirical Evidence

5.1 Magnitude of Moral Hazard

The magnitude of moral hazard associated with the purchase of health insurance depends directly on the elasticity of the demand for medical care, i.e., how responsive consumers are to changes in the price. Manning et al. (1987) report estimates from a randomized trial indicating that the overall elasticity of medical care is -0.17 , that is, a 0.17 percent increase in utilization occurs for every 1 percent decrease in price. Demand for hospital services was estimated to be the least elastic (-0.14) and demand for well visits was estimated to be the most elastic (-0.43). These estimates are broadly consistent with other estimates derived from non-randomized settings (Feldstein 1999).

Estimated elasticities for other services tend to be somewhat higher than those for hospital and physician services. For example, elasticity of demand for pharmaceuticals has been estimated at about -0.40 (Smith and Garner 1974). Demand elasticity for nursing home care has been estimated at -0.76 (Lamberton et al. 1986).

5.2 The Demand for Insurance

The theory outlined above suggests that insurance demand is likely to be greatest for services with a high expected expenditure. In general, the evidence is consistent with the theory: services with a large estimated loss were more likely to be covered with insurance (Feldstein 1999). In addition, we would expect insurance demand to be greatest for services with highly inelastic demand. Again, evidence is consistent with theory. That is, the less price responsive the demand for the services, and therefore the smaller the moral hazard and associated welfare loss, the higher the level of insurance coverage (Phelps 1997).

Taken together, these factors explain much of the variation in service coverage witnessed among working adults in the U.S. For example, 76 percent of all employees participate in the medical care benefit offered by their employer, whereas only 59 percent participate in the dental care benefit, and 26 percent in the vision care benefit (USDOL, BLS, p. 6). Of those participating in the medical care benefit, 100 percent have coverage for hospital room and board, inpatient surgery, and inpatient physician visits; 85 percent have

coverage for home health care, 60–66 percent for well-baby care, physical exams, and hospice care, 52 percent for immunization and inoculation, and only 35 percent for hearing care (same source, p. 48).

Although measuring risk aversion is difficult, Barsky et al. (1997) used survey responses from the Health and Retirement Study regarding a respondent's willingness to take a new job with uncertain earnings to classify individuals based on their degree of risk aversion. They then related this measure of risk aversion to a variety of behaviors, including the purchase of health insurance. Stratifying by employment status (self-employed, employed, and not employed), they find that within each strata, the more risk averse respondents are more likely to purchase insurance.

5.3 Adverse Selection

Considerable evidence suggests that when given a choice individuals make health plan decisions based on their anticipated medical expenditures, thus generating nonrandom and, from the insurer's perspective, adverse selection. Much of the literature examines enrollment patterns when individuals face multiple insurance options that vary in terms of generosity. Even though a variety of data sources and methods have been used to investigate the self-selection into a range of plan options, the evidence suggests that adverse selection is quantitatively large (Cutler and Zeckhauser 2000).

6. Conclusions

Health insurance provides value to individuals by reducing the risk of a large financial loss associated with an adverse health event. This allows individuals to obtain medical care that they otherwise might not be able to afford. The payment from the insurer serves as a transfer of income to individuals who become ill.

However, if left unchecked, health insurance markets generate two deleterious effects. First, by lowering the patient's cost of consuming medical care, insurance leads to an increase in utilization (moral hazard). This increased consumption results in welfare loss. Second, insurance markets have a tendency to be unstable because of adverse selection. In the extreme, non-random selection could cause insurance markets to unravel, leaving some or all individuals uncovered and distorting the set of plans offered in the market.

Most health insurance systems, whether private or public, incorporate a variety of strategies to limit the welfare loss from the overconsumption of medical care and the effects of adverse selection. Inherently these strategies involve restrictions on individual behaviors.

The challenge in most systems is to mitigate the unwanted consequences of insurance as much as possible while still accomplishing the fundamental goal of reducing risk.

Bibliography

- Akerlof G A 1970 The market for 'lemons.' *Quarterly Journal of Economics* **84**: 488–500
- Arrow K 1963 Uncertainty and the welfare economics of medical care. *American Economic Review* **53**: 941–73
- Barsky R B, Juster F T, Kimball M S, Shapiro M D 1997 Preference parameters and behavioral heterogeneity: An experimental approach in the health and retirement study. *The Quarterly Journal of Economics* (May): 537–79
- Cutler D M, Reber S J 1998 Paying for health insurance: The trade-off between competition and adverse selection. *The Quarterly Journal of Economics* **113**(2): 433–66
- Cutler D M, Zeckhauser R J 2000 The anatomy of health insurance. In: Culyer A J, Newhouse J P (eds.) *Handbook of Health Economics*. Elsevier, Amsterdam, Vol. 1A, Chap.11, pp. 564–643
- Deaton A, Muelbauer J 1987 *Economics and Consumer Behavior*. Cambridge University Press, Cambridge, UK
- Feldman R, Escrignano C, Pellise L 1998 The role of government in health insurance markets with adverse selection. *Health Economics* **7**: 659–70
- Feldstein P J 1999 *Health Care Economics*, 5th edn. In: Williams S J (ed.). Delmar publishers.
- Lamberton C E, Ellingson W D, Spear K R 1986 Factors determining the demand for nursing home services. *Quarterly Review of Economics and Business* **26**: 74–90
- Manning W G, Newhouse J P, Duan N, Keeler E B, Leibowitz A, Marquis M S 1987 Health insurance and the demand for medical care: Evidence from a randomized experiment. *The American Economic Review* **77**(3): 251–77
- Miyazaki H 1977 The rat race and internal labor markets. *Bell Journal of Economics* **8**(3): 394–418
- Newhouse J P 1992 Medical care costs: How much welfare loss? *Journal of Economic Perspectives* **6**(3): 3–21
- Nyman J A, Maude-Griffin R 2000 The welfare economics of moral hazard. Presented at the 11th Annual Economics Conference, Washington, DC
- Nyman J A, Maude-Griffin R 2001 The welfare economics of moral hazard. *International Journal of Health Care Finance and Economics*. **1**: 23–42
- Pauly M V 1968 The economics of moral hazard. *American Economic Review* **58**(3): 531–7
- Phelps C E 1997 *Health Economics*, 2nd edn. Addison-Wesley, Reading, MA
- Rothschild M, Stiglitz J 1976 Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics* 629–49
- Smith M C, Garner D D 1974 Effects of a medicaid program on prescription drug availability and acquisition. *Medical Care* **12**(7): 571–81
- Spence A M 1978 Product differentiation and performance in insurance markets. *Journal of Public Economics* **10**(2): 427–47
- US Department of Labor, Bureau of Labor Statistics 1997 *Employee Benefits in Medium and Large Private Establishments, 1993, 1994*

Wilson C 1977 A model of insurance markets with incomplete information. *Journal of Economic Theory*. **16**: 167–207

C. G. McLaughlin and M. E. Chernew

Health Interventions: Community-based

This article considers theories and methods underlying successful community-based health interventions, emphasizing four exemplars. Three exemplars relied heavily on modern means of mass communication and social marketing, relatively low-cost approaches with the potential to reach and change lifestyle behaviors of entire populations, in contrast to traditional individual or group counseling. One exemplar, studying alcohol-involved trauma, used community organizing for adherence to existing laws, rather than public education for behavior change. The article also describes barriers to change and lessons learned for future extensions of such endeavors.

By definition a community-based program is organized locally, achieving its goals through the community's institutions and communication networks. A community here is defined as a residential area with legally defined geographic boundaries, where a governmental system regulates certain local aspects of schools, businesses, transportation, policing, and recreational activities. A community is ordinarily last in a nation's regulatory chain, where education must ultimately occur, although for rural areas the *county* becomes the governing agent for education.

The relevance and intellectual context of community-based health education are addressed through these questions. (a) What is the rationale for community-wide interventions rather than in more limited locales, such as clinics, hospitals, work sites, or schools? (b) What theories, methods, and resources underlie success? (c) How does a community become committed to a novel health education campaign? (d) How can barriers to success be overcome in such interventions? (e) What roles do community activism, laws, and regulations play? (f) What is the future of community-based health education?

1. Rationale

Health education in schools, work sites, and clinics have a long history of individual success, but synergistic interactive effects can occur when they are imbedded in a total community campaign that adds inherently cost-effective mass media and environmental change (Schooler et al. 1997). Lifestyle changes, so powerfully influenced by custom and by the media in developed countries, cannot be countered effectively through simple means. Community-wide approaches fit the public health model because the traditional medical model can neither prevent most chronic

disease nor reach the entire population in need. The community provides influence through locally produced electronic and print mass media and through work sites, schools, medical and hospital settings, and social gatherings of many sorts. It offers opportunities for health-promoting regulations, such as providing opportunities for physical activity for all, school fitness classes, healthful school lunches, alcohol sales limits, and preventing tobacco product marketing to children (see *Health Promotion in Schools*).

Studies at the Stanford Center for Research in Disease Prevention (SCRDP) have shown that only multiple influences produce meaningful changes in the dietary, exercise, or tobacco-use behavior of adults, adolescents, and children. For adolescents, always resistant to health behavior change, the following influenced success: parent and teacher roles; school environments; amount and quality of school-based education on tobacco, fitness, and nutrition; peer influence; and amount and quality of community-wide health education. (Also, certain personal characteristics were influential, for example, self-efficacy toward one's behavior-change abilities; Fortmann et al. 1995.)

2. Theoretical Formulations

Community-based health education carried out by SCRDP and by some analogous projects have been guided by three major theoretical and management models (Flora et al. 1989).

(a) The community organization framework provides methods of identifying the health problem (and resources needed or available), mobilizing the community's opinion leaders and organizations, gaining populace support, forming coalitions, launching and maintaining education programs, achieving regulatory changes, and empowering communities to reach and maintain their goals.

(b) Social marketing provides methods of designing and sequencing messages for population subgroups based on their health needs, cultural attributes, social networks, media habits, and readiness for change.

(c) The mass communication-behavior change model describes theories underlying educational content, such as social cognitive theory (Bandura 1986) and methods of reaching and influencing the total population in multiple ways (see *Self-efficacy and Health*).

3. Intervention Principles, Methods, and Resources Needed for Success

3.1 Acceptance

Both relevant organizations and political and opinion leaders must be recruited into a coalition that sanctions any novel health education project. This

coalition must identify the health problem (needs assessment), provide a resource inventory, obtain populace support, and plan the intervention.

3.2 Comprehensive

Interventions must go beyond attempting to change knowledge, the usual goal of an educational system, by providing training in behavior-change skills. They must also go well beyond the individual, enlisting multiple community organizations in campaigns for change and seeking changes in the social environment and in regulations that promote access to the facilities and resources needed for healthful practices. Multiple communication channels (i.e., both mediated and face-to-face) are needed to reach different audiences, recognizing their preferences, stage of knowledge, and readiness for change.

A comprehensive intervention should involve schools, work sites, churches, and facilities for sport, recreation, and health. These organizations and others can serve as education conduits, with the community's electronic and print mass media organizations assisting in message design and delivery. The Internet provides a new channel, whose community education role is as yet poorly defined. Interactive computer learning, now becoming common in classrooms, can be designed for large groups—an emerging variant of mass media.

3.3 Variety

Comprehensiveness requires variety. For example, in tobacco control components of Stanford's CVD campaigns, local medical clinics, dental offices, pharmacies, and libraries distributed a low-cost skills-training Quit Kit; a local smoking cessation class, was shown on television; many newspaper articles and columns appeared; a local business supported costs of a smoking cessation contest; and all newspapers and electronic media 'cross-advertised' activities designed for mass audiences.

3.4 Integration

Success requires an appropriate mix and sequence of programs delivered through varied channels. This integration, with goals set in advance and goal changes based on early results, is analogous to a commercial marketing campaign, hence the term 'social marketing.'

3.5 Message Design and Audience Segmentation

Social marketing requires message tailoring to fit each subgroup's needs and preferences, taking into con-

sideration cultural differences, learning styles, and preferred learning sites. A message sequence should increase awareness, then, increase knowledge, and last, increase motivation and provide training in the skills needed for adoption and maintenance of a new behavior (Bandura 1986). Electronic media can carry out the first two parts of this sequence and stimulate use of the more information-dense print media of newspapers and booklets, which are inherently more effective in skills training than are electronic media (Flora et al. 1997).

Messages must be clear, focused, and salient. Salience requires broad reaching media, arousing interest and awareness—topics must break through passive indifference engendered by the information overload of many societies and become 'on the public agenda.' Given the large advertising budgets of today's mass media, health agencies' messages must be of sufficient production quality to compete for the public's attention.

3.6 The Message Must be Effective

Behavioral research confirms the rather self-evident rule that 'learning by doing' is more effective than 'learning from observing' (modeling a behavior), and both are more effective than an 'information-only' approach that changes knowledge alone. These principles are contained in Bandura's social cognitive theory, which posits that guided practice in a new behavior can lead to increased self-efficacy and to greater behavior change (Bandura 1986). Thus, 'knowledge-only' campaigns have been found less effective than those that apply Bandura's recommendations (see *Self-efficacy and Health*).

3.7 Intervention Dose

The dose needed depends on many factors: lesser amounts are needed in smaller communities, at earlier stages in a country's adoption of a 'health innovation,' and when the advocated behavior change is reasonably simple (such as mammography and hypertension screening and immunization campaigns). Clearly, more complex changes are needed in individuals and in society's norms to alter eating or exercise patterns or to control tobacco use. Complexity in respect to nutrition arises from many sources, including longstanding cultural beliefs and practices; entrenched methods in agriculture, food production, and retailing; advertising of 'unhealthful' foods; and the advent of widespread fast-food chains that are dominated by commercial interests unresponsive to local demands and needs.

Few projects have measured intervention dose. One excellent method records the total number and duration of messages distributed over a defined time period, albeit with a defect due to lack of message

quality measures (Farquhar et al. 1990). However, even this rough method, were it widely adopted, would be of great value to public health practitioners.

3.8 Role of Activism and Advocacy

Successful community-based health education requires effective leaders, community activists with the courage and charisma to advocate health innovations. Waves of advocacy filtering down from international, national, state, or provincial sources can wash up on the shore of an inert community to provide a local activist leader with the popular support to 'fight city hall' or other entrenched bureaucrats who defend the status quo. Tobacco control in Australia and the US provide examples. National and state advocacy groups with access to mass media created a strong mass movement for change, allowing advocates to enlist popular support for local tobacco control measures. In both California and Australia's State of Victoria this popular support led to statewide increases in tobacco taxes, with some retained for education against tobacco, a measure that had been resisted by state legislators who had long been influenced by tobacco lobbyists—an example of community activism filtering up the chain to a higher political level (Catalonia Declaration 1996).

3.9 Role of Policy, Laws, and Regulations (PLR)

Local actions can affect alcohol and tobacco sales and create environments that improve nutrition and enhance physical activity. However, national, state, or provincial actions can magnify local PLR and education efforts on topics such as tobacco taxation, automobile seat-belt laws, food and drug safety, school nutrition, school physical activity policies, and (in the United States) laws on firearms. As described in Section 4.8, widespread popular attitude changes in numerous communities can also affect the political process at the state or federal level.

4. History of Comprehensive Community Health Education

The history since the 1970s is restricted largely to 13 formal research projects designed to affect CVD risk factors. They involved entire populations of at least one education community, compared with at least one control community. Three of the four exemplars to follow are drawn from these 13 projects. Dissemination into practical applications of community organizing and mass communication technologies, derived in part from these research projects, occurred throughout many parts of the world.

4.1 The First Two Exemplars: The First Decade

The first and second exemplars, Stanford Three-community Study (TCS), in three small agricultural marketing towns in California (total population 45,000), and the North Karelia Study (NKS), in two adjoining predominately rural Finnish counties (North Karelia population about 180,000), each began in 1972. TCS, the first Stanford project, was carried out from 1972–75 in both English and Spanish, comparing effects of mass media alone in one community and mass media plus 10-session risk reduction classes for high-risk adults in a second, with a third as a control (Schooler et al. 1997). Groups exposed to varied education amounts showed a dose-response change in smoking, blood pressure, and blood cholesterol, with a proportionately larger effect in the Spanish-speaking residents than in the Anglo majority. (This minority population outcome required a compensating increase in intervention resources.) A composite CVD risk reduction of about 23 percent and 30 percent occurred in the mass media—only and mass media—plus classroom conditions, respectively. Thus, a relatively modest amount of mass media (about 30 television and radio 'spots,' weekly newspaper columns on heart health, and four separate mass mailings of booklets) was sufficient to change the population's body weight, cholesterol and blood pressure levels, and smoking prevalence. The education followed Bandura's principles but adapted to the mass media (as was also the case with exemplars two and three).

NKS, evaluated two, matched, rural Finnish counties that contained many villages with farming and lumbering as the main occupations. North Karelia (about 180,000 population) received an education campaign that began in 1972, continuing to the present. After 5 years, CVD risk factor changes comparable to the TCS occurred, and significant net reductions in CVD events occurred (Schooler et al. 1997). This study was marked by extensive community organizing, resulting in strong partnerships with residents and their organizations. The NKS influence on its country's policies was unparalleled among the CVD projects, providing its most important lesson—that a well executed project led by respected scientists can move an entire country. As examples, Finland's food and agricultural industries made large changes: Fertilizers were supplemented with selenium (a substance low in Finland's soil that is needed for health), milk pricing was changed (based on protein instead of fat content), programs were created to replace dairy farms with berry farms, a new canola industry replaced jobs lost in dairying, and increased production of low-fat foods occurred (Catalonia Declaration 1996).

In 1972 population-wide nutrition change and smoking cessation interventions were an innovation internationally, which may partly explain the success of these two pioneering programs (TCS and NK).

4.2 Early International Diffusion, 1977–1983

Studies patterned after the TCS done in Italy, Australia, Switzerland, and South Africa all had significant risk factor changes, adding evidence for the effectiveness of the TCS model (Farquhar et al. 1991).

4.3 The Second and Third Decades: Projects Begun in the 1980s and 1990s

The third exemplar is the Stanford Five-city Project (FCP) (USA). Its intervention phase from 1980 to 1986 extended TCS methods to larger populations (total population of about 360,000) with multifactor CVD prevention directed at two northern California cities. There were three control cities (Farquhar et al. 1990). It differed from TCS in greater use of community organizing and in greater collaboration with the communities' health, media, and education organizations in planning and implementing programs. It was similar in generous use of mass media, both print and electronic (63 percent television and 36 percent print). It was unique in measuring total dose of education (about 5 hours/year and about 100 episodes of exposure/year to all forms of media and classroom education).

FCP's initial year of television messages (defined as 'high reach/low involvement') stimulated the public's use of print media, which supplied more effective skills training than television (Flora et al. 1997). Results were comparable with the TCS (about a 15 percent fall in composite risk of CVD), with a major impact on blood pressure (Farquhar et al. 1990). Health bureaucracies, usually timid, should gain courage from the FCP's 'David and Goliath' demonstration that only 3 hours/year of high-quality television health education can counteract the public's exposure to about 100 hours/year of television advertising devoted to unhealthy nutrition. Although all preceding CVD studies showed effects in small towns and/or rural districts, FCP showed benefit in cities (with populations as high as 100,000). Also, these effects occurred despite the advent (at least in the state of California) in the 1980s of dual working families and increasing public use of fast food, factors that made achieving nutrition-behavior change more difficult.

FCP's modest resources and greater effects, as compared with some comparable projects, support the benefits of the skills training aspects of Bandura's social cognitive theory. Perhaps the most practical lesson to policymakers is that adult residents saved 30 times more money from their decreased cigarette purchases (\$120/adult/year) than the cost of the campaign (\$4/adult/year)—savings retained by the community. Lastly, the communities adopted FCP's technologies, later applying them to seat-belt promotion and violence and adolescent pregnancy prevention.

Other successful CVD projects, both large and small, occurred in these decades in the US, Sweden, Denmark, Canada, the Czech Republic, and China. Also, World Health Organization-sponsored projects began in about 23 other countries (Schooler et al. 1997). In all instances, they borrowed heavily from the experiences of the three exemplars and in many instances received training from either the Stanford or the North Karelia groups.

4.4 Community Projects in Other Health Topics

Interventions on alcohol, mammography, tobacco control, motor vehicle injuries, and HIV/AIDS are prominent examples of other community-based projects, but with fewer well-controlled studies than in CVD. One of these, Preventing Alcohol Trauma (PAT), is the fourth exemplar. This 5-year study of three US communities showed a 10 percent decrease in alcohol-related traffic injuries and a 50 percent decrease in adolescent alcohol use (Holder et al. 1997). Coalition building and organizational behavior change among the police, alcohol sales outlets, and alcohol servers were the main interventions. The public responded to a fear, instilled through publicity, of the penalties of greater enforcement of existing laws on underage alcohol purchases or drinking and driving. Therefore, in contrast to the three CVD exemplars, PAT showed that major public education is not needed for large public behavior changes. PAT estimated a cost saving of \$2.88 for every dollar invested, a number close to that found in many work-site health promotion studies. PAT found their communities had the required infrastructure for the campaigns, requiring only training provided by one indigenous community coordinator, a part-time clerk, an imaginative plan, and the will to proceed.

TCS, the first, and PAT, the fourth exemplar, are opposites: TCS had a maximum of mass-media education and a minimum of organizing, whereas PAT had the opposite. Thus, either model works, but to change complex behaviors, one needs sophisticated behavior-change methods, such as those adapted from Bandura (see *Self-efficacy and Health*). If fear of arrest for breaking existing laws on alcohol use suffices to change personal drinking-and-driving habits, then the education process is simpler.

5. Lessons Lead to a 'Master Plan'

A five-step approach emerges from these studies.

(a) Define the problem. Does the community need a health education campaign? A decision can be made from national or local survey data.

(b) Organize the community, creating a campaign that includes education, continued organizing, training of community organizations, empowerment of the public and the community's organizations, and future

maintenance of programs developed during an initial phase of about 2 years.

(c) Implement an intervention, delivering 3–5 hours of education exposure/year for 2 years, using initially about 50 percent television to arouse interest, followed by about 25 percent from more specific print media (emphasizing newspapers, if available) and 25 percent from community events and programs (such as health fairs, contests, and classes).

(d) Institutionalize programs. The community becomes a demonstration project, with its 'empowered' organizations functioning as health promotion resource centers for a wider region.

(e) Use the new community resources and its residents' potential power to advocate for local, regional, and national governmental regulations and laws that will increase local intervention effects and extend them beyond the community.

Any individual or organization that wishes to engage in community-based health education should gain courage from the words of Margaret Mead, as quoted in the closing passage of the Catalonia Declaration (1996, p. 75), 'Never doubt the capacity of a few dedicated individuals to change the world, in fact, it is the only way it ever has.'

6. Barriers and Challenges

A healthy city has been defined by the absence of crime, crowding, and poverty and by the presence of educated residents and enlightened (and trained) organizations. Together these lead to a community empowered to solve its social problems (i.e., to increase its social capital) (Travers 1997).

Considering barriers, MacIntyre (2000) found Glasgow's environmental factors to be major barriers to healthful exercise behavior. Certain macro-economic factors inherent in globalization, such as capital flight and increased wealth and income gaps, have been described as barriers to planned change (Bezruchka 2000, Cahill 1983). All such barriers threaten community stability, inhibiting greatly the success of health promotion attempts. However, wise compensatory resource allocation can overcome many barriers, as was shown in the Hispanic minority of the TCS. Therefore, the challenge for the future is: Responsibility for success in community-based health interventions lies with the interventionist, not with the community's residents!

7. Conclusions

Three decades of the 'total community' health education approach strongly support the feasibility, at relatively low cost, of achieving transfer of public education technologies to a community's infrastructure (public health, media, schools, etc.), resulting in significant changes in health habits of populations.

Although most studies derive from small communities, recent successes in Tianjin (China), a city of 400,000 (two exemplars examined populations of > 100,000), suggest that the model also works in large populations (Schooler et al. 1997). Organizing and educating communities requires advocacy, activism, coalition building, and leadership; success is enhanced by regulatory change. Theory matters: when the population gains self-efficacy through education, the result—community efficacy—enhances capacity to change institutional policy and practice, thus maintaining community change.

Science cannot serve society if its evidence for educational benefit is ignored. This is not a new concept. As written over 2,000 years ago, 'If a virtuous and learned scholar aims to influence the people as a whole, one must first educate the people' (Legge 1983).

See also: Community Health; Drugs: Illicit Use and Prevention; Health Behaviors; Health Education and Health Promotion; Health Psychology; Mental Health: Community Interventions; Public Health; Public Health as a Social Science; Smoking Prevention and Smoking Cessation; Socioeconomic Status and Health Care; Substance Abuse in Adolescents, Prevention of; Sun Exposure and Skin Cancer Prevention

Bibliography

- Bandura A 1986 *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice-Hall, Englewood Cliffs, NJ
- Bezruchka S 2000 Is globalization dangerous to our health? *Western Journal of Medicine* **172**: 332–4
- Cahill J 1983 Structural characteristics of the macroeconomy and mental health: Implications for primary prevention research. *American Journal of Community Psychology* **11**: 553–71
- The Catalonia Declaration: Investing in Heart Health 1996 Declaration of the Advisory Board, Second International Heart Health Conference 1995. Department of Health and Social Security of the Autonomous Government of Catalonia, Barcelona, Spain
- Legge J 1983 *The Chinese Classics: Confucian Analects, The Great Learning, The Doctrine of the Mean*. Southern Materials Center, Taipei, Republic of China
- Farquhar J W, Fortmann S P, Flora J A, Maccoby N 1991 Methods of communication to influence behaviour. In: Holland W W, Detels R, Knox G (eds.) *Oxford Textbook of Public Health*, 2nd edn. Oxford University Press, Oxford, UK, Vol. 2
- Farquhar J W, Fortmann S P, Flora J A, Taylor C B, Haskell W L, Williams P T, Maccoby N, Wood P D 1990 The Stanford Five-city Project: Effects of community-wide education on cardiovascular disease risk factors. *Journal of the American Medical Association* **264**: 359–65
- Flora J A, Maccoby N, Farquhar J W 1989 Communication campaigns to prevent cardiovascular disease: The Stanford studies. In: Rice R, Atkin C (eds.) *Public Communication Campaigns*. Sage, Newbury Park, CA, pp. 233–52
- Flora J A, Saphir M N, Schooler C, Rimal R N 1997 Toward a framework for intervention channels: Reach, involvement and impact. *Annals of Epidemiology* **S7**: S104–S112

- Fortmann S P, Flora J A, Winkleby M A, Schooler C, Taylor C B, Farquhar J W 1995 Community intervention trials: Reflections on the Stanford Five-city Project. *American Journal of Epidemiology* **142**(6): 576–86
- Holder H D, Salz R F, Grube J W, Treno A J, Reynolds R I, Voas R B, Gruenewald P J 1997 Summing up: Lessons from a comprehensive community prevention trial. *Addiction* **92**(S2): S293–S301
- MacIntyre S 2000 The social patterning of exercise behaviours: The role of personal and local resources. *British Journal of Sports Medicine* **34**: 6
- Schooler C, Farquhar J W, Fortmann S P, Flora J A 1997 Synthesis of findings and issues from community prevention trials. *Annals of Epidemiology* **S7**: S54–S68
- Travers K D 1997 Reducing inequities through participatory research and community empowerment. *Health Education and Behavior* **24**: 344–56

J. W. Farquhar

Health Outcomes, Assessment of

Health is probably the most valued human asset. In fact, studies on the preference for different states of being show that virtually everyone rates health as most important. Despite the perceived importance of health, health status has remained difficult to define. There are two common themes in definitions of health. First, premature death is undesirable, so one aspect of health is the avoidance of mortality. The health status of nations is often evaluated in terms of overall mortality rates or infant mortality rates. Second, quality of life is also very important. Disease and disability are of concern because they affect either life expectancy or life quality. For example, cancer and heart disease are the two major causes of premature death in the United States. In addition, disease or disability can make life less desirable. A person with heart disease may face restrictions in daily living activities and may be unable to work or participate in social activities. Even relatively minor diseases and disabilities affect quality of life. A cold, for example, may interfere with the ability to concentrate, work, or attend school. The cold, however, lasts only a short time. A chronic disease, such as arthritis, may affect the quality of your life for a long time (Brown et al. 2000).

Within the last few years, medical scientists have come to realize the importance of measuring patient-reported health. Many major diseases, including arthritis, heart disease, and diabetes, or digestive problems are evaluated in terms of the degree to which they affect life quality and life expectancy (Koloski et al. 2000). One can also evaluate treatments for these conditions by the amount of improvement they produce in everyday life. The Food and Drug Admin-

istration now considers patient-reported outcomes in their evaluations of new products and nearly all major clinical trials in medicine use quality of life assessment measures. Several approaches to health outcome measurement will be reviewed. Chapter 130 by O'Boyle covers quality of life measures in more detail.

1. Historical Perspective

For many years, researchers and clinicians have struggled to reach consensus on the measurement of health and health outcomes. Different definitions emphasize related, but divergent, concepts. Some terms used to describe health are positive concepts, such as wellness or normality. Other definitions emphasize negative concepts, such as disability and illness. Much of the debate has centered on whether health is a continuum. Are disability and illness distinct from health, or are they opposite ends of the same continuum? (Patrick and Erickson 1993).

Because the concept of health lacked conceptual clarity, the World Health Organization (WHO) proposed a comprehensive definition of health in their charter document. Health was defined as, '... a state of complete physical, mental, and social well-being and not merely the absence of infirmity' (World Health Organization 1948). Since the introduction of this definition there has been some convergence of thought. It is now widely recognized that health has multiple dimensions. Further, it is now accepted that measures of mortality alone cannot summarize the health status of populations (Field and Gold 1998). However, there is still considerable debate about what constitutes health and health status.

Some of the debate about the definition of health originates from sociologists. From the sociological perspective, illness represents deviation from society's standards for physical and mental well-being. Often, this deviation is recognized when individuals cannot perform activities usual for their social roles. In addition to observable functioning, illness may cause reports of symptoms of pain and these deviate from society's norms on wellness. Parsons defined illness as '... a state of disturbance in normal functioning of the total human individual including both the state of the organism as a biological system, and/or his personal and social adjustments' (Parsons 1951, p. 431). For Parsons, health represented the capacity to perform valued tasks.

The traditional biomedical model is predicated on finding a specific biologic problem and repairing it. In contrast, measures of health outcome consider benefit from the patient's perspective. Successful treatments are those that improve the quality of life or extend life. Most surgical interventions are justified on the basis of saving lives. However, the great majority of surgical procedures have no effect on life expectancy. For

example, four of the most common indications for surgery in the United States are transurethral resection prostatectomy (TURP) for benign hyperplasia of the prostate (BPH), extractions of cataracts, joint replacement for severe osteoarthritis of the knee or hip, and hysterectomy for diseases of the uterus. Although, it is often argued that these procedures prevent deaths, reviews of the evidence suggest that these procedures are rarely performed to extend life expectancy. In each case, however, the surgery may lead to improvements in functioning or the reporting of symptoms. For example, TURP may reduce the symptom of urinary frequency in men. Cataract extraction may help older patients function better with reading, night-driving, and other activities of daily living. Joint arthroplasty may increase mobility for older patients. For most of these cases, traditional physiological measures are limited as outcome assessment. Patients seek these procedures because they want relief of symptoms. Outcomes assessment must consider whether the symptoms are, indeed, relieved. Ultimately, measures of health outcome can be used to evaluate the benefits of treatment from the patient's perspective.

2. Assessment of Health Outcomes

Methods of assessment of health-related quality of life represent at least two different conceptual traditions. One grows out of the tradition of health status measurement. Several efforts to develop measures of health status were launched in the late 1960s and early 1970s. All the projects were guided by the World Health Organization's (WHO) definition of health status. The projects resulted in a variety of assessment tools, including the Sickness Impact Profile (Bergner et al. 1981) the Quality of Well-Being Scale (Kaplan et al. 1996) the SF-36 (Ware and Gandek 1998) and the Nottingham Health Profile (Lowe et al. 1990). Many of the measures examined the effect of disease or disability on the performance of social roles, the ability to interact in the community, and physical functioning. Some of the systems have separate components for the measurement of physical, social, and mental health. The measures also differ in the extent to which they consider subjective aspects of life quality (Brown et al. 2000).

Perhaps the most important distinction between methods to assess quality of life is the contrast between psychometric and decision theory approaches. The psychometric approach attempts to provide separate measures for the many different dimensions of quality of life. Perhaps the best known example of the psychometric tradition is the Sickness Impact Profile (SIP). The SIP is a 136-item measure that yields 12 different scores displayed in a format similar to an MMPI profile (Bergner et al. 1981).

The decision theory approach attempts to weight the different dimensions of health in order to provide

a single expression of health status. Supporters of this approach argue that psychometric methods fail to consider that different health problems are not of equal concern. A runny nose is not the same as severe chest pain. In an experimental trial using the psychometric approach, one will often find that some aspects of quality of life improve while others get worse. For example, a medication might reduce high blood pressure but also produce headaches and impotence. Many argue that the quality of life notion is the subjective evaluation of observable or objective health states. The decision theory approach attempts to provide an overall measure of quality of life that integrates subjective function states, preferences for these states, morbidity, and mortality.

3. Common Methods for the Measurement of Quality of Life

A variety of methods have been proposed to measure quality of life, but we cannot review and critique them all here. Instead, we present some of the most widely used psychometric and decision theory based methods. Readers interested in more detailed reviews should consult Walker and Rosser (1993) and McDowell & Newell (1996). Quality of life measures are also covered in Chapter 130 (by O'Boyle).

3.1 Psychometric Methods

(a) *SF-36*. Perhaps the most commonly used outcome measure in the world today is the Medical Outcome Study Short Form-36 (SF-36). The SF-36 grew out of work by the RAND Corporation and the Medical Outcomes Study (MOS) (Ware & Gandek 1998). Originally, it was based on the measurement strategy from the RAND Health Insurance Study. The MOS attempted to develop a very short, 20-item instrument known as the Short Form-20 or SF-20. However, the SF-20 did not have appropriate reliability for some dimensions. The SF-36 includes eight health concepts: physical functioning, role-physical, bodily pain, general health perceptions, vitality, social functioning, role-emotional, and mental health (Kosinski et al. 1999). The SF-36 can either be administered by a trained interviewer or self-administered. It has many advantages. For example, it is brief, and there is substantial evidence for its reliability and validity. The SF-36 can be machine scored and has been evaluated in large population studies. The reliability and validity of the SF-36 are well documented (Keller et al. 1999, Scott-Lennox et al. 1999, Stewart and Ware 1992, Ware and Gandek 1998).

Despite its many advantages, the SF-36 also presents some disadvantages. For example, it does not have age-specific questions and one cannot clearly

determine whether it is equally appropriate at each level of the age continuum. The items for older retired individuals are the same as those for children (Stewart and Ware 1992). Nevertheless, the SF-36 has become the most commonly used behavioral measure in contemporary medicine.

(b) *Nottingham Health Profile.* The Nottingham Health Profile (NHP), is another profile approach that has been used widely in Europe. One of the important features of the NHP is that the items were originally generated on the basis of extensive discussions with patients. The NHP has two parts. The first includes 38 items divided into six categories: sleep, physical mobility, energy, pain, emotional reactions, and social isolation. Items within each of these sections are rated in terms of relative importance. Items are rescaled in order to allow them to vary between 0 and 100 within each section. The second part of the NHP includes seven statements related to the areas of life most affected by health: employment, household activities, social life, home life, sex life, hobbies and interests, and holidays. The respondent indicates whether or not a health condition has affected his or her life in these areas. Used in a substantial number of studies, the NHP has considerable evidence for its reliability and validity.

An important strength of the NHP is that it is based on consumer definitions of health derived from individuals in the community. The language in the NHP is simple and the scale requires only a low level of reading ability. Psychometric properties of the NHP have been evaluated in a substantial number of studies. However, the NHP, like most profile measures, does not provide relative-importance weightings across dimensions. As a result, it is difficult to compare the dimensions directly with one another (Lowe et al. 1990).

3.2 Decision Theory Approaches

Within the last few years, interest has grown in using quality of life data to help evaluate the cost/utility or cost-effectiveness of healthcare programs. Cost studies have gained in popularity because healthcare costs have grown so rapidly in recent years. Not all healthcare interventions equally return benefit for the expended dollar. Objective cost studies might guide policymakers toward an optimal and equitable distribution of scarce resources. Cost-effectiveness analysis typically quantifies the benefits of a healthcare intervention in terms of years of life, or quality-adjusted life years (QALYs). Cost/utility is a special case of cost-effectiveness that weights observable health states by preferences or utility judgments of quality. In cost/utility analysis, the benefits of medical

care, behavioral interventions, or preventive programs are expressed in terms of years of life adjusted for reduced quality of life or quality-adjusted life years (QALYs).

If a man dies of heart disease at the age of 50 and we expected him to live to age 75, we might conclude that the disease precipitated 25 lost life-years. If 100 men died at age 50 (and also had a life expectancy of 75 years), we might conclude that 2,500 (100 men $\times$ 25 years) life-years had been lost. Yet death is not the only relevant outcome of heart disease. Many adults suffer myocardial infarctions that leave them somewhat disabled for a long time. Though still alive, they suffer diminished quality of life. Quality-adjusted life-years take into consideration such consequences. For example, a disease that reduces quality of life by one-half will take away 0.5 QALY over the course of each year. If the disease affects two people, it will take away one year (2×0.5) over each year. A medical treatment that improves quality of life by 0.2 for each of five individuals will result in the equivalent of 1 QALY if the benefit persists for one year. This system has the advantage of considering both benefits and side effects of programs in terms of the common QALY units.

The need to integrate mortality and quality of life information is clearly apparent in studies of heart disease. Consider hypertension. People with high blood pressure may live shorter lives if untreated, longer if treated. Thus, one benefit of treatment is to add years to life. However, for most patients, high blood pressure does not produce symptoms for many years. Conversely, the treatment for high blood pressure may cause negative side effects. If one evaluates a treatment only in terms of changes in life expectancy, the benefits of the program will be overestimated because one has not taken side effects into consideration. On the other hand, considering only current quality of life will underestimate the treatment benefits, because information on mortality (death) is excluded. In fact, considering only current function might make the treatment look harmful because the side effects of the treatment might be worse than the symptoms of hypertension. A comprehensive measurement system takes into consideration side effects and benefits and provides an overall estimate of the benefit of treatment (Russell 1986).

Of the several different approaches for obtaining quality-adjusted life years, most are similar. The three most commonly used methods are the EQ-5D, the Health Utilities Index (HUI) and the Quality of Well-being Scale (QWB).

(a) *EQ-5D.* The approach most commonly used in the European community in the late twentieth century was the EQ-5D. This method has been developed by a collaborative group from Western Europe, led by Paul Kind and associates, known as the EuroQol group

(Kind 1997). The intention of this effort was to develop a generic currency for health that could be used commonly across Europe. The concept of a common EuroQol was stimulated by the desire for a common European currency—the Euro dollar. The original version of the EuroQol had 14 health states in six different domains. In addition, respondents placed their health on a continuum ranging from death (0.0) to perfect health (1.0). The method was validated in postal surveys in England, Sweden, and the Netherlands. More recent versions of the EuroQol, known as the EQ-5D, are now in use in a substantial number of clinical and population studies (Gudex et al. 1996, Hurst et al. 1997). Although the EQ-5D is easy to use and comprehensive, there have been some problems with ceiling effects. Substantial numbers of people obtain the highest possible score.

(b) *Health Utilities Index*. Another approach has been developed in Canada by Torrance et al. (1996). This method, known as the Health Utilities Index (HUI), is derived from micro-economic theory. There have been several versions of the measure, typically identified by 'Mark.' The HUI Mark I was developed for studies in the neonatal intensive care unit. The measure had 960 unique health states. In 1992, the HUI Mark II was developed and included 24,000 unique health states. The HUI Mark III, released in 1995, had 972,000 health states. Eight components of the HUI Mark III include vision (six levels), hearing (six levels), speech (five levels), ambulation (six levels), dexterity (six levels), emotion (five levels), cognition (six levels), and pain (five levels). Multiplying the number of levels across the eight dimensions gives the 972,000 states.

Using multiattribute utility scaling methods, judges evaluate levels of wellness associated with each level of each domain. A multiattribute model is used to map preference for the 972,000 possible states on to the 0.0 to 1.0 continuum. The HUI has been used in many population and clinical studies. For overall health status, men obtain higher scores early in the life cycle. However, after about the age of 45, women obtain higher scores and this difference grows systematically through the remainder of the life span (Kaplan and Erickson 2000).

(c) *Quality of Well-being Scale (QWB)*. A third method, known as the Quality of Well-being Scale integrates several components into a single score. First, patients are classified according to objective levels of functioning. These levels are represented by the scales of mobility, physical activity, and social activity. Once observable behavioral levels of functioning have been classified, each individual is placed

on the 0 to 1.0 scale of wellness, which describes where a person lies on the continuum between optimum function and death.

To accomplish this, the observable health states are weighted by quality ratings for the desirability of these conditions. Human value studies have been conducted to place the observable states onto a preference continuum, with an anchor of 0 for death and 1.0 for completely well. Studies have shown that the weights are highly stable over a one-year period and that they are consistent across diverse groups of raters. Finally, one must consider the duration of stay in various health states. Having a cough or a headache for one day is not the same as having the problem for one year. A health measure must take these durations into consideration. Using this information, one can describe health-related quality of life in terms similar to years of life. For example, one year in a state assigned the weight of 0.5 is equivalent to 0.5 of a quality-adjusted life-year.

3.3 Cost-effectiveness Evaluations

Utility-based measures of quality of life are often used to evaluate the cost-effectiveness of health-care programs (Kaplan et al. 1997). For example, one study using the QWB showed a new medication for patients with arthritis produced an average of 0.023 QALY per year, whereas a new medication for AIDS produced nearly 0.46 of these units per year. However, the benefit of the arthritis medication may last as long as 20 years, ultimately producing $0.023 \times 20 \text{ years} = 0.46$ year. The AIDS treatment produced a benefit for only one year, so its total effect was $0.46 \times 1 \text{ year} = 0.46$ year. In other words, the general system allows the full potential benefits of these two very different treatments to be compared (Kaplan et al. 1995).

QALYs are generic measures of life expectancy with adjustments for quality of life (Gold 1996). QALYs consider both benefits and side-effects of programs in terms of the common outcome units. Although QALYs are typically assessed for patients, they can also be measured for others, including care givers who are placed at risk because they experience stressful life events. The Institute of Medicine recommended that population health metrics be used to evaluate public programs and to assist the decision making process (Field and Gold 1998).

In addition to health benefits, programs also have costs. Resources are limited and good policy requires the allocation of resources to maximize life expectancy and health-related quality of life. Thus, in addition to measuring health outcomes, costs must also be considered. Methodologies for estimating costs have now become standardized (Gold 1996). From an administrative perspective, cost estimates include all costs of treatment and costs associated with caring for any side

effects of treatment. From a social perspective, costs are broader and may include the costs of family members staying off work to provide care. Comparing programs for a given population with a given medical condition, cost-effectiveness is measured as the change in costs of care for the program compared with the existing therapy or program, relative to the change in health measured in a standardized unit such as the QALY. The difference in costs over the difference in effectiveness is the *incremental cost-effectiveness* and is usually expressed as the cost/QALY. Since the objective of all programs is to produce QALYs, the cost/QALY ratio can be used to show the relative efficiency of different programs (Gold 1996).

Some traditional interventions, such as bypass surgery for one-vessel coronary heart disease, may cost as much as \$700,000 to produce a QALY. Screening programs, such as mammography, may also require many resources to produce a QALY. On the other hand, public health programs such as pneumonia vaccines for the elderly or laws requiring children to be in infant seats and adults to be in seat belts in motor-cars may produce a QALY at a very low cost. It might be argued that programs to the left of the pay line should be funded, but those with cost/QALY ratios to the right of the line should be examined more carefully.

4. Disease-specific Measures

This chapter has focused on generic measures of health outcome. These measures can usually be applied to any age, gender, or ethnic group. In most cases age and gender-related norms are available. However, there are a significantly larger number of situation-specific measures. Some of these measures are appropriate for a particular age, gender or ethnic group. For example, several measures are designed specifically to assess health outcomes in children. Most often, the measures are disease-specific, meaning that they are designed to assess outcomes of a particular health problem. Validated measures of health outcome are available for illnesses such as arthritis, diabetes, heart disease, kidney failure, and virtually every other major health condition. Two examples are the Arthritis Impact Measurement Scale (Meenan 1982) and the UCSD Shortness of Breath Questionnaire used in studies of patients with emphysema (Eakin et al. 1998).

The Arthritis Impact Measurement Scales (AIMS) is a health index designed at the Multipurpose Arthritis Center at Boston University. It is intended to measure physical health and social well-being for patients with rheumatoid arthritis (Meenan 1982). The resultant scale includes 67 items, with questions about functioning, health perceptions, morbidity, and demographics. The AIMS contains scales for mobility,

physical activity, social activity, activities of daily living, depression and anxiety and arthritis-related symptoms. In effect, it is an adaptation of an early version of the QWB (Eakin et al. 1998), with a series of items designed to tap more specifically the effect of arthritis upon functioning and the quality of life. Factor analysis of the AIMS has produced three subscales: physical function, psychological function, and pain. Most current applications of the AIMS use composite scores for these three areas.

The University of California, San Diego, Shortness of Breath Questionnaire (SOBQ) includes 25 items that evaluate self-reported shortness of breath during the performance of various activities of daily living. Evaluations of the measure show it highly correlated with other quality of life measures such as the Quality of Well-Being Scale and the Center for Epidemiologic Studies Depression Scale. The measure has high internal consistency ($\alpha = 0.96$) and is significantly correlated with performance measures such as the distance that can be walked in six minutes (Eakin et al. 1998).

5. Summary

The assessment of patient-reported experience has become a standard component of health outcome evaluation. There are several important distinctions between commonly-used approaches. Generic measures are used to evaluate the health outcome for any illness or disease of state. Generic measures can typically be classified as derived from psychometric or decision theory. Psychometric approaches include the Sickness Impact Profile, the SF-36, and the Nottingham Health Profile. Decision theory approaches are used to estimate outcomes in terms of quality-adjusted life years (QALYs). Methods used for this purpose include the EQ-5D, the HUI, and the QWB. Disease-specific measures are available for a wide variety of health conditions. Although disease-specific measures may be more sensitive for outcomes of a particular condition, they cannot be used for cross-illness comparisons or for cost-effectiveness analysis. The assessment of health outcome is a rapidly-developing field and we anticipate major new developments over the next decade.

See also: Chronic Illness: Quality of Life; Quality of Life: Assessment in Health Settings; Well-being (Subjective), Psychology of

Bibliography

- Bergner M, Bobbitt R A, Carter W B, Gilson B S 1981 The sickness impact profile: Development and final revision of a health status measure. *Medical Care* **19**(8): 787-805

- Brown M, Gordon W A, Haddad L 2000 Models for predicting subjective quality of life in individuals with traumatic brain injury. *Brain Injury* **14**(1): 5–19
- Eakin E G, Resnikoff P M, Prewitt L M, Ries A L, Kaplan R M 1998 Validation of a new dyspnea measure: The UCSD Shortness of Breath Questionnaire. University of California, San Diego. *Chest* **113**(3): 619–24
- Feeny D, Furlong W, Mulhern R K, Barr R D, Hudson M 1999 A framework for assessing health-related quality of life among children with cancer. *International Journal of Cancer Supplement* **12**(5): 2–9
- Field M J, Gold M R 1998 *Summarizing Population Health*. Institute of Medicine. National Academy Press, Washington, DC
- Gold M R 1996 *Cost-effectiveness in Health and Medicine*. Oxford University Press, New York
- Gudex C, Dolan P, Kind P, Williams A 1996 Health state valuations from the general public using the visual analogue scale. *Quality of Life Research* **5**(6): 521–31
- Hurst N P, Kind P, Ruta D, Hunter M, Stubbings A 1997 Measuring health-related quality of life in rheumatoid arthritis: Validity, responsiveness and reliability of EuroQol EQ-5D. *British Journal of Rheumatology* **36**(5): 551–9
- Kaplan R M, Anderson J P, Patterson T L, McCutchan J A, Weinrich J D, Heaton R K, Atkinson J H, Thal L, Chandler J, Grant I 1995 Validity of the Quality of Well-Being Scale for persons with human immunodeficiency virus infection. HNRC Group HIV Neurobehavioral Research Center. *Psychosomatic Medicine* **57**(2): 138–47
- Kaplan R M, Erickson P 2000 Gender differences in quality-adjusted survival using a Health-Utilities Index. *American Journal of Preventive Medicine* **18**(1): 77–82
- Kaplan R M, Sieber W J, Ganiats T G 1997 The Quality of Well-being Scale: Comparison of the interviewer-administered version with a self-administered questionnaire. *Psychology & Health* **12**(6)
- Keller S D, Ware J E Jr, Hatoum H T, Kong S X 1999 The SF-36 Arthritis-Specific Health Index ASHI: II Tests of validity in four clinical trials. *Medical Care* **5**(Suppl MS): 51–60
- Kind P 1997 The performance characteristics of EQ-5D, a measure of health related quality of life for use in technology assessment [abstract]. *Annual Meeting of International Society of Technology Assessment in Health Care* **13**(5): 81
- Koloski N A, Talley N J, Boyce P M 2000 The impact of functional gastrointestinal disorders on quality of life. *American Journal of Gastroenterology* **95**(1): 67–71
- Kosinski M, Keller S D, Ware J E Jr, Hatoum H T, Kong S X 1999 The SF-36 Health Survey as a generic outcome measure in clinical trials of patients with osteoarthritis and rheumatoid arthritis: Relative validity of scales in relation to clinical measures of arthritis severity. *Medical Care* **37**(5 Suppl MS): 23–39
- Lowe D, O'Grady J G, McEwen J, Williams R 1990 Quality of life following liver transplantation: A preliminary report. *Journal of the Royal College of Physicians of London* **24**(1): 43–6
- McDowell I, Newell C 1996 *Measuring Health: A Guide to Rating Scales and Questionnaires*, 2nd edn. Oxford University Press, New York
- Meenan R F 1982 The AIMS approach to health status measurement: Conceptual background and measurement properties. *Journal of Rheumatology* **9**(5): 785–8
- Parsons T 1951 *The Social System*. Free Press, Glencoe, IL
- Patrick D L, Erickson P 1993 *Health Status and Health Policy: Quality of Life in Health Care Evaluation and Resource Allocation*. Oxford University Press, Boston, MA
- Russell L B 1986 *Is Prevention Better Than Cure?* Brookings Institution, Washington, DC
- Scott-Lennox J A, Wu A W, Boyer J G, Ware J E Jr 1999 Reliability and validity of French, German, Italian, Dutch, and UK English translations of the Medical Outcomes Study HIV Health Survey. *Medical Care* **37**(9): 908–25
- Stewart A L, Ware J E 1992 *Measuring Functioning and Well-being: The Medical Outcomes Study Approach*. Duke University Press, Durham
- Torrance G W, Feeny D H, Furlong W J, Barr R D, Zhang Y, Wang Q 1996 Multiattribute utility function for a comprehensive health status classification system. Health utilities index mark 2. *Medical Care* **34**(7): 702–22
- Walker S R, Rosser R 1993 *Quality of Life Assessment: Key Issues in the 1990s*. Kluwer Academic Publishers, Dordrecht
- Ware J E Jr, Gandek B 1998 Overview of the SF-36 Health Survey and the International Quality of Life Assessment IQOLA Project. *Journal of Clinical Epidemiology* **51**(11): 903–12

R. M. Kaplan

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Health Policy

In Western industrialized nations, the term health policy has come to mean government policies that subsidize and regulate the *economic* relations between patients, insurers, and health care providers—doctors, hospitals, pharmacies, pharmaceutical companies. Within this general framework—an economic model of medicine—the role of government in health care varies considerably from country to country and has been the subject of considerable and continuing political contestation. In order to understand these differences, one must return to the political circumstances that led to the enactment—or rejection—of specific health programs.

1. Types of Health Policy

The role of government in health care may be classified according to methods of financing health care provision. Governments may subsidize medical treatment by providing for tax exemptions for medical expenditures. They may provide government subsidies to private insurers. They may provide health insurance coverage by creating government insurance programs. Or they may provide health services through government-employed doctors and government-owned hospitals. Because each type of financing has implications for the role of government in regulating the medical market, government financial provisions in the health area can be ranked from less to more

invasive: tax expenditures are considered the least interventionist form of health policy; government health services are considered the most interventionist.

1.1 Government Subsidies

Tax deductions for medical expenditures were generally introduced after the Second World War, and were often the result of political conflicts preventing a more ambitious role for government in health, such as the failure of the Truman national health insurance proposal of 1948 in the United States. A policy of tax expenditure impacts the medical market by reducing the costs of health care to the consumer. This financial incentive allows governments to promote health policy goals, for example, by deciding which types of medical treatment will be deductible, or which types of insurance benefits qualify as tax exempt. Tax deductibility has been an important element in the spread of health insurance coverage as part of collective bargaining agreements since the 1940s, although the current trend is towards reduction rather than expansion of insurance benefits provided by employers.

A second form of government financing for health care is the payment of *subsidies* to insurance carriers. This policy dates from the mid-nineteenth century and was a reaction to the growth of voluntary mutual aid societies (also known as friendly societies or sickness funds) that provided their members with death and sickness benefits (cash payments to cover the costs of burial and income loss during illness, respectively). Typically, governments provided the mutual societies with a fixed subsidy per member. In return, the societies were required to follow specific financial practices—for example to build up larger financial reserves—and to provide medical benefits in addition to cash payments. In addition, political authorities used financial resources as a means of political control, monitoring the activities of the societies, and sometimes intervening in the selection of the members of the societies' governing boards, as was the case in nineteenth century France, Italy, and Belgium.

1.2 National Health Insurance

Under national health insurance programs, governments provide *compulsory* health insurance to all or parts of the population. Historically, such programs were first introduced to provide insurance to industrial workers at the end of the nineteenth century, with Bismarck's social insurance laws of 1883 as the first example (see *Social Security; Welfare State, History of Social Insurance: Legal Aspects*). Over the course of the twentieth century, compulsory health insurance programs have been expanded to cover ever greater portions of national populations—hence the term 'national.' National health insurance programs are in

place in most of Europe, Latin America, Canada, Australia, and Japan. When the entire population is covered as a right of citizenship, these programs are said to be universal. Nevertheless, while the term 'national health insurance' connotes the compulsory provision of health insurance to broad sectors of the population, there is no generally accepted definition for the portion of the population to be covered or for the exact benefits to be provided. These are matters of political negotiation and political choice.

National health insurance programs provide governments with greater incentives and instruments for regulating the medical market than programs of government subsidies. Government decisions determine who is to be publicly insured—often defined by income or age limits, or by specifying occupational groups, such as farmers, salaried employees, civil servants, or the unemployed, that are to be compulsorily insured. These decisions, in turn, set limits to the market for private insurance. National health insurance programs may be directly administered by government agencies, or by independent insurance carriers, often the voluntary mutual aid societies or nonprofit group plans that existed prior to the government program. In either case, however, governments regulate benefits, conditions of membership, portability of benefits, and contribution rates. Rather than paying a premium for their insurance coverage, citizens are said to pay a contribution, one which is often collected in the form of a payroll tax, as part of the nation's social insurance or social security system. Employers are generally required to pay for half or more of the insurance contribution for their employees. Consequently, many public insurance programs provide for some representation of these contributors on the boards of the insurance carriers, a practice that has been termed 'self-administration'.

The regulation engendered by national health insurance is not limited to citizens and insurers. Political decisions—often the result of negotiations between representatives of providers and insurance authorities—determine the specific treatments and medications that will be covered by a particular national health insurance program, as well as the price that may be charged. Governments may simply set maximum fees to be charged to national health insurance patients, or they may negotiate fees with representatives of medical or hospital associations. These lists of negotiated fees are known as fee schedules. Hospitals or doctors may be required to accept the national health insurance fee as payment in full, and thus prohibited from adding surcharges (balance-billing or extra payments).

Because government price controls may be circumvented by increasing the volume of services, newer methods of cost containment have been developed. Case-based payment or diagnostic-related groups remunerate hospitals a fixed sum per patient with a particular diagnosis rather than reimbursing the full

costs of treatment for a specific individual. In this way, providers are given an incentive to reduce the volume (and hence the costs) of treatment. The global budget bases hospital payments on a larger entity. Either a department or even the entire hospital will be given a yearly (or quarterly or semi-annual) budget, within which the hospital managers may allocate funds at their discretion. In the area of pharmaceuticals, common cost-containment measures include the development of positive (allowed) and negative (prohibited) lists of prescription drugs, as well as efforts to negotiate directly with manufacturers to obtain discounts on pharmaceutical and other medical products. In order to reduce the volume of prescriptions, some national health insurance authorities have introduced economic monitoring of physicians (reviews whereby physician behavior is compared to statistical averages) and prescription budgets for physicians. Further, many national health insurance programs rely on cost-sharing on the part of patients to reduce the demand for health services.

1.3 National Health Services

In a national health service, governments provide health services directly. Citizens are guaranteed a range of health care benefits, which are generally provided at no or at low cost at the time of treatment. Medical benefits are typically financed by tax revenues (although there may be a health insurance component as well) and provided under government auspices. Just what this 'public' provision is to entail, however, is highly controversial. In the British National Health Service, founded in 1948, and in Sweden, a country that can be considered to have a *de facto* national health service, the overwhelming majority of hospitals are government owned, and hospital doctors are paid a full-time salary. In Southern Europe (Italy, Portugal, and Spain), the more recently created national health services have incorporated privately-owned hospitals, such as those owned by churches and social security institutions, into the public service on a contractual basis. Within national health service hospitals, patients may or may not be allowed to pay additional monies for treatment in a 'private' bed or private room. Similarly, hospital doctors (especially chiefs of staff) may or may not be allowed to treat 'private' patients for additional fees.

The outpatient or ambulatory sector has traditionally been a battleground for the demarcation of the public-private boundary. Local health clinics or hospital outpatient departments compete with private office practice by providing health treatments at virtually no charge through doctors that are paid on a salaried basis. Even in private offices, doctors in national health services are often under contract to the government, and paid according to fee schedules or on

a capitation basis (a set fee per patient per year). Efforts to pay private office practitioners a yearly salary for their national health service work have generally failed in market economies. In state socialist countries, on the other hand (China, Cuba, the former German Democratic Republic, and the former Soviet Union, for example), health care was long provided through government hospitals and outpatient clinics (polyclinics or Ambulatorien) that employed doctors as salaried civil servants.

In comparison to systems of government subsidies and national health insurance—both of which increase the demand for health services—national health services provide governments with relatively effective tools of cost-containment and health planning. Particularly in a system such as the British, where health expenditures are financed by the national government and the national health budget constitutes a hard constraint on health spending, less regulation of providers has been necessary to hold expenditures in check. Moreover, centralized public authority for investment decisions has facilitated national planning for medical technology and the regionalization of the hospital system. These policies limit and guide the distribution of expensive technologies, such as nuclear-magnetic-resonance-imaging devices, and create a division of labor among hospitals, with more highly-specialized personnel and equipment concentrated at regional centers (see *Health Care Technology*). As expenditures are curtailed, however, waiting lists and complaints about the quality of services may arise.

1.4 Health Policy Issues

From this description, it is clear that the relationship between state and market has been a central issue of health policy. But, while this division dovetails neatly with political ideologies—and hence with political discourse—it is too simplistic for describing the complex mixture of private and public provision and financing of health care in place today. Even in relatively 'market-oriented' systems, government financing and regulation of health care is extensive, including funds for hospital construction and research, as well as a myriad of regulations governing health insurance, hospitals, and medical equipment. Furthermore, health care decision-makers—whether public or private—are turning increasingly to quite similar policy instruments. Private insurance companies, hospital corporations, and planners interested in regulating the private market promote managed care and managed competition that involves the same budgetary controls, fee schedules, and salaried employment that have been developed to contain costs under national health insurance and in national health services. On the other hand, government health systems now strive to use market incentives and increasing

competition to improve the quality, responsiveness, and efficiency of public systems in programs of public competition and internal markets. Consequently, terms like 'socialized medicine' and 'medical rationing' are better considered as political slogans than as accurate descriptions of health policy. The public component of medical care is very large in all countries—and has been so historically—and efforts to influence the distribution of health care services are prevalent throughout both the public and the private sectors. Instead, the politically more neutral monopsony (or existence of a single buyer) may better capture differences in health systems. To the extent that purchasing power becomes concentrated, the buyers of health care can exert more leverage over the providers. This tendency is at the root of conflicts between providers, such as doctors, and the organized buyers of medical services—private insurance companies, voluntary mutual aid societies, hospitals, and governments.

In sum, health policy entails government's pursuit of three goals through both public and private means: the provision of citizens with access to health care; the assurance of the quality of health care; and the control of the costs of health care. Because it is virtually impossible to meet these three goals simultaneously, health policy will continue to be contested. (See also: *Health Care Systems: Comparative; Health Care Delivery Services; Health Care Markets: Theory and Practice; Health Economics; Managed Care; National Health Care and Insurance Systems; Health Care, Rationing of*.)

2. Determinants of Health Policy

2.1 Public Health Provision

Conflicts over the role of the state in medicine go back as far as the Reformation. At that time, responsibility for health care was divided among an assortment of authorities with overlapping and contradictory jurisdictions. Medical services were delivered by a diversity of practitioners, with widely varying degrees of training, and through a range of institutions, including the Catholic church. In their quest for external and internal sovereignty (see also *Absolutism, History of and Policy History: Origins*), monarchs such as Henry VIII of England or Gustav Wasa of Sweden appropriated hospices along with other Church lands, and began to offer health services under government auspices. Particularly in countries with an absolutist tradition, the rise of the nation state (see *Nations and Nation-states in History; State Formation; State and Society; State, Sociology of the; State, History of*) was associated with an expansion of military medicine, and the establishment of public health officers as well as government bureaucracies for tracking contagious

diseases. In Russia, the state established public health clinics called *Zemstvos*; in many other countries, public health centers were blocked by medical opposition. By the end of the nineteenth century, governments throughout the West had acknowledged some role for the state in the provision of health services, but its scope varied widely.

2.2 The Medical Profession

Political authorities also interceded in the relationships between churches, universities, and medical corporations by giving government approval to specific institutions of medical training and licensing, and, in some cases, by outlawing alternative practitioners. The Royal College of Physicians of London, for example, was granted a charter in 1518 to license physicians. This charter did not interfere, however, with the rights of Cambridge and Oxford universities or ecclesiastical authorities. As in many European countries, three strata of practitioners co-existed: the academically-trained physicians; the more practically-trained surgeons—who had originally been associated with barbers, and whose development was linked to military medicine, like the *Feldshers*, *Wundärzte*, and *Fältskäraren* on the Continent; and the apothecaries—who, expanding their trade to prescribing medicines, had split off from the grocer's guild in 1617. The medical profession was not unified until the Medical Act of 1858, which provided for a General Council of Medical Education and Registration, composed of representatives of the medical corporations and the universities, that was to supervise medical education and register holders of medical licenses. All previous licensing bodies could continue to grant degrees, but there were no restrictions on who could practice medicine other than penalties for improper use of a medical title.

The process by which specialized routes of training and barriers to entry were established has been called professionalization (see *Medical Sociology; Professions, Sociology of; Professionalization/Professions in History*). In some countries, such as Spain and Sweden, professional unification came earlier than in Britain, and professional monopoly was 'tighter,' as unlicensed practice was actually outlawed. In many cases, liberal ideology—with its stress on free markets, individual freedom, and a minimal state—was a barrier to professionalization. In the United States and Switzerland, liberal and democratic ideals precluded medical licensing, limits on medical school admissions, and even prohibitions on medical quackery until well into the twentieth century. In Prussia, the liberal Berlin Medical Society opposed the introduction of mandatory licensing in 1869, which led to a system of state certification for doctors with no penalties for unlicensed practice when Germany was

unified. In France, the revolution (which firmly placed hospitals in the hands of the state) led to new laws that annulled the *ancien régime* requirement of a university degree for the practice of medicine, outlawed medical associations and abolished the university medical faculties, as well as a host of special rights and privileges granted to various subgroups of the profession. Only at the end of the nineteenth century was medical monopoly firmly re-established in France and placed under the jurisdiction of the state universities, without a role for professional bodies. Perhaps for this reason, professional organization in France has been weak, although medical monopoly is strong (see also *Western European Studies: Health*).

2.3 Medical Associations

One can distinguish between several different types of medical organization. Medical corporations, like the British Royal Colleges or the German *Bundesärztekammer*, are professional bodies with legal jurisdiction over medical licensing and professional ethics. Medical societies are associations formed for the purposes of disseminating medical knowledge, and have often founded medical journals. Despite their efforts to maintain some distance from economic and political issues, they have also promoted the interests of their members. Medical associations and medical unions are most directly involved with economic negotiations and lobby efforts, having been founded explicitly as interest associations. In addition, some nations require doctors participating in national health insurance schemes to belong to associations that represent doctors in bargaining with health authorities, such as the *Kassenärztliche Bundesvereinigung* in Germany. Separate organizations often exist for hospital doctors and for various medical specialties. Medical licensing is now only rarely left to medical corporations, but specialty licensing is often left to specialty societies, which may explain their high membership rates and influence.

Professionalization and the development of medical organizations both spurred and was spurred by the expansion of the market for medical care. By defining medicine as a distinct body of knowledge—in order to gain public legitimacy for professional privilege, on the view of some interpretations—medical professions promoted the demand for their services. As governments acknowledged these claims to expertise, they moved to provide medical services to the public, turning increasingly to programs of medical insurance to supplement direct public provision. To be sure, these policies of state medical care were not motivated purely by health policy concerns—the fear of working-class unrest and the need for healthy soldiers were just as important. Indeed, the voluntary mutual societies themselves were not formed solely to provide sickness

and death benefits, but instead served as a way to form associations at a time when many working-class organizations were prohibited by various anticombination acts. In any case, the provision of medical benefits through state-subsidized mutual associations and compulsory health insurance plans further stimulated the formation of medical associations, as doctors sought to prevent the development of disadvantageous working conditions.

2.4 Professional Autonomy

The increasing prevalence of collective payment for medical services stimulated normative and theoretical discussions about the relationship between public service and professional freedom (see for example the debate between Harold Laski and T. H. Marshall described in Marshall 1964 [1939]). Freidson's seminal work defines the key feature of professional freedom as 'technical autonomy.' Occupations that have been able to convince the public that their practitioners are privy to indispensable knowledge gain a privileged position, defined by the exclusive competence to judge professional work. Whereas Freidson specifically states that economic autonomy ('control over the socioeconomic terms of work,' 1970, p. 25) has nothing to do with professional freedom, this has been a central issue in health policy. Medical associations across Western Europe and North America fought for economic autonomy, arguing that economic freedom such as the right to set fees and receive private patients was a precondition of therapeutic freedom (the physician's ability to choose freely among medical therapies) and of the right of patients to a 'free choice of doctor.' As private medical practice has become subject to burdensome government regulation and corporate management controls, however, some argue that physicians are better off in national health systems, where more generalized budgetary controls obviate the need for such detailed micromanagement, preserving a greater degree of clinical autonomy (the ability of physicians to make individual medical decisions).

2.5 The Politics of Health

National health policies are thus the result of centuries of political conflict over the proper role of government in health. Both governmental decisions and those made in private negotiations have influenced the complex set of institutions and regulatory framework that governs contemporary health care financing and provision. In order to make sense of these patterns, scholars have turned to some of the same sociological and political theories that have been applied to the politics of the welfare state (see *Social Welfare Policy*:

Comparisons; Welfare State, History of). The development of different national health policies has been understood in terms of differences in: leading political ideologies and political cultures; social structure and class relations; state traditions and state structures; left and right political representation; national political institutions. A particular focus of the study of health politics has been the issue of professional dominance, that is, whether the medical profession as the sole experts qualified to deliver medical services can parlay their professional expertise into sufficient political power to dominate health policy decisions.

2.6 Future Challenges

Because of its historical development, health policy has been focused very much on the role of nation-states in facilitating and regulating medical services that are provided by national medical professions. This traditional definition now faces two central challenges: the internationalization of the economy, and the revolution in molecular biology. Internationalization poses a challenge to national institutions and programs of health care. Not only the free movement of persons and services across borders, but the tremendous international differences in access to medical treatment, highlighted by crises such as the Aids pandemic, cannot be easily addressed through national health policies. Nor is the Western-industrial approach, which is focused on individual medical services, necessarily appropriate for the problems of the developing world.

Even more profound, perhaps, are the implications of the cracking of the genetic code. Discussions about whether the goal of health policy should be to provide access to medical services or whether investments in public health—hygiene, vaccines, and nutrition—might not have more impact on the health states of populations have been perennial, to be sure. Nevertheless, the discovery of the genetic basis for disease poses a new threat both to the position of doctors as the holders of medical expertise and the role of curative medicine. Furthermore, the potential for genetic engineering, and even the development of more effective pharmaceuticals through recombinant DNA methods, entails tremendous ethical and cost problems.

See also: Health Care Delivery Services; Health Care: Legal Issues; Health Care Organizations: For-profit and Nonprofit; Health Economics; Health in Developing Countries: Cultural Concerns; Health Insurance: Economic and Risk Aspects; Health Outcomes, Assessment of; Medical Sociology; National Health Care and Insurance Systems; Professions, Sociology of; Welfare State

Bibliography

- Abelsmith B 1985 Who is the odd man out? The experience of Western Europe in containing the costs of health care. *Milbank Memorial Quarterly/Health and Society* 63: 1–17
- Anderson O W 1972 *Health Care: Can There be Equity? The United States, Sweden and England*. John Wiley, New York
- Björkman J W, Altenstetter C 1998 *Health Policy*, The International Library of Comparative Public Policy 7. Edward Elgar, Cheltenham, UK
- Freidson E 1970 *Profession of Medicine: A Study of the Sociology of Applied Knowledge*. Dodd, Mead & Co., New York
- Glaser W A 1991 *Health Insurance in Practice: International Variations in Financing, Benefits and Problems*. Jossey-Bass, San Francisco
- Hollingsworth J R, Hage J, Hanneman R A 1990 *State Intervention in Medical Care: Consequences for Britain, France, Sweden, and the United States, 1890–1970*. Cornell University Press, Ithaca, NY
- Klein R 1989 *The Politics of the National Health Service*, 2nd edn. Longman, London
- McLachlan G, Maynard A (eds.) 1982 *The Public/Private Mix for Health: The Relevance and Effects of Change*. The Nuffield Provincial Hospitals Trust, London
- Marmor T R 1994 *Understanding Health Care Reform*. Yale University Press, New Haven, CT
- Marshall T H 1964 [1939] The recent history of professionalism in relation to social structure and social policy. In: Marshall T H (ed.) *Class, Citizenship and Social Development: Essays*. Doubleday, Garden City, NY
- Maxwell R J 1981 *Health and Wealth: An International Study of Health Care Spending*. Lexington Books, Lexington, MA
- Organisation for Economic Co-operation and Development (OECD) 1994 *The Reform of Health Care Systems: A Review of Seventeen OECD Countries*. OECD, Paris
- Powell F D, Wessen A F 1999 *Health Care Systems in Transition: An International Perspective*. Sage, Thousand Oaks, CA
- Raffel M W (ed.) 1997 *Health Care and Reform in Industrialized Countries*. Pennsylvania State University Press, University Park, PA
- Rimlinger G V 1971 *Welfare Policy and Industrialization in Europe, America and Russia*. John Wiley, New York
- Rodwin V 1984 *The Health Planning Predicament: France, Québec, England and the United States*. University of California Press, Berkeley, CA
- Schieber G J, Poullier J-P 1991 International health spending: Issues and trends. *Health Affairs* 10(1): 106–16
- Schneider M, Biene-Dietrich P, Gabanyi M, Huber M, Köse A, Scholtes L, Sommer J H 1993 *Gesundheitssysteme im internationalen Vergleich*, Ausgabe 1992. Beratungsgesellschaft für angewandte Systemforschung (BASYS), Augsburg
- Starr P 1982 *The Social Transformation of American Medicine*. Basic Books, New York
- Stone D A 1980 *The Limits of Professional Power*. University of Chicago Press, Chicago
- White J 1995 *Competing Solutions: American Health Care Proposals and International Experience*. The Brookings Institution Press, Washington, DC
- Wilkerson J D, Devers K J, Given R S 1997 *Competitive Managed Care: The Emerging Health Care System*. Jossey-Bass, San Francisco

E. M. Immergut

Copyright © 2001 Elsevier Science Ltd. 6591
All rights reserved.

Health Professionals, Allied

Health care services experienced remarkable growth in the latter half of the twentieth century in industrialized countries. Advances in medical science and technology, the expansion of health insurance, the aging of the population, increased public expectations and disposable income, and expanding social definitions of the boundaries of medical care all contributed to increased utilization of health services and changing patterns of care. These changes fostered the development of an expanded division of medical labor in which allied health professionals emerged to complement and compete with physicians in the provision of a broader array of health services. Globalization of health care has fostered renewed interest in strengthening the roles of allied health providers in developing countries and in the post-Soviet states in order to increase access to primary medical care and improve population health.

1. The Changing Context of Health Care

Health and medical services have undergone dramatic change since the mid-twentieth century. Medical care has become more specialized while also expanding in scope. At the beginning of the twentieth century, physicians accounted for about one in three healthcare workers; by the 1980s the comparable ratio was about 1 in 16 (Mechanic 1983). Health care occupations are highly sex-segregated with most nonphysician occupations dominated by women. Indeed, in many parts of the world physicians are also predominantly women reflecting societal sex role links between women and caring/health functions.

1.1 Changing Disease Patterns

Advances in medical science including antibiotics and vaccines as well as improved standards of living contributed to shifting the burden of illness from acute, infectious illnesses to chronic conditions. While advances in modern medicine have produced important treatments for chronic illnesses, most have resulted in 'halfway' technologies that extend life but do not cure the underlying disease (Thomas 1977). Examples include long-term renal dialysis for the treatment of kidney failure and, more recently, anti-retroviral drugs that extend the lives of people with acquired immunodeficiency syndrome (AIDS). Chronic conditions are accompanied by a myriad of functional limitations and social support needs that fall outside the conventional definition of medical practice. Nurses, social workers, occupational therapists, physical therapists, and other nonphysician health

professionals have demonstrated interest and expertise in helping the chronically ill and frail elderly manage the sequella of chronic conditions and medical treatments, thus creating new roles for themselves in the medical division of labor.

1.2 Medicalization of Problems of Everyday Living

The successes of modern medicine in treating life-threatening illnesses have resulted over time in the redefinition of a range of health-related issues and problems of everyday life as medical concerns. The boundaries of medicine have expanded to encompass health promotion, diet and exercise, child rearing, school performance, occupational health, and a wide array of personal and social problems. Medicalization of problems of everyday living lend legitimacy to social interventions (such as behavior control of disruptive school children), allow those seeking help to adopt a socially sanctioned sick role (Parsons 1951), and provide access to health insurance funds for financing treatment. Physicians are not well suited by interests or training to providing quasimedical social services, thus opening up new roles for allied health professions including social workers, psychologists, and chiropractors, as well as a myriad of new occupations such as personal exercise trainers and weight loss consultants.

Indeed, physicians have continued to pursue more specialized medical training and practice as help-seeking patterns in the general population have redefined the boundaries of medical care to encompass a much broader array of health concerns. Thus, in addition to the expanded opportunities afforded the allied health professions, alternative and complementary health practitioners such as naturopaths, acupuncturists, and herbalists have grown in number and availability. In 1997, American consumers paid an estimated \$27 billion for unconventional therapies, which is roughly comparable to out-of-pocket expenditures for physician services (Eisenberg et al. 1998). Increasingly, conventional physician-dominated settings, including prestigious teaching hospitals, are establishing alternative and complementary health clinics employing nonphysician providers.

1.3 Organizational Changes Conducive to Expansion of Nonphysician Professions

Professions routinely aspire for full control over well-documented jurisdictions in which workplace authority is secured formally through legislation or courts of law and validated by public opinion (Abbott 1988). Physicians have successfully defended their dominance over the medical division of labor through legislative

and legal actions, controlling referral patterns, and by limiting direct access to health services payments (Sage and Aiken 1997). Yet, nonphysician health professions are constantly seeking to define themselves, grow, survive, and contest with medicine within an organizational environment that is continually subject to change due to exogenous social and technological trends.

The organizational context of medical practice changed substantially in the latter third of the twentieth century. The corporatization and consolidation of health services resulted in loss of autonomy for physicians (Starr 1982), but new opportunities and a more favorable environment for the allied health professions. Managerial priorities on cost containment gave allied health professions their first opportunity to compete with physicians on the basis of price and client satisfaction, and they have been successful on both.

1.4 Rapid Growth of Nonphysician Clinicians

The changing nature of health care and its organization has resulted in the rapid growth of allied health professions even in a context of a rapidly increasing physician supply. An analysis of growth of nonphysician clinicians in the USA revealed that the aggregate number of nonphysician clinicians in 10 disciplines (nurse practitioners, physician assistants, nurse-midwives, chiropractors, acupuncturists, naturopaths, optometrists, podiatrists, nurse anesthetists, and clinical nurse specialists) doubled between 1992 and 1997, and was expected to increase by two-thirds more by 2005 (Cooper et al. 1998). These 10 occupations were selected for study because they most overlap the scope of medical and surgical services provided by physicians. The growth of allied health professions is not unique to the USA, and evidence of expansion of nonphysician clinicians can be found in many industrialized countries and also in the developing world as health care becomes an increasingly global enterprise.

1.5 Globalization of Health Services

Global trade and business interests have increasingly impacted upon health services and health care personnel through cross-national commercial ventures, trade agreements affecting the movement of workers across national boundaries, and international aid activities. The North American Free Trade Agreement (NAFTA), linking Mexico, the USA, and Canada, and the European Union (EU) created the impetus to begin reconciling the training and practice of health professions across countries (Freeman et al. 1995). Health services is a major vehicle for transforming economic markets in the now independent states of the former Soviet Union, and many Western concepts of organization and financing of health services, and

education of medical personnel are being introduced (Mossialos and McKee 1998/99). The growing threat of new emerging infectious diseases in parts of the developing world has attracted international attention and focus on strengthening public health and primary care in poor countries (Institute for the Future 2000). Since the health workforce is the infrastructure upon which health services initiatives are built, global initiatives inevitably impact on the education and composition of the medical workforce.

2. Nursing

Nursing is the largest of the health professions. Nursing historically has been closely tied to the profession of medicine, and the boundaries between the two have blurred over time in interesting ways. Nurses now participate in activities that in the past were the sole domain of physicians, such as physical examinations, diagnosis, treating illness, and prescribing drugs. The obverse is also true. Physicians have claimed jurisdiction over some types of services provided routinely in the past by nurses, as in the case of attending at normal births (Fagin 1990), and provision of anesthesia for surgery which only since the 1980s has become a physician specialty (Bankert 1989).

Expanded roles for nurses is a global phenomenon. The USA has led in developing expanded roles for nurses in care exploiting high technology (Fairman and Lynaugh 1998) and in independent practice roles in ambulatory care with the development of advanced practice nurses (Mezey and McGivern 1999). Other countries are now following similar patterns. British nurses have been granted wider authority by the National Health Service to prescribe drugs (Norman 2000). Canada is training nurse practitioners based on the US model of advanced practice nursing. Role expansion for nurses is occurring in a context of sufficient or excess supply of physicians, suggesting that nurses are not substituting for physicians as much as they are responding to a broader definition of health and medical care and market forces that provide new opportunities for role expansion for nonphysicians.

2.1 Historical Perspectives

Modern nursing dates back to mid-nineteenth century England when hospitals were disorganized, unsanitary, staffed by unskilled and undisciplined workers, and were generally held in low esteem. Only the destitute sought help in hospitals, except in times of war. Florence Nightingale, a wealthy and educated English woman, accepted the challenge of bringing order to military hospitals during the Crimean War and in doing so began a social experiment that would change the nature of hospitals around the world. Nightingale identified trained nurses as the vehicle for transforming hospitals but she recognized that nurses capable of skilled work and responsible ward man-

agement would have to be recruited from the better-educated classes. To legitimize the work of nursing to more highly educated women, she drew on the well-respected traditions of religious education and military-style authority. Young nurses motivated by a sense of dedication and service to mankind worked long hours for few economic rewards under strict discipline with total obedience expected (Abel-Smith 1960).

Nightingale combined her vision of order and management led by trained nurses, with the opportunity created by the Crimean War and her social position and statistical prowess to forge a new future for hospitals. Florence Nightingale, while well known as the founder of modern nursing, was also among the first to use proportional statistics (Cohen 1984). By tracking deaths by cause in military hospitals in Crimea she demonstrated that more soldiers died of hospital environmental causes than of injuries received in battle. Her volume, *Notes on Hospitals*, has been reissued in several versions as an example of one of the first medical outcomes studies (Rosenberg 1989).

2.2 Vestiges of History in Contemporary Nursing Roles

Nursing's history leaves a legacy that gives contemporary professional nurses special legitimacy in the medical division of labor, and a high level of public trust. However, that same legacy creates substantial barriers to the continued evolution of the profession of nursing in contemporary health care. Nursing developed as an ancillary occupation supportive to physicians. Nurses gained much of their professional status and public legitimacy from their close association with physicians. The subordinate status of nursing relative to medicine in the medical labor hierarchy, however, has hindered their quest for greater autonomy (Friedson 1970). The dominance of medicine over nursing is often cited as the archetype of settlement of competing occupational jurisdictional claims by subordination (Abbott 1988). Medical dominance over nurses is linked to the organizational frame of the modern hospital where the majority of nurses practice and where physicians exercise considerable professional authority. Indeed, nursing's quest for professional autonomy was not realized to any great extent until they began to practice in larger numbers and in a broader array of roles outside the hospital.

The association of early nurse training with religious orders has reinforced a continuing expectation of service and charity. These attributes are no longer valued as highly in Western culture, and no longer serve as sufficient motivating forces to recruit and retain professional nurses in today's world. Despite a shift from caring and nurturing functions to more technological responsibilities akin to those associated with male roles in other occupations, nursing has

retained a feminine image and is thus subject to sex-role stereotyping and sex discrimination in professional relationships and labor practices. The large-scale entry of women into the labor force and into higher education has facilitated the advancement of nursing by increasing the absolute number of nurses, raising overall educational levels, and broadening scope of practice. However, women now have many more career options than in years past and nursing is attracting a smaller share of new career aspirants as suggested by a steady rise in the average age of practicing nurses (Buerhaus et al. 2000). Prospects for attracting enough qualified nurses to meet projected future demands for health care are uncertain, which poses a threat to the professional advancement of nursing. In the past when the supply of nurses has been insufficient to meet market demand, nursing has been vulnerable to the substitution of less-educated personnel and pressures to lower educational standards.

2.3 Nursing Education

The education of professional nurses has undergone a dramatic transformation since the 1970s that is still underway world-wide (Lynaugh and Brush 1996). Nursing education developed in hospital apprentice-like programs with a large service component. Indeed, until the Great Depression forced many graduate nurses into hospital employment, hospital nursing care was supplied largely by students in training. Leaders in American nursing advocated early in the twentieth century the standardization of nursing education within institutions of higher education. The Flexner Report of 1910 had brought about similar reform in medical education in the USA. The analogue of the Flexner Report for nursing, known as the Goldmark Report, was published in 1923 and recommended the baccalaureate degree as the minimum entry-level educational requirement for professional nursing. Subsequent reports and commissions supported the recommendation culminating in the American Nurses Association in the late 1970s setting 1985 as the target year for requiring the baccalaureate degree. While progress has been made toward the goal of baccalaureate education for nurses, the new century dawned without this goal having been achieved in the USA, although other countries including Australia, New Zealand, Iceland, Chile, and United Kingdom have implemented an academic degree requirement for new entrants into nursing.

Friedson (1970) described nursing as an incompletely closed profession because of its inability to set a minimum educational requirement. Three generic types of education programs prepare professional nurses: three-year hospital diploma programs, two-year associate degree programs, and four-year college/university programs. Neither employers nor legal practice acts distinguish between professional nurses by type of education. The proportion of US nurses

trained in hospital diploma programs declined from over 80 percent in 1960 to less than 8 percent in 1996. Community college programs instead of baccalaureate education replaced the diploma programs and now train over 60 percent of US nurses (Aiken and Gwyther 1995). By 1996, only about 40 percent of US nurses held a baccalaureate or higher degree, and 10 percent held master's or doctoral degrees (Moses 1997).

3. Nurse Practitioners and Physician Assistants

Nurse practitioners, also known as advanced practice nurses, and physician assistants are two relatively new additions to the medical division of labor, both originating in the USA in the 1960s. Nurse practitioners represent more of an evolution in the profession of nursing than a distinct new occupation while physician assistants constitute a new allied health profession.

Nurse practitioners (NPs) are registered nurses with advanced clinical training (Mezey and McGivern 1999). In the USA a master's degree is required for certification in advanced practice nursing. Nurse practitioners provide generalist care including routine physical assessment, treatment of uncomplicated episodic illness, continuing care for persons with stable chronic conditions, and acute care in hospital settings. Nurse midwives, a type of advanced practice nurse, provide routine obstetrical care including normal deliveries as well as routine health care for women. Nurse practitioners have successfully defined a legal scope of autonomous practice although attempts by physicians to exert control over their practice and remuneration is a continuing source of conflict (Sage and Aiken 1997).

Physician assistant (PA) education is generally 24 months in length. Programs exist in four different academic pathways: certificate programs, associate degree programs, bachelor's degree programs, and master's degree programs although the PA component is the same regardless of degree level (Hooker and Cawley 1997). A major difference between physician assistants and nurse practitioners is that physician assistants have accepted a legally dependent role working under the direction of a physician while nurse practitioners have fought for and obtained a legal scope of practice that does not require physician supervision (Sage and Aiken 1997).

3.1 Historical Context

While nursing's history and evolution is strongly tied to the development of the modern hospital, nurses have a long and distinguished history in public health and home care, settings in which they exercised considerable autonomy and professional leadership. Lillian Wald invented the term 'public health nurse' in 1893 to describe a role for nurses that was a com-

bination of health promotion, illness care, and social reform (Buhler-Wilkerson 1993). Public health nursing and public health in general underwent substantial erosion in the USA with the rise of health insurance after World War II. However, the foundation for the emergence of nurse practitioners was the legacy of public health nursing.

The evolution of expanded roles for nurses and the development of physician assistants is often linked to a perceived shortage of generalist physicians in the USA in the 1960s and the market demand for physician substitutes. In the case of nurse practitioners, however, these factors created a favorable environment for expanding the scope of nursing practice that had long been advocated by nursing's leaders and that was exemplified by the early public health nurses. Thus, nurse practitioners, while often considered new allied health professionals, are really not new at all but nurses with legally sanctioned expanded scope of practice including some domains formerly controlled solely by physicians, such as drug prescription authority (Lynaugh and Brush 1996, Mezey and McGivern 1999). While educational and legal requirements for advanced nursing practice vary across countries, there is a global trend to replicate the successful US model (Fagin 1990, *British Medical Journal* 2000).

The origins of the new occupation of physician assistants in the USA stem from the return to the civilian workforce of military medical corps personnel in the 1960s at a time of a perceived domestic physician shortage. Most initial recruits to physician assistant programs were experienced medical corps personnel and the initial training programs were short-term on-the-job training. While the profession was predominantly male in its inception, it is now almost equally male and female, and the most recent graduating classes are over 60 percent female (Hooker and Cawley 1997). The early programs were located at medical schools but did not have academic standing or lead to degrees. Over time, degree-granting programs have been incorporated into schools of allied health.

Nurse practitioner education was rapidly incorporated into the large existing university-based infrastructure of schools of nursing. Enrollments in nurse practitioner programs grew steadily and by 1996 over 70,000 nurses in the USA had formal preparation as nurse practitioners, primarily at the master's level (Moses 1997). There was no comparable existing educational infrastructure for physician assistant education, and thus their numbers have grown more slowly than those of nurse practitioners, reaching 29,000 in 1997 (Cooper et al. 1998).

3.2 Practice Issues

A large research literature has developed on the safety and effectiveness of care provided by nurse practitioners and physician assistants. The Office of Tech-

nology Assessment of the US Congress conducted a synthesis of research published through 1985 and concluded that nurse practitioners and physician assistants working within their areas of competence provided comparable care to that provided by physicians, and in some aspects of care, such as patient satisfaction, nurse practitioners and physician assistants had better outcomes than physicians (Office of Technology Assessment [OTA] 1986). A special issue of the *British Medical Journal* (2000) reports similar findings from four randomized clinical trials evaluating care provided by nurse practitioners in the UK.

Competing jurisdictional claims between professions are unlikely to be solved by scientific evidence. Indeed, no matter how many studies have been undertaken including randomized controlled trials (Mundinger et al. 2000), many physicians have not accepted nurse practitioners (Kassrer 1994, Sox 2000). The opposition of physicians notwithstanding, the market demand for nurse practitioners has grown steadily in the USA, a trend that appears to be spreading to other countries (Norman 2000).

4. Other Allied Health Professionals

Over 200 different occupations and professions comprise the allied and auxiliary health care workforce. In the USA, this broad category accounts for 60 percent of the nation's 10.5 million health care workers and includes personnel who provide therapeutic, diagnostic, informational, and environmental services across the full array of health care delivery settings (Ruzek et al. 1999). The non-nursing allied health occupations are diverse in roles, functions, and education (Hamburg 1975). Many of these occupations are products of the twentieth century, having evolved in response to demands created by new technologies and niche roles created by changing patterns of medical care. Some of the alternative and complementary medicine occupations such as acupuncturists have roots in cultures that predate the development of modern Western medicine but are nevertheless considered newcomers within the Western division of medical labor (Eisenberg et al. 1998).

Education, work setting, and license/regulatory requirements vary vastly across the allied health occupations. Some fields require six months to one year training and are institutional on-the-job training programs, while others are university based. Physical therapy (PT) and occupational therapy (OT) are illustrative of the allied health professions that emerged primarily in the twentieth century.

4.1 Roles and Work Settings

Physical therapists plan and administer treatment to relieve pain, improve functional mobility, maintain cardiopulmonary functioning, and limit the disability of people suffering from physically disabling injury or

disease. Approximately one-third are employed in hospitals; other major employers include rehabilitation facilities, home health agencies, nursing homes, school systems, and clinics. About 20 percent of physical therapists are in private practice.

Occupational therapists promote learning of skills necessary to perform daily tasks, diminish or correct pathology, and promote and maintain health. Occupational therapy focuses on both mental and physical disabilities. About 60 percent of occupational therapists work with people with physical disabilities. Some specialize in particular age groups such as the elderly or children. Occupational therapists working with the mentally ill or disabled in institutional and community-based settings typically provide activities that help mentally ill and disabled people learn to cope with daily stresses and manage their work and leisure time more efficiently (Institute of Medicine 1989). Most occupational therapists are employed by healthcare organizations and private practice is not common.

4.2 Evolution of Occupational and Physical Therapy

Occupational therapy traces its origins back 200 years to a French physician's observation that mental patients given menial tasks to perform improved in functional capacity faster than those who were idle. In the USA at the end of the eighteenth century, physician Benjamin Rush introduced work as a treatment for mentally ill patients at Philadelphia's Pennsylvania Hospital, and in 1906 the first training course for occupational therapists was established in Boston (Institute of Medicine 1989). World War I created a demand for rehabilitation of large numbers of injured soldiers, which served as a major impetus for expanding the scope of practice of occupational therapy to physically as well as mentally disabled persons, and for establishing the new occupation of physical therapy.

4.3 Education and Characteristics

Not surprisingly, given the growth impetus for physical and occupational therapists associated with the two World Wars, the early recruits to both occupations were predominately women. Over time both have remained predominately female occupations, although 25 percent of physical therapists are men compared with 5 percent of occupational therapists and 6 percent of nurses. The first training program in physical therapy in the USA began in 1917 as a short course for women aides. The success of these aides in the aftermath of World War I resulted in federal funding to increase the supply of therapists during World War II through accelerated programs to train college graduates in related fields such as physical education for roles in therapy. University-based degree-granting programs in physical therapy were

established concurrently. Physical therapy has thus been a university-based educational program almost from its inception.

Since 1979, the American Physical Therapy Association has advocated graduate level preparation for the profession. As of 2002, the Association will no longer accredit baccalaureate professional programs and is moving to master's and doctoral level professional education. While practitioners with a baccalaureate degree may still practice, the purpose of the change in accreditation is to shift the educational patterns for future cohorts. Physical therapy does not support career mobility from the physical therapy aide category to the professional category and has established non-overlapping educational requirements for the two groups thus avoiding the problem nursing has faced. Physical therapy has also developed viable options for private practice, which allows the occupation to continue to evolve independently, to some extent, from institutional managerial control of salaried practice. Physical therapy is considered to be more visible and autonomous than occupational therapy for these reasons (Institute of Medicine 1989).

Occupational therapy in many ways more closely resembles nursing in its composition and education than physical therapy, perhaps because its roots are in the nineteenth century with nursing and its knowledge and intervention domains are less discrete than physical therapy. The first formal 'professional' standards for education in occupational therapy were established in 1923 as 12-month training programs for high school graduates, in contrast to physical therapy, which moved directly to university-based education. There are presently two educational levels for practice: technical training at the two-year associate degree level and professional education at the baccalaureate degree for registered occupational therapists.

Occupational therapy has had difficulty attracting students reportedly because of its focus on mental health services, which are highly stigmatized, and poorer remuneration opportunities than alternative fields such as physical therapy. In 1995, there were 54,000 registered occupational therapists compared to 102,000 physical therapists in the USA (Bureau of Labor Statistics 1996).

5. Global Perspectives

Developing countries, which contain 84 percent of the world's population, account for only 11 percent of the world's health professionals (Schieber and Maeda 1999). Physicians, the most expensive category of health personnel, are often in relatively generous supply compared to nurses and allied health personnel in countries with modest annual health expenditures. An imbalance in the ratio of physicians to allied health workers has a number of consequences including excess expenditures on secondary and tertiary care leaving inadequate resources for primary care and

public health as well as geographic imbalances in health care because physicians tend to locate in urban areas. The range in the ratio of nurses to doctors is enormous across countries: from one doctor to three or four nurses in northern European countries, to one doctor to 1.5 nurses in south-eastern Europe, and from one nurse to two doctors in Argentina, to one nurse to five doctors in Pakistan (Bankowski and Fülöp 1987).

Allied health workers provide much of the medical care in poor countries but are in short supply and often poorly trained. The potential of nurses and nurse midwives to improve access to and quality of health care in developing countries is well recognized. However, many countries have insufficient resources to create employment opportunities for qualified nurses (Bankowski and Fülöp 1986). Thus, nurses trained in developing countries migrate in substantial numbers to more affluent countries. Additionally allied health workers with some medical training such as feldshers in Russia are also in short supply because of the practice of admitting experienced health care workers to physician training programs (Ryan and Thomas 1996).

Developing countries tend to emulate the medical care systems in industrialized countries by building hospitals, purchasing technology, and training large numbers of physicians. Ultimately the prospects for improving health in the developing world are in investments in frontline allied health workers, primary care, and public health and safety.

Thus, nursing and other allied health occupations have good prospects for future growth in developing countries as well as in the sophisticated and highly technological health systems of industrialized countries.

See also: Alternative and Complementary Healing Practices; Health Care Delivery Services; Health Care: Legal Issues; Health Care Markets: Theory and Practice; Health Care Organizations: For-profit and Nonprofit; Health Care, Rationing of; Health Insurance: Economic and Risk Aspects; Managed Care; Medical Profession, The; National Health Care and Insurance Systems; Nursing Homes

Bibliography

- Abbott A 1988 *The System of Professions*. University of Chicago, Chicago
- Abel-Smith B 1960 *A History of the Nursing Profession*. Heineman, London
- Aiken L H, Gwyther M E 1995 Medicare funding of nurse education: the case for policy change. *Journal of the American Medical Association* **273**(19): 1528-32
- Bankert M 1989 *Watchful Care: A History of America's Nurse Anesthetists*. Continuum, New York
- Bankowski Z, Fülöp T 1987 *Health Manpower Out of Balance*. Council for International Organizations of Medical Sciences, Geneva, Switzerland

- British Medical Journal 2000 Doctors and nurses. Special issue (7241: 15 April)
- Buerhaus P I, Staiger D O, Auerbach D I 2000 Implications of an aging registered nurse workforce. *Journal of the American Medical Association* **283**(22): 2948–54
- Buhler-Wilkerson K 1993 Bringing care to the people: Lillian Wald's legacy to public health nursing. *American Journal of Public Health* **83**: 1778–86
- Bureau of Labor Statistics, US Department of Labor 1996 *Occupational Outlook Handbook*. US Government Printing Office, Washington, DC
- Cohen I B 1984 Florence Nightingale. *Scientific American* **250**: 128–32
- Cooper R A, Laud P, Dietrich C L 1998 Current and projected workforce of nonphysician clinicians. *Journal of the American Medical Association* **280**(9): 788–94
- Eisenberg D M, Davis R B, Ettner S L, Appel S, Wilkey S, Van Rompay M, Kessler R C 1998 Trends in alternative medicine use in the United States, 1990–1997: results of a follow-up national survey. *Journal of the American Medical Association* **280**(18): 1569–75
- Fagin C M 1990 *Nursing Leadership: Global Strategies*. National League for Nursing, New York
- Fairman J, Lynaugh J 1998 *Critical Care Nursing: A History*. University of Pennsylvania Press, Philadelphia, PA
- Freeman P, Gómez-Dantés O, Frenk J 1995 *Health Systems in an Era of Globalization*. National Academy of Sciences, Washington, DC
- Freidson E 1970 *Professional Dominance: The Social Structure of Medical Care*. Aldine, Chicago
- Hamburg J (ed.) 1974 *Review of Allied Health Education*. University of Kentucky Press, Lexington, KY
- Hooker R S, Cawley J F 1997 *Physician Assistants in American Medicine*. Churchill Livingstone, New York
- Institute for the Future 2000 *Health and Health Care 2010: The Forecast, the Challenge*. Jossey-Bass, San Francisco
- Institute of Medicine 1989 *Allied Health Services: Avoiding Crises*. National Academy Press, Washington, DC
- Lynaugh J E, Brush B L 1996 *American Nursing: From Hospitals to Health Systems*. Blackwell, Cambridge, UK
- Kassirer J P 1994 What role for nurse practitioners in primary care? *New England Journal of Medicine* **330**: 204–5
- Mechanic D 1983 *Handbook of Health, Health Care, and the Health Professions*. Free Press, New York
- Mezey M D, McGivern D O (eds.) 1999 *Nurses, Nurse Practitioners: Evolution to Advanced Practice*. Springer, New York
- Moses E B 1997 *The Registered Nurse Population: Findings from the National Sample Survey of Registered Nurses*. Division of Nursing, US Bureau of Health Professions, Rockville, MD
- Mossialos E, McKee M (eds.) 1998/99 Health care reform, public health, pharmaceutical policy, focus on Russia and Central Asia. *Eurohealth* **4**(6): Special Issue, Winter
- Mundinger M O, Kane R L, Lenz E R, Totten A M, Tsai W, Cleary P D, Friedewald W T, Siu A L, Shelanski M L 2000 Primary care outcomes in patients treated by nurse practitioners or physicians: A randomized controlled trial. *Journal of the American Medical Association* **283**: 59–68
- Norman S 2000 Making sense of higher level practice. *Nursing Standard* **14**(35): 47–56
- Office of Technology Assessment (OTA) 1986 *Nurse Practitioners, Physician Assistants, and Certified Nurse Midwives: A Policy Analysis*. Health Technology Case Study No. 37. US Government Printing Office, Washington, DC
- Parsons T 1951 *The Social System*. Free Press, Glencoe, IL
- Rosenberg C E (ed.) 1989 *Notes on Hospitals: Florence Nightingale on Hospital Reform*. Garland, New York
- Ruzek J Y, Bloor L E, Anderson J L, Ngo M 1999 *The Hidden Health Care Workforce: Recognizing, Understanding and Improving the Allied and Auxiliary Workforce*. University of California, San Francisco Center for the Health Professions, San Francisco, CA
- Ryan T M, Thomas R 1996 Trends in the supply of medical personnel in the Russian Federation. *Journal of the American Medical Association* **276**(4): 339–42
- Sage W M, Aiken L H 1997 Regulating interdisciplinary practice. In: Jost T (ed.) *Relation of the Health Professions*. Health Administration Press, Ann Arbor, MI, pp. 71–101
- Schieber G, Maeda A 1999 Health care financing and delivery in developing countries. *Health Affairs* **18**(3): 193–205
- Sox H C 2000 Independent primary care practice by nurse practitioners. *Journal of the American Medical Association* **283**: 106–8
- Starr P 1982 *The Social Transformation of American Medicine*. Basic Books, New York
- Thomas L 1977 On the science and technology of medicine. *Daedalus* **106**: 35–46

L. H. Aiken

Health Promotion in Schools

Health promotion in schools is a topic of both practical and empirical interest. Schools which provide comprehensive school health promotion programs are more effective in encouraging children to adopt health-enhancing behaviors and in reducing health-compromising behaviors than schools that provide health education alone. Research into the effectiveness of comprehensive school health programs provides evidence of the positive influence that such programs can exert on student health behaviors. However, conditions that enable effective school health promotion programs to be initiated and maintained depend on several school organizational and program implementation factors.

1. Changing Concepts of Health Promotion

The meaning of health and health promotion have been transformed over the 1990s. Influenced by the World Health Organization (WHO), health is seen as more than avoiding illness. Similarly, health promotion involves more than helping people to choose healthy lifestyles.

Health promotion in the USA, Canada, and Western Europe has been influenced by WHO concepts contained in the Ottawa Charter for Health Promotion (1986). Health is viewed as a resource of daily living, a positive concept emphasizing social and personal

resources, as well as physical capacities. Responsibility for health rests not only with the health sector, but includes building healthy public policy, creating supportive environments, strengthening community action, developing personal skills, and reorienting health services. Accordingly, health promotion professionals are expected to work for health through advocacy, enabling individuals, and mediating among diverse sectors. Developments since the Ottawa Charta suggest increasing awareness of the role values play in health-related activity as well as increasing emphasis on social justice, reducing inequalities, and enhancing social cohesion. The Jakarta Declaration on Health Promotion in the 21st Century (1997) finally outlined priorities for promoting social responsibility for health, increasing investments for health development, increasing community capacity, empowering the individual, and securing an infrastructure for health promotion.

Currently, the ever-widening gap between rich and poor is an issue of great concern, and compelling evidence confirms that exclusion, both societal and material, is highly linked to health problems. The potential effects of inequalities on the health of populations have first been stressed by The United Kingdom Charta for Health Promotion, followed by reports from Canada and the US which reported profound variation in health and well-being between poor and nonpoor children. Health inequalities exist in level of mortality and morbidity, accidents and injuries, risk behavior, mental health, school achievement, and family violence. These differences have traditionally been interpreted in terms of risk behaviors resulting from lack of education, and material deprivation. Yet, careful analyses indicate that most of the variance in health outcomes among differing social classes is not accounted for by lifestyle factors. Recent analyses challenge the 'risk group' metaphor, showing that American and European children and adolescents in high-risk settings (low income families, high unemployment of the caregivers, poor neighborhoods) were in poorer health than those living in wealthier settings.

2. Health-promoting Programs

Numerous health-promoting programs have been published; the focus diverges with some publications emphasizing public health concerns and other reflecting an education perspective. In addition, the programs vary with respect to age range, comprehensiveness, theoretical background, and the importance of evaluation. While some studies comprised primarily of survey data, others examine the effect of intervention programs. Several papers incorporate a substantial body of epidemiological and demographic data as basis for their conclusions. Most discuss policy options and recommendations. Thus, substantial vari-

ation exists with respect to such factors as level of detail, use of research data, intended audience, and political viewpoint.

Rather than giving an overview of the current state of research in school-based health promotion (see David and Williams 1987), in the following, some examples of comprehensive programs will be given, prominent trends will be presented, and, finally, frequent domains in health promotion programs will be evaluated.

To date, not many prevention and intervention programs have incorporated the important developments in health promotion values and principles. The Dartmore Health Promotion Study (McIntyre et al. 1996) is one of the rare programs that coordinated school health instruction, health services, and a healthful environment to enhance the program's effect on the health of the heart and mental health of children in grades four, five, and six. Several studies suggest that children are not engaging in physical activity of sufficient length and intensity to improve cardiovascular fitness; during adolescence, the amount of physical activity further decreases. This has resulted in the development of community-wide Youth Exercise Promotion in order to prevent heart diseases and to influence blood pressure in pupils with risk for hypertension.

To give an example, the CATCH (Child and Adolescent Trial for Cardiovascular Health) is the most comprehensive multilevel program which has been extensively evaluated (Perry et al. 1997). The interventions targeted changes at the school level (food services offerings and minutes of class time devoted to physical activity) and at the student level (psycho-social factors, dietary behaviors, physical activity levels, smoking intentions, physiological indicators). The CATCH interventions consisted of: (a) Eat Smart, a school food service program to provide children with meals reduced in fat and sodium; (b) Catch PE, to increase the amount of time students spent in physical activity at school and for other times of the day; (c) Hearty Heart and Friends, 15 classroom lessons to teach dietary and physical activity information through a series of stories; and (d) Family Fun Nights, during which pupils and family members participated in 90 minutes of games, fitness-related activities, and healthy snacks.

Generally, health promotion programs have been confined to highly selective programs designed to identify specific health risks, such as tobacco smoking, alcohol consumption or drug use, and to train health-promoting behaviors, for example with respect to prevention of HIV. Most of these programs are designed for late childhood and early adolescence and pursue multiple strategies for preventing health damage, including more traditional strategies (for example, teaching drug information and/or general communication and problem-solving skills) and approaches based on the social influence model (which emphasizes

strategies for identifying and resisting social pressure to use drugs or alcohol). Time reserved for cognitive instruction was usually low, with most curricula ranging between 5 and 15 lessons. Although the overall cognitive progress in most programs was impressive, this was not sufficient to produce corresponding reduction in use of alcohol, drug or nicotine. Mostly, the effects were stronger and longer lasting for girls than for boys. Frequently, developmentally relevant variables are incorporated, for example increasing peer pressure or a risky lifestyle during adolescence. Resistance-Skill Training(s), however, was not always found to be effective strategies for preventing the onset of alcohol, drug or tobacco use; sometimes counterproductive effects were found.

Health promotion in middle schools is a neglected area, although the middle school years are a time of increasing risk for negative health related outcomes. Health promotion in the middle school mostly focuses on two approaches, a social decision approach (e.g., teaching strategies and skills that help children to make favorable decisions when confronted with health damaging situations and offers) and social skills training (in which children are prompted, coached, and guided to practice using the skills). These approaches are based on the observation that children are unlikely to retain and use information conveyed solely through formal lessons; they need opportunities for reinforcement and practice. Further, not many health promotion or health education programs are developed for elementary school children. The Claxton Elementary School Health Program (Landis and Janes 1995) is one of the rare examples. In this project, school administrators, parents, and local physicians worked together to discover what types of healthy and unhealthy activities children and parents are engaged in, and identified which health problems are of greatest concern to parents and teachers.

Professionals in school health education have frequently emphasized parent and family involvement in school health education. Unfortunately, preparing prospective educators for such involvement often has been overlooked. Teacher training for involving families in school health education were designed to provide staff members with the background necessary to incorporate family involvement concepts and activities into their classes. This may include, for example, (a) newsletters to provide information about health and the school health program to family members, (b) family calendars about quick activities—less than five minutes—that families can do daily, and (c) information about children's books, television programs and movies that address health topics.

Recently, the interaction with families of high-risk children became a major focus, particularly in settings outside school. The Triple P Program, for example, aims at helping parents of kindergarten children with sleeping and eating disorders to cope with these

challenges by developing positive parenting attitudes and behaviors. Although it targets categorical symptoms rather than their common antecedents, it is highly valuable for preventing more severe future health problems in offspring.

To summarize, health promotion and education programs focus on late childhood and adolescent subjects and target problem behavior (which is highly inter-related such as drug and alcohol use) and unhealthy life styles (with respect to nutrition and physical activity). Stress prevention programs or programs teaching functional coping skills are less frequently applied. Most programs are centered in and around schools, but preventive approaches sometimes reach out to families with preschool children at risk

3. Evaluation of School-based Prevention and Intervention Programs

Quantitative research summaries are needed to deal with questions such as the relative efficacy of prevention and intervention programs with different aims and program orientations. In assessing program outcomes, five criteria should be met: (a) the use of appropriate comparison groups; (b) the use of appropriate pretests, post-tests, and follow up assessments; (c) examining research attrition; (d) the validity of dependent variable assessments; and (e) the use of statistical procedures appropriate for program evaluation research. In the following, quantitative reviews of prevention and intervention programs targeting adolescent drug use, smoking, and alcohol consumption were presented. They represent a new approach in that program classification schemes were used to categorize outcome, which were then employed in a quantitative meta-analysis of study effect size, based on the above criteria.

3.1 Meta-analyses of Smoking Prevention Programs

A substantial number of research report evaluating programs designed to prevent adolescent smoking have been published in the 1980s and 1990s. A very important set of questions deals with: (a) how successful these programs have been in general; (b) whether programs of different orientation show different levels of success; and (c) what variables, in addition to program orientation, may account for different levels of success. The meta-analyses conducted by Bruvold (1993) was based on 94 separate intervention studies published between 1970 and 1990. These studies evaluated school-based programs designed to deter adolescents from smoking tobacco products and were screened for methodological rigor. Each study reviewed was independently evaluated on the five methodological criteria detailed above. Forty-eight studies with adequate and defensible method-

ology were included in the meta-analyses. Program classification revealed four different types of orientation of the prevention program: (a) rationally-oriented information approach; (b) developmentally-oriented affective education; (c) social norms-oriented alternatives; and (d) social reinforcement-oriented social pressures approach. The rational model supports traditional classroom interventions where information about the short- and long-term health effects of tobacco is didactically presented. Reinforcement, social norms, and developmental interventions have two features in common: all are relatively new, and all were developed because of dissatisfaction with the uncertain effectiveness of the traditional didactic approach. Weighted average effect sizes were computed for all 48 better methodology studies for knowledge, attitude, and behavior, respectively. Positive results with similar high effect sizes for knowledge were found in programs with developmental, social norms, and social orientation. Results for attitude were generally not as positive. Programs classified primarily as social reinforcement-oriented seem to have had the most impact on attitudes, followed by those classified primarily as developmental. This suggests that newer orientations were more successful in modifying attitudes than more traditional rational approaches. Behavioral effect sizes were found to be largest for interventions with a social reinforcement orientation, moderate for interventions with either a developmental or a social norm orientation, and small for interventions with the traditional rational orientation. Altogether, the results of this meta-analysis suggest that rational orientation had very little if any impact on behavior and attitudes, and that the developmental and social reinforcement orientation revealed similarly positive effects on attitudes and behavior. Not much difference, depending on program orientation, emerged with respect to the cognitive outcome.

3.2 Comparison of Tobacco and Alcohol Intervention Programs

Frequently, school-based interventions were designed to deter children and adolescents from starting to consume alcohol and tobacco, and to convince those already consuming these substances to stop or reduce consumption. Bruvold and Rundall (1988) conducted a meta-analysis on those programs that combined tobacco and alcohol prevention. Again, they classified the programs according to their theoretical orientation into rational, social reinforcement, developmental, and social norms. In addition, the major methodological requisites for research evaluating school-based tobacco and alcohol interventions detailed above were applied to the studies incorporated in the meta-analysis. Altogether, 19 studies were found which had a control or comparison group, reported attrition rate, employed pretest, post-test and follow-up measures,

and used adequate statistical analyses. A very consistent result was that the intervention program, based upon the rational model produced fewer positive behavior and attitude effects but more positive knowledge effects than those in which the rational model was *not* dominant. Knowledge effects were consistently positive over all studies reviewed, but attitude effects were mixed, some positive and some negative. For smoking behavior, the effect size was satisfactory for long- and short-term measures, but behavioral results for alcohol were meager for long- and short-term measures. Thus, while the alcohol interventions had an effect in the desired direction on alcohol consumption, the average impact was not as large as that for smoking.

3.3 Effectiveness of Drug Abuse Resistance Education

School-based drug prevention programs have been an integral part of the US and European antidrug campaign during the 1980s and 1990s. Although programs have proliferated, Project DARE (Drug Abuse Resistance Education) is the most widely implemented school-based, drug use prevention program in the United States, and has been adopted by approximately 50 percent of school districts nationwide. The drug prevention curriculum offers 17 lessons, usually once a week, and is applied in elementary schools, as well as in junior and senior high schools. Lessons focus on teaching pupils the skills needed to recognize and resist social pressures to use drugs. In addition, lessons focus on providing information about drugs, teaching decision-making skills, building self-esteem, and choosing healthy alternatives to drug use. Didactics include lectures, group discussion, question and answer sessions, audiovisual material, workbook exercises, and role playing. Ennett et al. (1994) used meta-analysis to review methodological rigorous evaluations of short-term effects. The DARE effect size for drug use behavior was very low across the studies. For all outcomes considered, the DARE effect size means were substantially smaller than those of programs emphasizing social and general competencies and using interactive techniques. Considering the enormous amount of public funding for DARE, the short-term effectiveness of DARE for reducing or preventing drug use behavior can not be considered as satisfactory.

4. Implications for School Health Programs and Personnel

The School Health Education Evaluation indicated that for school health education to promote health for children, schools should devote at least 50 hours of classroom time each year to health instruction. How-

ever, in many industrialized countries, particularly in southern and eastern Europe, these standards are not met (Takashini and Hamburg 1997). Limited resources, including curricular time, make it unlikely that health promotion in schools will realistically increase in strength. Thus, it is important to develop more efficient ways to promote children's health, for example by adding complementary interventions to school health education from other sources.

Meta-analyses of more recent school-based studies show that it is necessary to carefully evaluate the programs. Overall, improvements of methodology and conceptualization of programs is noticeable, but still a need exists to upgrade substandard procedures regarding experimental attrition and establishing the validity of dependent variable measures. Further, there is a need for developing program intervention models that clearly and fully follow tenets of the guiding theory. In addition, prevention and intervention programs which have been designed and evaluated so far are largely confined to certain age groups (frequently to adolescents) and limited to the treatment of problem behaviors (such as drug, alcohol, or tobacco use). The health needs of younger ages, for example preschool children or children from grades one to five, are largely neglected.

Based on the recent developments of health promotion values and principles detailed in the Jakarta Declaration, health promotion in schools should acknowledge societal factors such as inequalities and poverty as important determinants of health in children and adolescents. Health promotion in schools therefore has to incorporate the community level of life quality in designing programs. School health personnel need to deal with issues that influence children and youth's health, but is beyond the traditional focus of health education and behavior change. They need to assess the quality of life as it is perceived by children and adolescents, and identify the specific needs of individual children. Then, community-based action can be developed to address these issues.

See also: Adolescent Health and Health Behaviors; Alcohol Use Among Young People; Childhood Health; Health Education and Health Promotion; Health Interventions: Community-based; Mental Health: Community Interventions; Physical Activity and Health Promotion; Public Health; Smoking Prevention and Smoking Cessation

Bibliography

- Bruvold W H 1993 A meta-analysis of adolescent smoking prevention programs. *American Journal of Public Health*. **83**: 872–80
- Bruvold W H, Rundall T G 1988 A meta-analysis and theoretical review of school-based tobacco and alcohol intervention programs. *Psychology and Health*. **2**: 53–78
- David K, Williams T (eds.) 1987 *Health Education in Schools*. Harper and Row, London
- Ennett S T, Tobler N S, Ringwalt C L, Flewelling R L 1994 How effective is drug abuse resistance education? A meta-analysis of project DARE outcome evaluations. *American Journal of Public Health*. **84**: 1394–401
- Jakarta Declaration on Health Promotion into the 21st Century 1997 World Health Organization, Geneva
- Landis S E, Janes C L 1995 The Claxton elementary school health program: Merging perceptions and behaviors to identify problems. *Journal of School Health*. **65**: 250–4
- McIntyre L, Belzer E G, Manchester L, Blanchard W, Officer S, Simpson A C 1996 The Dartmouth Health Promotion Study: A failed quest for synergy in school health promotion. *Journal of School Health*. **66**: 132–7
- Ottawa Charta on Health promotion 1986 World Health Organization, Geneva
- Perry C L, Sellers D E, Johnson C, Pedersen S, Bachman K J, Parcel G S, Stone E J, Luepker R V, Wu M, Nader P R, Cook K 1997 The child and adolescent trial for cardiovascular health (CATCH): Intervention, implementation, and feasibility for elementary schools in the United States. *Health Education & Behavior*. **24**: 716–35
- Takanishi R, Hamburg D A (eds.) 1997 *Preparing Adolescents for the Twenty-First Century: Challenges Facing Europe and the United States*. Cambridge University Press, New York, pp. 82–107

I. Seiffge-Krenke

Health Psychology

Health psychology can be defined as the application of psychological knowledge and methods to the study, prevention, and management of physical diseases and disorders. Evident in this definition is the fact that health psychology is both a science and an applied discipline intended to promote human welfare. However, unlike many psychological disciplines, such as clinical or counseling psychology, health psychology addresses issues of physical rather than mental health. Emotional adaptation and mental health are considered in health psychology only to the extent that they influence—or are influenced by—physical health and illness (Schwarzer and Gutiérrez 2000).

The formal emergence of health psychology in the 1970s was influenced by a variety of factors. By that time, for industrialized nations infectious illnesses such as smallpox and influenza were no longer the leading cause of death. Through advances in basic biomedical science and its application in public health, vaccines, antibiotics, and acute clinical care, infectious illnesses had been replaced by several chronic diseases as the leading sources of morbidity and mortality. Cancer, coronary heart disease, cerebrovascular disease, and diabetes differ from infectious diseases in three fundamental ways that assured the emergence of health psychology. First, these illnesses are influenced by behavior. Daily habits or 'lifestyle' factors are

major contributors to their development and course. Second, unlike the acute crisis of infectious illness, chronic illnesses have far-reaching impacts on emotional, social, and vocational adaptation over long periods of time. Finally, behavior change is a key component of care for such conditions. Hence, behavior plays a prominent role in the development, impact, and management of the current major health threats in industrialized nations, making psychological research and practice essential in comprehensive medical research and care. This is the case for our most common chronic diseases, such as cancer and heart disease, and for more recently emerging diseases such as HIV/AIDS.

1. History, Foundations, and Related Disciplines

The emergence of health psychology was made possible by a variety of developments at the interface of behavioral and biomedical science. The work of Walter Cannon, Hans Selye, and other physiologists working in the 1930s, 1940s, and 1950s described physiological changes induced by environmental stressors and negative emotions (e.g., the 'fight or flight' response). This identified a pathway through which psychological factors could impact physical health. Shortly thereafter, epidemiologists determined that several behaviors—notably smoking, excessive intake of fat and calories, and low levels of physical activity—were risk factors for the subsequent development of cancer and cardiovascular disease. Elsewhere, psychologists were beginning to explore the application of learning principles to the treatment of physical conditions, as in the seminal work by Neal Miller on the operant conditioning of physiological responses and the work of William Fordyce on the use of behavioral contingencies in the management of pain and disability. At the same time, clinical psychologists—led primarily in the United States by Joseph Matarazzo—were beginning to establish research and clinical roles within academic medicine through appointments to medical school faculties and the establishment of related academic medical departments and services.

In response to a variety of government and organizational initiatives, health psychology became a recognized subdiscipline in the late 1970s. For example, it was established as a formal division within the American Psychological Association in 1978, and the first health psychology handbook was published in 1979 (Stone et al. 1979). At the same time, an interdisciplinary field—behavioral medicine—was established, sharing the focus on behavior in the development, prevention, and management of medical conditions (Schwartz and Weiss 1978). Although most health psychologists view themselves as engaging in behavioral medicine, that field includes concepts,

methods, and professionals from medicine, public health, and other disciplines. Hence, health psychology is a component of this broader field. Because health psychology addresses the impact of behavior on health beyond its role in established disease and disability, its emergence represented a departure from older fields such as medical and rehabilitation psychology. The former refers traditional clinical psychological services (e.g., assessment, consultation, and psychotherapy) within medical settings, whereas the latter refers to similar services and research in the context of physical disabilities (e.g., spinal cord or brain injuries).

The biopsychosocial model is a conceptual cornerstone in health psychology and behavioral medicine. In the traditional biomedical model which underlies much of modern medical research and practice, disease is seen as stemming from disordered biological systems. Aberrations in biochemistry or physiology cause disease, alter its course, and are intervention targets. Psychological or social factors are seen as largely irrelevant. In a landmark paper, Engel (1977) articulated the biopsychosocial model as an alternative. Here biological, psychological, and social processes are seen as inter-related influences on and effects of health and illness. The onset, course, and treatment of illness is best understood as involving each of these interacting, reciprocally determined levels of analysis.

Current research in health psychology consists of three general topics. The first—health behavior and prevention—focuses on the determinants of behaviors that influence disease (e.g., diet, exercise, smoking), and the effectiveness of interventions to reduce these risks. The second—stress and illness—focuses on direct connections between psychological processes and the development of illness. Rather than the intervening effects of habits and lifestyle factors, this research examines the impact of environmental stress and related personality and social processes on psychobiological mechanisms and the pathophysiology of disease. The third research area—the psychosocial impact and management of illness—focuses on the emotional and social consequences of illness and psychological interventions intended to improve adjustment and supplement traditional medical care.

2. Health Behavior and Prevention

Decades of epidemiological research have demonstrated the impact of daily habits on health (see *Health Behaviors*). A variety of conceptual models have been developed to describe the determinants of these behaviors, and thereby provide targets for interventions intended to reduce health risks (Armitage and Conner 1999). For example, the health belief model suggests that several cognitive factors influence health behav-

ior: the value or importance an individual places on health; beliefs about vulnerability conferred by a specific behavior; beliefs about the severity of the related illness; the individual's beliefs about the effectiveness of a behavior change in reducing risk; and a view of the relative costs and benefits of that change. From this perspective, interventions designed to increase the perceived importance of good health, strengthen the belief that a specific behavior impacts health, make clear the severity of the disease or condition to be prevented, emphasize the risk reducing impact of behavior change, and make salient the relative benefits of this change in promoting more healthy lifestyles. Based on this and closely related models of health behavior, interventions ranging from individual counseling to national advertising campaigns have been developed.

Other models of health behavior describe the adoption of a healthier lifestyle as a cyclical process, in which people consider and initiate behavior change, often revert to less healthy habits, and perhaps later renew their efforts (Weinstein et al. 1998). Different interventions may be optimal at different stages in this process. Informational strategies derived from the health beliefs model may be best for individuals who are in the earliest stages, either before or as they begin to consider a change in lifestyle (see *Health Behavior: Psychosocial Theories*). In contrast, active planning about the nature of 'high risk' situations where a return to old habits is likely and rehearsal of strategies for avoiding or managing such relapses may be useful after they have initiated changes. A specific health behavior change can also be addressed to many different intervention levels. For example, to increase regular physical activity, individual level counseling can be supplemented by small group or organizational level interventions such as couple, family, or workplace programs. Further, public policy level interventions can be effective, such as taxes to increase the cost of cigarettes or the construction of public recreational facilities.

For many health behavior changes, the biological level of analysis is a critical consideration. Nowhere is this more evident than in the case of smoking cessation. It has become quite clear that nicotine is an addictive component of tobacco, and that habitual smoking can be seen, in part, as the self-regulation of this addictive agent. Hence, nicotine replacement therapy (e.g., nicotine gum or trans-dermal patches) is a valuable addition to psychological or behavioral approaches.

The biopsychosocial approach to health behavior change is also evident in our current understanding of the nature and treatment of obesity. Body weights 25 percent or more above normal are associated with increasing risk of diabetes, high blood pressure, coronary heart disease, and stroke. Further, in industrialized countries, the prevalence of obesity is increasing. Vulnerability to obesity is influenced by

genetic factors. There are several likely physiological 'sites' of these genetic vulnerabilities, including resting metabolic rate, the degree of weight gain in response to overfeeding, and the location of fat deposition in response to excess energy intake. The energy balance model identifies two additional influences on obesity—the amount of energy (i.e., calories) consumed and the amount expended. Simply, individuals gain weight when the amount of energy consumed exceeds energy expended. Hence, caloric restriction (i.e., dieting) and increased energy expenditure through physical activity are critical determinants of obesity and essential components of treatment. However, larger sociocultural and economic factors also affect obesity. In industrialized nations, the availability of high-fat, energy-rich foods has increased dramatically in recent decades, and levels of leisure time and work-related physical activity have decreased. Especially for the genetically vulnerable, this sociocultural and economic context for the individual's energy balance can be a 'toxic environment' promoting obesity (Hill and Peters 1998). Hence, our understanding of health behavior influences on obesity is facilitated by the comprehensive perspective of the biopsychosocial model.

Smoking and obesity are very difficult to treat in adulthood, even with multicomponent, multilevel interventions. Research and practice on health behavior and prevention has begun to address the prevention of unhealthy behavior. That is, rather than focus on improved methods of smoking cessation and weight loss in adults, more attention has been paid recently to methods of preventing smoking and obesity in childhood and adolescence, and on the modification of these and other risk factors in youth. This trend has necessitated a developmental view of health habits and behavioral risk factors, as well as age-appropriate interventions at the individual, family, school, and community levels. Even though the modification of health behaviors in adulthood remains a priority in health psychology, early prevention is likely to become more widespread in the future.

The emergence of health psychology and its focus on health behavior change preceded by only a few years the identification of Acquired Immune Deficiency Syndrome (AIDS) and the HIV infectious agent. Quickly, epidemiological and clinical research identified intravenous drug use and unprotected sexual activity as primary risk factors. As a result, health psychologists have been closely involved in efforts to prevent HIV infection (see *HIV Risk Interventions*). As in other health behavior change efforts, multiple levels of intervention have been employed. Individual level programs include educational interventions based on the health behavior model and related approaches such as enhancing the individual's skill in negotiating condom use during intercourse. Community level approaches have focused on changing norms regarding sexual behavior and advocacy of 'safer sex' by

visible members of high-risk communities. Institutional approaches have involved increased access to clean needles for drug users.

3. Stress and Disease

A second major focus of research and practice in health psychology involves the impact of stress, negative emotions, and other psychosocial factors on the development of physical illnesses (see *Stress and Coping Theories*). In contrast to the effects of health behavior described above, this area addresses more direct connections between mind and body. In the general view guiding this research area, the stresses and strains of everyday life exact a toll on the individual's physiology. Accumulated over many days, months, or even years, such physiological changes can initiate, hasten, or exacerbate the development of illness. Further, some characteristics of the individual's personality or social environment can make them more or less susceptible to these psychological influences on disease.

A growing array of environmental and personality characteristics have been identified in epidemiological research as risk factors for physical illness and premature death. For example, individuals who are socially isolated or report low levels of available social support are at increased risk of cardiovascular disease, cancer, and premature death (House et al. 1988). High levels of job stress also place people at greater risk of illness (Schnall et al. 1994), as does the chronic stress of caring for seriously ill family members (Schulz and Beach 1999). Individuals who are prone to anger and hostility are at increased risk of developing cardiovascular disease and dying prematurely, as are people who report symptoms of anxiety, depression, pessimism, and hopelessness (Smith and Gallo 2001).

People who are socially isolated, experiencing high levels of job stress, or prone to negative emotions may be more likely to smoke, overeat, consume too much alcohol, or avoid regular exercise. These health behaviors could account for some of the effect of the social environment and personality on health. However, the association of these psychosocial risk factors with subsequent morbidity and mortality remains significant even when the effects of health behavior are controlled (Adler and Matthews 1994). The prevailing view in this research area is that the physiological effects of stressful environments and negative emotions are the link between psychosocial risk factors and subsequent disease. These mechanisms are best understood in two general pathways—the effects of stress on the cardiovascular system (Rozanski et al. 1999) and its effects on the immune system (Kiecolt-Glaser and Glaser 1995).

Environmental threats and demands evoke transient increases in heart rate, blood pressure, and concentrations of various hormones (e.g., epinephrine,

norepinephrine, cortisol, etc.). In human and animal research, over time these stress-induced physiological changes appear to promote more enduring levels of high blood pressure, and initiate and hasten the development of atherosclerosis in the coronary and carotid arteries. Atherosclerosis in these sites increases the risk of coronary heart disease and stroke, respectively. Environmental stressors and the brief physiologic changes they evoke can also precipitate acute manifestations of cardiovascular disease (e.g., temporary reductions in oxygen supply to the heart muscle) among individuals with pre-existing atherosclerosis (Rozanski et al. 1999). Further, psychosocial risk factors identified in epidemiological studies as risk factors for cardiovascular disease tend to be related to this psychophysiological mechanism. For example, social support generally reduces the magnitude of these stress-induced physiological reactivities (Uchino et al. 1996), and personality characteristics associated with increased risk of cardiovascular disease such as hostility are associated with more pronounced cardiovascular and neuroendocrine reactivity (Smith and Gallo 2001). Additional evidence of this mechanism comes from nonhuman, primate models of environmental stress, individual differences in social behavior, and the development of atherosclerosis. For example, subjected to the chronic stress of being housed with new cage-mates every few weeks, socially dominant but not subordinate male monkeys develop more severe atherosclerosis than when they are left in stable, less stressful housing conditions. Presumably, the recurring threats to status among dominant animals and the effort required to reassert status evoke physiological reactivity. Further, this effect can be prevented when the input of the sympathetic nervous system to the heart is blocked pharmacologically (Manuck et al. 1995). That is, reduction of the usual cardiovascular 'fight or flight' response through pharmacologic blockade prevents the otherwise unhealthy effect of psychosocial factors on atherosclerosis. Such findings provide the outlines of an answer to age-old questions about the influence of the mind on the body, and suggest that stress reduction could be useful in the treatment of cardiovascular disease. Interventions designed to reduce psychological stress have been found to produce beneficial effects on the health of patients with coronary heart disease (Rozanski et al. 1999).

A variety of illnesses from the common cold to some forms of cancer are influenced by the immune system, as is the rate at which wounds heal. The immune system is a complex set of structures and processes, the overall function of which is to identify and destroy foreign bodies (e.g., viruses) and aberrant cells (e.g., malignancies), and to facilitate repair of damaged tissues. Reductions in immune system functioning render the individual more susceptible to conditions it otherwise controls. A growing body of evidence suggests that environmental stress and factors that

influence the individual's susceptibility to stress impact immune functioning (Rabin 1999). Factors as diverse as the stress of final examinations to marital discord have been found to reduce immune functioning. This interdisciplinary field, psychoneuroimmunology (PNI), is the second area in which the study of mind-body connections is quite advanced. Nearly all of the psychosocial variables identified in epidemiological research as risk factors for disease have been found to be related to immune functioning (Kiecolt-Glaser and Glaser 1995).

Health behaviors might be responsible for some effects of stress on immune functioning. Smoking, inactivity, poor diet, and even sleep deprivation can produce deficits in immune functioning, and are therefore plausible links between stress and immune functioning. However, as in the case of research on cardiovascular disease, statistical control of health behaviors does not eliminate the effect of life stress and chronic negative emotions on immune functioning. As described above, environmental stress and negative emotions are associated with heightened circulating levels of several neuroendocrine factors such as cortisol and catecholamines. These neuroendocrine responses can disrupt immune functioning. It is not clear that the type and degree of immune disruption produced by stress is sufficient to account for the effects of such psychosocial risk factors on the development and course of even minor illnesses, let alone conditions as serious as cancer. Further, although some studies have demonstrated that stress-reducing interventions can have positive effects on immune functioning even among high-risk groups such as individuals with HIV, the evidence of substantial health benefits of interventions based on the PNI literature is preliminary. Nonetheless, the sophistication and clinical implications of this research area are growing rapidly.

4. Psychosocial Aspects of Medical Illness and Care

The third major topic in health psychology involves the psychosocial impact of acute and chronic medical illness, and the usefulness of psychological interventions as adjuncts to standard medical care. Some research is intended to identify the type and extent of psychosocial impacts of medical illness such as pain, disability, emotional distress, and disruption of personal relationships. Other work seeks to identify modifiable psychosocial influences on these outcomes (e.g., social support, coping styles, etc.), as well as psychosocial influences on the disease process itself (e.g., depression). Such studies can guide the development of psychosocial interventions. Although the patient is typically the focus of such interventions, some are directed at family members, caregivers, or even members of the healthcare team.

An early example of this topic was Irving Janus's psychological study of surgical patients. Based on patient interviews, Janus described a curvilinear association between the patient's anxiety before an operation and subsequent physical recovery. Those with moderate levels of presurgical anxiety experienced the least pain and the most rapid recovery compared to patients with very low or high levels. Janus argued that moderate anxiety facilitated the 'work of worry,' helping patients to develop realistic expectations and prepare to cope with later discomfort. In contrast, patients with high levels of preoperative anxiety were too distressed to engage in adaptive preparations, and patients with very low levels preoperative distress were not sufficiently motivated to prepare. Decades of subsequent research have indicated that presurgical anxiety is related to postoperative recovery in a simple linear manner (i.e., greater preoperative distress is associated with slower and more complicated recovery), providing the impetus for now routine psychological preparation of patients undergoing many potentially stressful medical procedures (Contrada et al. 1994). The provision of social support and information, hypnosis and relaxation therapies, and rehearsal of distraction and other coping strategies have been found to be effective in reducing emotional and physical discomfort, reducing medical complications, and accelerating wound healing and recovery after invasive diagnostic procedures, chemotherapy, childbirth, and even open heart surgery (Kiecolt-Glaser et al. 1998). In the case of young children, interventions intended to reduce parents' distress can facilitate the child's recovery, presumably by enhancing the parent's ability to comfort the child. These interventions are cost effective; interventions as brief as 30 minutes can reduce the length of hospitalizations by a day or more (Devine 1992). Research in this area suggests it is important to match interventions and the patient's coping style. Additional information is useful for patients who prefer involvement in their care, whereas training in distraction is best for patients who manage stress with avoidance.

Chronic illnesses such as coronary heart disease, cancer, and diabetes create many burdens for patients. These conditions pose the threat of further limitations and early death, often are associated with discomfort, disrupt social and vocational functioning, and involve unpleasant medical interventions and demands for patient behavior change. These threats to quality of life are in and of themselves important targets for health psychology interventions. However, emotional distress and poor adherence to prescribed medical regimens can contribute directly to poor medical outcomes. Interventions to facilitate coping with chronic disease, managing unpleasant symptoms, and making necessary changes in health behavior can improve functioning among these individuals and in some cases improve prognoses (Compas et al. 1998, Dusseldorp et al. 1999). For some diseases (e.g.,

essential hypertension, diabetes), behavioral interventions for changes in diet, increased physical activity levels, weight loss, and adherence to medical regimens are particularly effective, as these patient behaviors are key components of medical management. The patient's success in implementing and maintaining these behavior changes is a central determinant of health (see *Chronic Illness, Psychosocial Coping with*).

In other conditions (e.g., arthritis, headache, chronic back pain), psychological interventions are useful additions to standard care. Stress management, training in pain coping techniques, reinforcement for increasing levels of physical activity, and interventions designed to improve adaptive communication in couples and families have been found effective for reducing pain and reliance on pain medication, and in increasing functional activity levels (Compas et al. 1998). HIV/AIDS poses nearly all of these challenges in that patients are faced with a life-threatening and sometimes disabling disease, painful and distressing symptoms, and the necessity of adherence to complex medical regimens. Behavioral interventions to reduce stress, control unpleasant symptoms and treatment side effects, and improve adherence are valuable additions to HIV/AIDS care (Sikkema and Kelly 1996).

5. Emerging Challenges and Opportunities in Health Psychology

The rapid growth and evolution of health psychology seems likely to continue. Developments in basic and applied research in each area of the field have included many contributions to our understanding of health and illness, and to efforts to prevent and manage illness. Three factors are likely to shape health psychology in years to come. First, rising costs of medical care will continue to encourage the development of effective methods to prevent illness, maximize the benefits of care and limit its costs, and reduce the economic impact of illness. Although psychological interventions face growing competition from traditional medical services in a tightening climate in healthcare financing, effective interventions of any type will be well received.

Second, all areas of medical science and practice, including health psychology and behavioral medicine, must increase attention to the nature and management of health problems across diverse segments of the population. The growing concern with women's health and the health of ethnic minorities are obvious examples of this issue. Behavioral influences on health, the impact of illness on psychosocial functioning, and optimal intervention strategies in prevention and management are likely to vary across these groups. More recently, it has become clear that socioeconomic status is also a critical consideration. Within indus-

trialized nations, socioeconomic status is related to health and life expectancy. Health behavior, environmental stress, and access to medical care are important factors in this association (Adler et al. 1993). Hence, research and applied agendas in health psychology are likely to shape and be shaped by the emerging emphasis on the socioeconomic gradient in health. Of course, behavioral issues are no less important in the health of poorer nations, and an emerging 'world health' perspective will identify new roles for health psychology.

Finally, advances in basic and applied medical science will continue to shape health psychology. For example, developments in medical imaging have made it possible to study the impact of stress and other psychological factors on disease in more definitive ways. The revolution in medical genetics is making it possible to study gene behavior and gene environment interactive effects on disease more precisely, and may create the opportunity to focus behavioral risk reduction on those at greatest risk. The development of the artificial kidney and organ transplantation procedures created the need to understand the psychosocial aspects of new medical technologies. Similarly, advances in genetic testing to assess risk of illnesses such as cancer have created a new area of psychosocial research. The last three decades of theory, research, and practice in health psychology have produced remarkable growth and innovation. Although the future is difficult to predict, it appears to hold even more progress and impact for this recent variety of social and behavioral science.

See also: Adolescent Health and Health Behaviors; Behavioral Medicine; Biopsychology and Health; Health and Illness: Mental Representations in Different Cultures; Health Behavior: Psychosocial Theories; Health Behaviors; Health Behaviors, Assessment of; Illness Behavior and Care Seeking; Pain, Health Psychology of; Public Health

Bibliography

- Adler N, Boyce W T, Chesney M A, Folkman S, Syme S L 1993 Socioeconomic inequalities in health. *Journal of the American Medical Association* **269**: 3140–5
- Adler N, Matthews K 1994 Health psychology: Why do some people get sick and some stay well? *Annual Review of Psychology* **45**: 229–59
- Armitage C J, Conner M 2000 Social cognition models and health behaviour: A structured review. *Psychology and Health* **15**: 173–89
- Compas B E, Haaga D A, Keefe F J, Leitenberg H, Williams D A 1998 Sampling of empirically supported psychological treatments from health psychology: Smoking, chronic pain, cancer, and bulimia nervosa. *Journal of Consulting and Clinical Psychology* **66**: 89–112
- Contrada R, Leventhal E, Anderson J 1994 Psychological preparation for surgery: Marshaling individual and social

- resources to optimize self-regulation. In: Maes S, Leventhal H, Johnston M (eds.) *International Review of Health Psychology*. Wiley, New York/Chichester, UK, Vol. 3, pp. 219–66
- Devine E 1992 Effects of psychoeducational care for adult surgical patients: A meta-analysis of 191 studies. *Patient Education and Counseling* **19**: 129–42
- Dusseldorp E, van Elderen T, Maes S, Meulman J, Kraaij V 1999 A meta-analysis of psychoeducational programs for coronary heart disease patients. *Health Psychology* **18**: 506–19
- Engel G L 1977 The need for a new medical model: Challenge for biomedicine. *Science* **196**: 129–36
- Hill J O, Peters J C 1998 Environmental contributions to the obesity epidemic. *Science* **280**: 1371–4
- House J S, Landis K R, Umberson D 1988 Social relationships and health. *Science* **241**: 540–5
- Kiecolt-Glaser J K, Glaser R 1995 Psychoneuroimmunology and health consequences: Data and shared mechanisms. *Psychosomatic Medicine* **57**: 269–74
- Kiecolt-Glaser J K, Page G G, Marucha P T, MacCallum R C, Glaser R 1998 Psychological influences on surgical recovery: Perspectives from psychoneuroimmunology. *American Psychologist* **53**: 1209–18
- Manuck S B, Marsland A L, Kaplan J R, Williams J 1995 The pathogenicity of behavior and its neuroendocrine mediation: An example from coronary artery disease. *Psychosomatic Medicine* **57**: 275–83
- Rabin B S 1999 *Stress, Immune Function, and Health: The Connection*. Wiley-Liss & Sons, New York
- Rozanski A, Blumenthal J A, Kaplan J R 1999 Impact of psychological factors on the pathogenesis of cardiovascular disease and implications for therapy. *Circulation* **99**: 2192–217
- Schnall P L, Landsbergis P A, Baker D 1994 Job strain and cardiovascular disease. *Annual Review of Public Health* **15**: 381–411
- Schulz R, Beach S R 1999 Caregiving as a risk factor for mortality: The caregiver health effects study. *Journal of the American Medical Association* **282**: 2215–19
- Schwartz G, Weiss S 1978 Yale Conference on Behavioral Medicine: A proposed definition and statement of goals. *Journal of Behavioral Medicine* **1**: 3–11
- Schwarzer R, Gutiérrez B 2000 Health psychology. In: Pawlik K, Rosenzweig M R (eds.) *The International Handbook of Psychology*. Sage, London, pp. 452–65
- Sikkema K J, Kelly J A 1996 Behavioral medicine interventions can improve the quality of life and health of persons with HIV disease. *Annals of Behavioral Medicine* **18**: 40–8
- Smith T W, Gallo L C 2001 Personality traits as risk factors for physical illness. In: Baum A, Revenson T, Singer J (eds.) *Handbook of Health Psychology*. Erlbaum, Mahwah, NJ, pp. 139–72
- Stone G, Cohen F, Adler N 1979 *Health Psychology*. Jossey-Bass, San Francisco
- Uchino B N, Cacioppo J T, Kiecolt-Glaser J K 1996 The relationship between social support and physiological processes: A review with emphasis on underlying mechanisms and implications for health. *Psychological Bulletin* **119**: 488–531
- Weinstein N D, Rothman A J, Sutton S R 1998 Stage theories of health behavior: Conceptual and methodological issues. *Health Psychology* **17**: 290–9

T. W. Smith

Health Research, Qualitative

Qualitative health research is social research that uses qualitative methods to study problems in health, illness, and medicine. These methods include participant observation, in-depth interviewing, document analysis, and focus groups. The data collected are qualitative in nature including fieldnotes, open-ended interviews, narratives, text, organizational documents, and news and media excerpts. The data are usually in text form, but also may be visual or pictorial, and are analyzed by a variety of strategies (see Miles and Huberman 1994, Strauss 1987).

Qualitative health research is distinguished from quantitative and clinical research. Quantitative health research typically utilizes some kind of numerical data (e.g., from censuses, surveys, or experiments) and relies on statistical techniques for analysis. Clinical research is usually based on patient populations and is conducted as an adjunct to the delivery of medical care, focusing on clinical (e.g., treatment) problems. Some clinical research may use qualitative data, but typically qualitative health research does not focus on clinical or medically defined problems.

1. Background and Context

The roots of modern qualitative research can be found in early twentieth century social science (Vidich and Lyman 1994). While anthropologists had used ethnographic field work to study nonliterate and tribal cultures, sociologists at the University of Chicago in the 1920s pioneered using participant observation studies for examining aspects of urban and industrial societies (Bulmer 1984). Very little of this early fieldwork focused on health or medicine, although the classic study of *Middletown* did devote a chapter to it (Lynd and Lynd [1929] 1956).

During the 1950s and 1960s a new generation of University of Chicago trained sociologists including Everett Hughes, Anselm Strauss, Howard S. Becker, Eliot Freidson, Julius Roth, and Fred Davis, among others, laid the groundwork for the modern development of qualitative health research (Charmaz and Olesen 1997). This research was characterized by ethnographic approaches and the use of qualitative data, especially from interviews and observation, focusing on the subjective and interactional aspects of social life.

Early studies focused on medical education, organization of health care, medical institutions (especially hospitals), doctor–patient interaction, and death and dying. Several of these studies, especially by Becker et al. (1961) and Glaser and Strauss (1967) contributed significantly to the development of more rigorous and systematic data gathering and analytic methods.

Others like Goffman's (1961) *Asylums* became classics in the field.

In the latter years of the twentieth century qualitative health research prospered and continued to evolve. While social scientific research on health and medicine was dominated by quantitative studies, qualitative studies first found an intellectual niche and later flourished.

While sociologists pioneered qualitative health research, other disciplines adapted it to their own research needs. Anthropologists, who had long used qualitative methods, began to focus more research on modern societies and conducted studies on ethnicity, illness, and health care in community settings (Good and Good 2000). Qualitative studies permeated nursing research, in part because of the development of a nursing Ph.D. program at the University of California, San Francisco by Anselm Strauss and his colleagues. Feminist researchers have used qualitative methods for research on gender and health, reproductive issues, care-giving and care-receiving (Olesen 1994). Beginning in the late 1960s, British academics adopted qualitative methods to conduct research in health and medicine. In ensuing years it was adopted by researchers in France, Australia, Germany and other countries (Anderson and Bury 1988).

Qualitative methods are now well established in social scientific health research. Any number of useful primers for qualitative research now exist (e.g., Lofland and Lofland 1995), as well as compendiums for more advanced researchers (e.g., Denzin and Lincoln 1994, Hammersley and Atkinson 1995). Several journals feature qualitative studies, especially *Sociology of Health and Illness*, *Medical Anthropology Quarterly*, *Health*, and *Qualitative Health Research*. Dozens of health-related books based on qualitative data from researchers in several fields appear each year.

2. Types of Qualitative Methods

2.1 Participant Observation

Participant observation is the classic form of qualitative research. In this mode, the researchers collect observational data by observation in a situation or organization, after gaining access to their research sites. The balance between participation and observation varies, depending on the researcher and the site. The goal of the research is to understand the situation from the perspective of the participants and gaining this type of social scientific empathy can be a lengthy process. Participant-observers take fieldnotes based on their observations, which become the core data and the basis of their analysis and report. Observation is often combined with informal and formal in-depth interviewing to broaden data sources.

Participant observation is very labor- and time-intensive, but allows a close-up view of social life. In health research it has been used with great benefit for

the study of medical education (Becker et al. 1961), the delivery of health care (Zussman 1992), uses of medical technology (Timmermans 1999), professional-patient interaction (Waitzkin 1991), patient and self-help groups, alternative health care, and interaction among professionals (Bosk 1979).

2.2 In-depth Interviewing

In-depth interviewing is a central method of qualitative data collection, characterized by the researcher asking respondents a series of open-ended questions or raising topics for discussion. This is contrasted to survey interviewing, where researchers use mostly fixed-response questions (e.g., yes or no, multiple choice, Likert scale questions). Many researchers see in-depth interviewing as a 'guided conversation,' lasting perhaps from 15 minutes to many hours. While some use a fixed list of questions, others prefer to have an interview guide of topics, and then allow respondents to present their view or 'tell their story.' In either case, the goal here is to obtain narrative or textual data in the respondent's own words. Some studies are longitudinal and conduct several interviews with the same subjects of a period of time. Researchers often use tape recorders or similar devices and then transcribe the tapes into written textual data.

Researchers need to be careful to construct their interview guides so as not to unnecessarily bias the data. Questions should be asked in as neutral a form as possible and interviews should take place in a situation that is comfortable for the researcher and the respondent. One of the benefits of this kind of interviewing is that it allows the researcher to be sensitive to respondents and adjust the interview to new paths in the research inquiry.

The number of respondents in such studies varies greatly. Some studies use just a few selected respondents and interview them in great detail and depth. At the other end of the spectrum, some studies use a randomly selected sample of hundreds of respondents, each of whom is interviewed (in this case, usually by members of a research team). Most frequently, in-depth interview samples range from 20 to 100 respondents, depending on research question and the researchers' resources.

In-depth interviews have been used to study experience of illness and disability; patient and/or family view of treatment or care giving; health care providers' views of their work; the development and dilemmas in health careers; biographical perspectives on health and illness; and medical issues like adherence, seeking care, use of services, and quality of life (e.g., Conrad 1985, Kleinman 1988, Schneider and Conrad 1983).

2.3 Focus Groups

Since the early 1980s, focus groups have come into fashion as a method for collecting qualitative data,

adapted from marketing research. Focus groups are a collection of people who meet together with the investigator to discuss some topic or issue. Depending on the research topic, the members of the focus group may be already acquainted or unknown to each other, may or may not have direct experience with the topic at hand, and the groups may be of different sizes and composition. Typically, focus groups are small (five to 10 members) and meet from a single to a few occasions with the researcher. The researcher proposes the question or topic and then lets the group discuss it; depending on varying strategies, the leader may guide groups' discussion toward the topic of interest or let the group take its own course. The conversation is usually tape-recorded and then transcribed.

One advantage of focus groups is that they are a very economical method of collecting data. Several well-designed focus groups meeting for a couple of hours each can produce interactive data representing dozens of people. Compared to interviews, focus groups allow for considerable interaction among participants and thus resemble a multifaceted conversation. However, focus groups are artificial occasions, created by the researcher, and thus may not very well reflect everyday life, tend to remain on a superficial level, and are constrained by the group experience.

Focus groups are useful for discovering what a range of views are on different issues, what people's reactions are to certain topics or stimuli (e.g., type of drug or treatment); how people discuss health issues collectively; and to get a sense of the experience related to an illness, care-giving situation, or some other health-related matter (e.g., Kerr et al. 1998).

2.4 Documents

Some qualitative health data comes from sources other than observation, interviews, and focus groups. Most of these data come in the form of some type of document. This can range from governmental or 'official' documents, to organizational memos, to patient records, to news media stories, to personal letters and diaries. They can include pictorial or visual documents but are usually in some form of text. Such documents are ubiquitous to our modern world and are common in health settings and related to health issues.

Document analysis may be used as an adjunct to other methods such as participant-observation or may be the central data of a research inquiry. In some studies the data are collected very systematically, with a rigorous sampling strategy and a tight coding scheme; in other studies, often with limited and scarce data, the data collection is more of a convenience sample meant to illustrate some empirical or conceptual point. In most cases, for textual materials, one of several forms of qualitative content analysis is used to analyze the data. Often the documents can be

analyzed with the same strategies used for other qualitative data.

One advantage of documents is that they already exist and are available if one can get access to them. However, documents were created for other purposes than research and thus may not contain all the information an investigator requires or may represent more the view of the producer than reflect the reality of the situation.

Document analysis has been used to examine the emergence of medical conceptions in the professional literature, the changes in official statements about an illness or treatment, the news media's coverage or depiction of a health problem, an organization's statements about operation or purpose, or the content and presentation in patient records.

3. Qualitative Analysis

Qualitative analysis tends to be inductive rather than deductive. Instead of working from specific hypotheses and with predetermined codes, the researcher codes the data while or after it has been collected. Coding schemes can often be very complicated but are key to the data analysis (see Lofland and Lofland 1995, Strauss 1987).

Once the data are coded, researchers use qualitative analytic strategies such as 'grounded theory' (Glaser and Strauss 1967) to develop a substantive and a conceptual analysis. Grounded theory is a strategy that uses elaborate coding and continuous comparison to develop the analysis, allowing the categories to 'emerge' from close reading and analysis. An important aspect of qualitative analysis is the development of concepts, abstractions from the data that can be used to generalize the findings to other situations or cases. Examples are Glaser and Strauss's (1964) concept of 'awareness contexts' which they developed to understand the different situations and stances around information about dying, and Conrad's (1985) 'self-regulation' as an alternative to the commonly held notion of 'noncompliance.' Qualitative studies rarely claim to develop entire theories but at their best create insightful and useful conceptual understanding of aspects of the empirical world.

The best qualitative health studies enlighten the substantive issue at hand by the application or development of a conceptual analysis. If qualitative studies are to be case studies, the researchers must ask, what is it a case of? This will necessitate a conceptual or theoretical abstraction from the particulars of the case.

4. Contributions of Qualitative Health Research

Qualitative health studies have made important contributions to our knowledge. By insisting that researchers stay close to the data in their analysis and by

emphasizing first-hand data collection, qualitative studies have enlightened areas around health and health care. Qualitative studies are well suited to examining the social context of health care, organizational culture (Anspach 1993), the impacts of technology on patients (Timmermans 1999, Baszanger 1998), bioethics in medical care (Bosk 2000), the experience of illness (Conrad 1987, Anderson and Bury 1988), the functioning of medical organizations (Zussman 1992), and the process of medical education (see Hafferty 2000). Many important studies on patient experience and professional–patient interaction are qualitative (see Conrad 1987, Waitzkin 2000).

For many health issues, qualitative studies provide a unique ‘insider’s approach’ to studying a problem. In their review of ethnographic studies in medical sociology, Charmaz and Olesen (1997) conclude that qualitative studies have contributed in six ways: (a) maintaining a critical focus on the institution of medicine; (b) bringing subjective experience into medical sociology; (c) providing empirical corrections to public, professional, and sociological assumptions about medical care and health and illness; (d) challenging accepted ideas in subfields of medical sociology; (e) deepening other social scientific analyses; and (f) extending the range of medical sociology.

Qualitative research usually stands on its own but can also be used in conjunction with quantitative research; typically, either the quantitative or qualitative research predominates. For example, quantitative research can be used to set the context of a qualitative study or qualitative data can be used as cases to illustrate quantitative findings. True integration of qualitative and quantitative data is achievable although difficult (e.g., Broadhead and Heckathorn 1994).

Qualitative health research is most often of the ethnographic or analytically descriptive form described here. But other varieties have developed, including conversational analysis (e.g., Frankel 1984), narrative research (Kleinman 1988) and biographical reflection (Zola 1982). While these are not discussed separately here, they have also made important contributions to health research.

5. Critiques of Qualitative Health Research

There are several critiques of general qualitative research that also apply to qualitative health research. Since most such studies are of specific sites or nonrandom samples, the question is often raised about how generalizable findings are from the case to the larger population. Generalizability in this sense of sampling is a concept derived from quantitative studies that is often misapplied to qualitative studies. Most qualitative studies attempt to examine social processes that can be revealed by a close investigation and analysis, and it is the processes, not the site or sample, that may be generalizable. Another critique, which

frequently is a self-critique, is that qualitative researchers rarely tend to study ‘up.’ That is, investigators tend to study on their own social level or below, but rarely study the ‘seats of power.’ In health, that means we have many studies of patients, doctors, and health service organizations, but few studies of health insurance organizations, the medical industries, or the high level policy makers. This is in part a problem of access, and, it can be said, true of quantitative social scientific studies as well.

In the late 1980s, a number of scholars advanced the so-called postmodern critique of ethnography, calling into question the authority of the research and researcher and creating what some have called ‘the crisis of representation’ of qualitative research (see Denzin and Lincoln 1994). These critics argued that ethnographies are stories (even ‘fictions’) and narratives rather than authoritative representations of reality. The postmodernist emphasis on the validity of multiple perspectives rather than privileging the authors’ authority and singular truth raises significant questions about data collection, presentation, and validity of research. Some critics even suggest social scientists should abandon the pretense that ethnography adequately represents the ‘real world.’ Postmodernism’s disenchantment with ‘realism’ is not evident in most qualitative health research. Most of the research data are still realist tales (Van Maanen 1988), albeit more reflexive and circumspect than the qualitative research of earlier decades. Qualitative research remains accepted as a social scientific method of investigating social reality.

Having survived challenges from quantitatively oriented disciplines and criticisms from postmodernists, social scientific qualitative health research is by now well-established. New data sources like the Internet, an increasing number of computer-based analytic packages, and adoption of the method by more academic disciplines should allow for continued expansion and development of qualitative health research in the twenty-first century.

See also: Archival Methods; Biographical Methodology; Psychological Perspectives; Case Study: Methods and Analysis; Clinical Assessment: Interview Methods; Clinical Treatment Outcome Research: Control and Comparison Groups; Ethical Dilemmas: Research and Treatment Priorities; Experimenter and Subject Artifacts: Methodology; Field Observational Research in Anthropology and Sociology; Interpretive Methods: Macromethods; Interpretive Methods: Micromethods; Laboratory Experiment: Methodology; Narratives and Accounts, in the Social and Behavioral Sciences; Objectivity of Research: Ethical Aspects; Person-centered Research; Qualitative Methods, History of; Research Subjects, Informed and Implied Consent of; Single-subject Designs: Methodology; Social Survey, History of

Bibliography

- Anderson R, Bury M (eds.) 1988 *Living with Chronic Illness: The Experience of Patients and Their Families*. Unwin Hyman, London
- Anspach R 1993 *Deciding Who Lives*. University of California Press, Berkeley, CA
- Baszanger I 1998 *Inventing Pain Medicine: From the Laboratory to the Clinic*. Rutgers University Press, New Brunswick, NJ
- Becker H S, Geer B, Hughes E C, Strauss A 1961 *Boys in White*. University of Chicago Press, Chicago
- Bosk C L 1979 *Forgive and Remember: Managing Medical Error*. University of Chicago Press, Chicago
- Bosk C L 2000 The sociological imagination and bioethics. In: Bird C, Conrad P, Fremont A (eds.) *Handbook of Medical Sociology*, 5th edn. Prentice-Hall, Upper Saddle River, NJ, pp. 398–410
- Broadhead R S, Heckathorn D D 1994 AIDS prevention outreach among injection drug users: Agency problems and new approaches. *Social Problems* 41: 473–95
- Bulmer M 1984 *The Chicago School of Sociology*. University of Chicago Press, Chicago
- Charmaz K, Olesen V 1997 Ethnographic research in medical sociology: Its foci and distinctive contributions. *Sociological Methods and Research* 25: 452–94
- Conrad P 1985 The meaning of medications: Another look at compliance. *Social Science and Medicine* 20: 29–37
- Conrad P 1987 The experience of illness: Recent and new directions. In: Roth J, Conrad P (eds.) *Research in the Sociology of Health Care*. JAI Press, Greenwich, CT, Vol. 6, pp. 1–32
- Denzin N K, Lincoln Y S (eds.) 1994 *Handbook of Qualitative Research*. Sage Publications, Thousand Oaks, CA
- Frankel R M 1984 Physicians and patients in social interaction: Medical encounters as a discursive process. *Discourse Processes* 7: 103–5
- Glaser B G, Strauss A L 1964 Awareness contexts and social interaction. *American Sociological Review* 29: 269–79
- Glaser B G, Strauss A L 1967 *The Discovery of Grounded Theory*. Aldine, Chicago
- Goffman E 1961 *Asylums*. Doubleday Anchor, New York
- Good M D, Good B 2000 'Parallel sisters': Medical anthropology and medical sociology. In: Bird C, Conrad P, Fremont A (eds.) *Handbook of Medical Sociology*, 5th edn. Prentice-Hall, Upper Saddle River, NJ, pp. 377–88
- Hafferty F 2000 Reconfiguring the sociology of medical education: Emerging topics and pressing issues. In: Bird C, Conrad P, Fremont A (eds.) *Handbook of Medical Sociology*, 5th edn. Prentice-Hall, Upper Saddle River, NJ, pp. 238–57
- Hammersley M, Atkinson P 1995 *Ethnography: Principles and Practice*, 2nd edn. Routledge, London
- Kerr A, Cunningham-Burley S, Amos A 1998 Drawing the line: An analysis of lay people's discussions of the new genetics. *Public Understanding of Science* 7: 113–33
- Kleinman A 1988 *Illness Narratives*. Basic Books, New York
- Lofland L, Lofland J 1995 *Analyzing Social Settings*. Wadsworth, Belmont, CA
- Lynd R S, Lynd H M 1956 [1929] *Middletown*. Harcourt Brace, New York
- Miles M B, Huberman A M 1994 *Qualitative Data Analysis: A Sourcebook of New Methods*, 2nd edn. Sage, Thousand Oaks, CA
- Olesen V 1994 Problematic bodies: Past, present and future. *Symbolic Interaction* 17: 231–7
- Schneider J W, Conrad P 1983 *Having Epilepsy: The Experience and Control of Illness*. Temple University Press, Philadelphia, PA
- Strauss A L 1987 *Qualitative Analysis*. Jossey-Bass, San Francisco
- Timmermans S 1999 *Sudden Death and the Myth of CPR*. Temple University Press, Philadelphia
- Van Maanan J 1988 *Tales from the Field*. University of Chicago Press, Chicago
- Vidich A J, Lyman S F 1994 Qualitative methods: Their history in sociology and anthropology. In: Denzin N K, Lincoln Y S (eds.) *Handbook of Qualitative Sociology and Research*. Sage, Thousand Oaks, CA, pp. 23–59
- Waitzkin H 1991 *The Politics of Medical Encounters*. Yale University Press, New Haven
- Waitzkin H 2000 Changing patient–physician relationships in the changing health-policy environment. In: Bird C, Conrad P, Fremont A (eds.) *Handbook of Medical Sociology*, 5th edn. Prentice-Hall, Upper Saddle River, NJ, pp. 271–83
- Zola I K 1982 *Missing Pieces*. Temple University Press, Philadelphia, PA
- Zussman R 1992 *Intensive Care*. University of Chicago Press, Chicago

P. Conrad

Health Risk Appraisal and Optimistic Bias

1. Importance of Health Risk Perceptions

Perceptions of health risks influence the actions people take to prevent illness and injury, the decisions they make to seek medical care, their choices among treatment options, and their adherence to treatment requirements. Public controversies over the risks that may be created by government and corporate actions are also familiar. During these controversies, scientists and members of the public frequently disagree, demonstrating that laypeople do not passively accept the risk judgments of experts.

These various reasons suggest that it is important to understand how risk perceptions are formed, how accurate they are, and how they influence decisions and behavior. Furthermore, as research continues to identify new risks and 'risk factors' (attributes of individuals or environments that alter a person's likelihood of harm), the need grows to explain this information clearly so that individuals can make better decisions about their health. Such education about risk is the goal of the specialty called 'risk communication.'

2. Defining 'Risk' and 'Risk Perceptions'

Before pursuing this topic it should be mentioned that the term 'risk' is used in several different ways. Sometimes 'risk' is used synonymously with the word 'hazard', denoting a specific undesirable outcome. In this sense, cancer, the common cold, and automobile accidents are all 'risks.' 'Risk' is also used as a synonym

for 'probability', referring to the probability (or likelihood) of the undesirable outcome. Thus, one can speak of the risk that a person will develop cancer. 'Risk' is frequently used in still another way, as a synonym for 'danger' or 'threat' (for example, 'How risky is smoking?'). The term 'risk perceptions', then, is a general term that refers to the various beliefs that people hold about health and safety hazards.

3. How do People Think about Risk?

3.1 Dimensions of Hazards

Hazards have many dimensions. According to classical theories of decision-making, the key attributes of a hazard are the probability that it will occur and the severity of the undesirable consequences should it occur.

Pioneering research by Lichtenstein et al. (1978) demonstrated that laypeople are also influenced by hazard attributes that are not so obviously important. When scientists are asked about the riskiness of different illnesses and technologies, their judgments strongly reflect the number of people who die annually from each one, thus reflecting both seriousness (deaths) and likelihood (number of events in a particular population). Lay judgments are only moderately related to the number of fatalities. In addition, laypeople respond to whether the risk is familiar or unfamiliar, manufactured or natural, voluntary or involuntary, and other factors. Summarizing the many attributes that influence risk perceptions, researchers have suggested that lay judgments of risk reflect two main dimensions, the extent to which the hazard is perceived to be known or unknown and the extent to which it is dreaded or not (Slovic 1987).

Leventhal and colleagues suggest that each person copes with illness threats in accord with their own 'common sense' model of that illness (see *Illness Behavior and Care Seeking*). These models involve five attributes: the identity of the risk, its timecourse (e.g., chronic vs. acute), its consequences, its causes, and its potential cure or control (Leventhal et al. 1992).

The preceding perspectives emphasize various hazard attributes. Another approach emphasizes the interconnections among the hazard properties. Beliefs about the cause of an illness, its prevalence, routes of exposure, and personal risk are all important, but it is also important to know how they are related to one another. Misunderstandings about the causes and prevalence of hazard, for example, may result in misperceptions about the likelihood of victimization. These different aspects of knowledge and belief are viewed as a 'mental model' that describes both the separate pieces and the connections between these pieces (Bostrom et al. 1992).

3.2 Hazard Probabilities

Much attention has been given to how people think about hazard probabilities. (Apparently researchers think that hazard severity is a simpler issue, for this topic has received little attention.) Although proportions (e.g., '0.25'), percentages (e.g., '25 percent'), and odds (e.g., '1 in 4') are the units used by scientists to measure probability, there is abundant evidence that laypeople have great difficulty understanding and using such information. For instance, people sometimes give a higher probability to the likelihood of dying of breast cancer than to the likelihood of developing breast cancer. Thus, the ability to recite the correct percentage, rate, or odds should not be used as the criterion to decide whether people truly understand the probability. Except for percentages given in weather forecasts, the public is rarely exposed to probability statistics, so it is not surprising that many people find numerical information about hazard probabilities unhelpful.

Some evidence indicates that people find verbal designations of probability (e.g., 'small chance of happening') to be easier to use and to better reflect their own views than numerical risk measures (Diefenbach and Weinstein 1993), but people vary greatly in the numbers they associate with particular verbal labels, making the conversion of numbers to words highly subjective. Expressions of probability in terms of odds (e.g., 1 in 50 chance) are especially disliked by laypeople.

Because it seems impossible to navigate an uncertain world without forming estimates of the likelihood that particular events will occur, we tend to assume that people must form and encode judgments of hazard probabilities. In fact, it is possible that people rarely do encode such information. Instead they may generate risk judgments from past experiences and from their own implicit theories as needed. It is also possible that people store in memory only very crude probability information (e.g., no chance, small chance, large chance, certain) and never use the finer distinctions that concern scientists and policy makers.

Because social comparison is a central feature of human existence, in everyday life people may find it more natural to make decisions on the basis of risk comparisons than on beliefs about absolute probability (the likelihood that an event will occur). Thus, a woman's interest in preventing heart disease may depend less on her beliefs about her absolute likelihood of a heart attack (e.g., 40 percent chance) than on her belief that her risk is greater than the risk of her friends. She may also compare one risk (e.g., heart disease) with another (e.g., cancer).

4. Goals of Risk Perceptions

Obviously, accurate risk perceptions can help people make appropriate decisions about actions to avoid

illness or injury. However, accuracy is only one of the goals guiding people as they formulate their risk judgments (Kunda 1990). Risk perception is not an unbiased appraisal of information, but an attempt to seek the most comforting view of one's personal vulnerability that fits within the bounds of the evidence. Beliefs about our vulnerability to harm can lead to positive or negative emotions (pleasure about successfully avoiding harm; anxiety about potential illness or disability). Perceptions of vulnerability can also affect our feelings of self-esteem (see *Vulnerability and Perceived Susceptibility, Psychology of*). Since we want to believe that we are intelligent, competent, and foresighted, to admit that our actions place us at greater risk than our peers would be a threat to our self-esteem. Thus, minimizing negative affect and maximizing self-esteem are also goals of risk perceptions, and these goals may be at variance with the goal to be accurate.

5. Optimistic Bias

This bias in risk perceptions is revealed especially clearly when people are asked to compare their risks with the risk of the rest of the population. Whatever they may believe about the risk of others, they tend to believe that their own risk is less (Weinstein 1987). For example, when people are asked how their risk of a particular illness compares with the risk of others of the same age and sex, many claim to have below average risk and few admit to having above average risk. Across a variety of hazards, it is typical to find 40–70 percent of a group asserting below average risk; another 30–50 percent saying that their risk is average; and less than 10 percent acknowledging that their risk is above average. Obviously, this distribution cannot be correct. The 'average' person has, by definition, an 'average' risk. So when the people who claim below average risk greatly outnumber those who say their risk is higher than average, implying that the average person is below average in risk, something must be wrong with their risk judgments.

Optimistic bias is not just denial, a blanket defense against anxiety. If it were, we would expect to find the greatest optimism among hazards, such as cancer, that generate the most fear. The amount of optimistic bias (that is, the extent to which people claiming below average risk outnumber those who say their risk is above average) does vary from hazard to hazard (Weinstein 1987), but it is not related to the severity of the hazard. Instead it is greatest for events that are believed to be preventable by individual action and events with which people have little personal experience. Bias is especially great if people think that vulnerability appears early, leading them to conclude that if they have not experienced the problem so far, they are free from risk.

The belief that we are less at risk than others is apparently reassuring, since people resist attempts to correct their misperceptions. Most interventions designed to eliminate bias in comparative risk judgments—for example, having people think of the ways in which the negative event could happen or think of the precautions that they have not taken—have proved unsuccessful (Weinstein and Klein 1995).

It may appear obvious that underestimating risk will expose people to unnecessary harm. However, a different view of this issue, known by the label 'positive illusions,' suggests that optimistic biases are really adaptive (Taylor and Brown, 1988). It points out that people who are overly optimistic about their ability to fight off the AIDS virus, for example, may be encouraged to maintain a healthy lifestyle and to bear with burdensome treatments. Thus, even if the perception of resistance is false, it may support other actions that are themselves beneficial. Although definitive evidence is lacking, it seems likely that underestimations of personal vulnerability decrease interest in prevention, especially when people see themselves as members of a group that does not need to be concerned about the risk, but optimism about surviving a life-threatening illness can help patients avoid depression and hopelessness (Wood et al. 1985).

6. Effects of Risk Perceptions on Health-related Behaviors

Perceived probability and/or perceived severity occupy an important role in nearly all theories that are used to explain the health-related behavior of individuals (see *Health Behavior: Psychosocial Theories*). The literature on health behavior generally supports the idea that increases in perceived risk increase precautionary behavior (Janz and Becker 1984), but there are also numerous instances in which risk perceptions and behavior turn out to be unrelated.

This should not be surprising. There are many factors other than risk perceptions that influence behavior, and people who think that they have a high risk may nevertheless doubt their ability to carry out precautions or the efficacy of the precautions. In addition, most health behavior studies have used correlational research designs, which are problematic for examining risk perception–risk behavior linkages. Finally, as mentioned earlier, experimental interventions may fail to alter peoples' perceptions of their invulnerability.

The link between perceptions and action may depend on the type of action. Precautions that prevent a health problem from occurring (such as vaccinations) should reduce both perceived risk and worry about that risk. In contrast, being tested for the presence of an existing health problem (as in cancer screening) offers the immediate potential of extremely distressing information, especially if the problem is

difficult to treat, so high perceived risk may be associated with high fear. The links between risk perceptions and different types of action still need to be elucidated.

See also: Attitudes and Behavior; Depression, Hopelessness, Optimism, and Health; HIV Risk Interventions; Personality and Risk Taking; Sexual Risk Behaviors

Bibliography

- Bostrom A, Fischhoff B, Morgan M G 1992 Characterizing mental models of hazardous processes: A methodology and an application to radon. *Journal of Social Issues* **48**: 85–100
- Diefenbach M A, Weinstein N D, Oreilly J 1993 Scales for assessing perceptions of health-hazard susceptibility. *Health Education Research* **8**: 181–92
- Janz N K, Becker M H 1984 The health belief model: A decade later. *Health Education Quarterly* **11**: 1–47
- Kunda Z 1990 The case for motivated reasoning. *Psychological Bulletin* **108**: 480–98
- Leventhal H, Diefenbach M A, Leventhal E A 1992 Illness cognition: Using common sense to understand treatment adherence and affect cognition interactions. *Cognitive Therapy and Research* **16**: 143–63
- Lichtenstein S, Slovic P, Fischhoff B, Layman M, Combs B 1978 Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory* **4**: 551–78
- Slovic P 1987 Perception of risk. *Science* **236**: 280–5
- Taylor S E, Brown J D 1988 Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin* **103**: 193–210
- Weinstein N D 1987 Unrealistic optimism about susceptibility to health problems: Conclusions from a community-wide sample. *Journal of Behavioral Medicine* **10**: 481–500
- Weinstein N D, Klein W M 1995 Resistance of personal risk perceptions to debiasing interventions. *Health Psychology* **14**: 132–40
- Wood J V, Taylor S E, Lichtman R R 1985 Social comparison in adjustment to breast cancer. *Journal of Personality and Social Psychology* **49**: 1169–83

N. D. Weinstein

Health: Self-regulation

Many health conditions are caused by risk behaviors such as problem drinking, substance use, smoking, reckless driving, overeating, or unprotected sexual intercourse. Fortunately, human beings have, in principle, control over their conduct. Health-compromising behaviors can be eliminated by self-regulatory efforts, and health-enhancing behaviors such as physical exercise, weight control, preventive nutrition, dental hygiene, condom use, or accident-preventive

actions can be adopted instead. Health self-regulation refers to the motivational, volitional, and actional process of removing such health-compromising behaviors in favor of adopting and maintaining health-enhancing behaviors (see *Health Behaviors*). Health self-regulation encompasses a broad range of social cognitions and behaviors that are subject to developmental changes. This article, however, is limited to models that address the current debate within the discipline of health psychology, excluding the larger fields of social and developmental psychology (Baltes 1997, Baumeister and Heatherton 1996, Brandtstädter and Greve 1994, Carver and Scheier 1998).

1. Health Motivation Factors

Before changing their habits, people need to become motivated. This is seen as a process towards an explicit intention (e.g., 'I intend to quit smoking this weekend'). Three variables are considered to play a major role in this process: (a) risk perception, (b) outcome expectancies, and (c) perceived self-efficacy (Bandura 1997, Carver and Scheier 1998, Conner and Norman 1996).

Perceiving a health threat seems to be the most obvious prerequisite for the motivation to remove a risk behavior. If one is not aware at all of the risky nature of one's actions, motivation cannot emerge. Usually, people are aware of some level of risk although the accuracy of their perception may be biased. When it comes to a comparison with similar others, one's view of the risk is somewhat distorted ('Compared to others of my age and sex my risk of getting lung cancer is low/medium/high'). This has become known as the 'optimistic bias' (see *Health Risk Appraisal and Optimistic Bias*). Nevertheless, persons also acknowledge some degree of risk when confronted with objective data. There is a realistic component that keeps the positive illusions in leash. For example, smokers not only know that smoking can cause adverse health in others, they also perceive that they are more at risk themselves for lung cancer and other diseases than nonsmokers.

Risk perception has two aspects: perceived severity of a health condition and personal vulnerability toward it (see *Vulnerability and Perceived Susceptibility, Psychology of*). The first refers to the amount of harm that could occur, and the second pertains to the subjective probability that one could fall victim to that condition. Thus, it has been recommended that people should be informed about the existence of a health risk, and moreover, that they should imagine themselves as possible victims if they do not take the necessary precautions.

Fear arousal is an emotion that can accompany risk perception. Scaring people into health behaviors, however, has not been shown to be effective. In general, initial risk perception seems to be advan-

tageous to put people on track for developing a motivation to change, but later on other variables are more influential in the self-regulation process, such as outcome expectancies, self-efficacy, and behavioral intentions.

People not only need to be aware of the existence of a health threat, they also need to know how to regulate their behavior by understanding the contingencies between their actions and subsequent outcomes. These *outcome expectancies* are influential beliefs in the motivation to change (Bandura 1997). A smoker may find more good reasons to quit than good reasons to continue smoking ('If I quit smoking then I will save money'). This imbalance does not lead directly to action but it can help to come up with the intention to quit. The pros and cons represent positive and negative outcome expectancies that are typical in rational decision making. However, such contingencies between actions and outcomes need not be explicitly worded and evaluated, they can also be rather diffuse mental representations, loaded with emotions. Outcome expectancies can also be understood as methods, or means-ends relationships, indicating that people know proper strategies to produce desired effects.

Moreover, the efficacy of such a method has to be distinguished from the belief in one's personal efficacy to apply the method. *Perceived self-efficacy* (Bandura 1997) portrays individuals' beliefs in their capabilities to exercise control over challenging demands and over their own functioning (see *Self-efficacy and Health*). Self-regulatory strength (Baumeister and Heatherton 1996) is a similar notion of one's capacity or stable resource to cope with demands. These beliefs are critical in approaching novel or difficult situations or in adopting a strenuous self-regimen. People make an internal causal attribution when imagining their behavior (e.g., 'I am certain that I can quit smoking even if my friend continues to smoke'). Such optimistic self-beliefs influence the goals people set for themselves, what courses of action they choose to pursue, how much effort they invest in given endeavors, and how long they persevere in the face of barriers and setbacks. Some people harbor self-doubts and cannot motivate themselves. They see little point in even setting a goal if they believe they do not have what it takes to succeed. Thus, the intention to change a habit that affects health is to some degree dependent on a firm belief in one's capability to exercise control over that habit (see *Control Beliefs: Health Perspectives*).

2. Self-regulatory Efforts

Changing one's health behavior is considered to be a difficult self-regulation process. After people have become engaged in a goal they need to prepare action and, later, maintain the changes in the face of obstacles and failures. Thus, goal setting and goal pursuit can be understood as two distinct processes, the latter of which requires a great deal of self-regulatory effort.

The process of self-regulation can be subdivided into sequences such as planning, initiation, maintenance, relapse management, and disengagement although these are not perfectly distinct categories. Plans specify the when, where, and how of a desired action, and carry the structure of 'When situation S arises, I will perform response R' (Gollwitzer 1999). People take *initiative* when the critical situation arises and give it a try. This requires firm self-beliefs to be capable to perform the action. People who do not hold such beliefs see little point in even trying.

The adoption and *maintenance* of the health behavior is not achieved through an act of will but involves the development of self-regulatory skills and strategies. This embraces various means to influence one's own motivation and behaviors such as the setting of attainable, proximal subgoals, creating incentives, drawing from an array of coping options, and mobilizing social support. Action control processes include focusing one's attention on the task, while avoiding attention to distractors, resisting temptations, and managing unpleasant emotions. Perceived self-efficacy is required to overcome barriers and to stimulate self-motivation repeatedly.

Competent *relapse management* is needed to recover from setbacks. Some people rapidly abandon their newly adopted behavior when they fail to get quick results. When entering high-risk situations (e.g., a bar where others smoke), they cannot resist due to a lack of self-efficacy. The competence to recover is different from the competence to initiate an action (Marlatt et al. 1995). Restoration, harm reduction, and renewal of motivation are serviceable strategies within the context of health self-regulation.

Disengagement from the goal can be evidence for lack of persistence and, thus, indication of self-regulatory failure (Carver and Scheier 1998). But in the case of repeated failures, disengagement or scaling back the goal might become an option which can be adaptive, depending on the circumstances. For example, if the goal was set too high or if the situation has changed and become more difficult than before, it is often not worthwhile to continue the struggle. In the case of health-compromising behaviors, giving up is not a tenable option. Better self-regulatory skills have to be developed and unique approaches to the problem need to be taken. The experience of failure can be a useful learning experience to build up more competence, on the condition that the individual makes a beneficial causal attribution of the episode, and practices constructive self-talk to renew the motivation.

3. Models of Health Behavior Change

In health psychology (see *Health Psychology*), models are built to describe, explain, and predict how people become motivated to change their risk behaviors, and how they become encouraged to adopt and main-

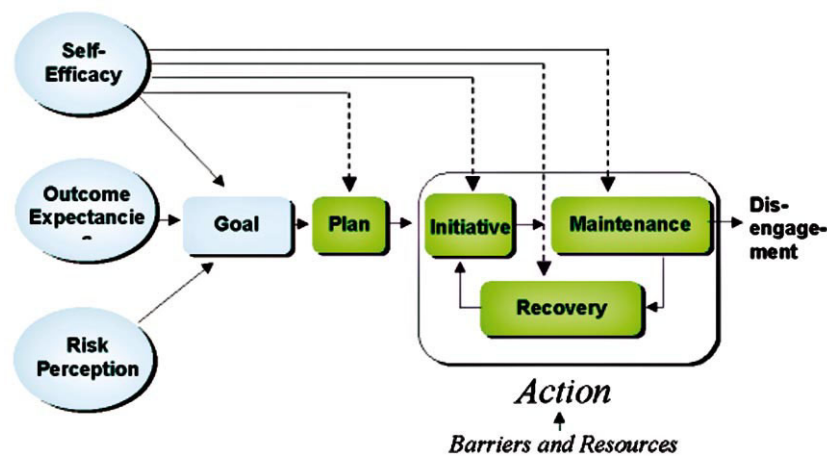


Figure 1

A social-cognitive process model of health self-regulation (Schwarzer 1992, 1999)

tain health actions (see *Health Behavior: Psychosocial Theories*).

3.1 The Transtheoretical Model of Behavior Change

Stage theorists have made an attempt to consider process characteristics by including the postintentional phase. The *Transtheoretical Model of Behavior Change* (DiClemente and Prochaska 1982), for example, has become the most popular stage model. Its main feature is the implication that different types of cognitions may be important at different stages of the health behavior change process. It includes five discrete stages of health behavior change that are defined in terms of one's past behavior and future plans (precontemplation, contemplation, preparation, action, maintenance). For example, at the precontemplation stage, a drinker does not intend to stop consuming alcohol in the future. At the contemplation stage, a drinker thinks about quitting sometime within the next six months, but does not make any specific plans for behavior change. At the preparation stage, the drinker resolves to quit within the next six months. The action stage includes individuals who have taken successful action for any period of time. If this abstinence has lasted for more than six months, the person is categorized as being in the maintenance stage.

The model, which also includes self-efficacy and other relevant features, has been applied successfully to a broad range of health behaviors, with particular success for smoking cessation. However, it has been argued that the notion of stages within this theory might be flawed or circular, in that the stages are not genuinely qualitative, but are rather arbitrary distinctions within a continuous process. Instead, these 'stages' might be better understood as 'process heur-

istics' to underscore the nature of the entire model. That is, the model can serve as a useful heuristic that describes a health behavior change process which has not been the major focus of health behavior theories so far. In redirecting the attention to a self-regulatory process, the transtheoretical model has served an important purpose for applied settings. The postintentional, preactional phase ('preparation') may be the most challenging stage for researchers and professionals because it is exactly in this phase where an intention is or is not translated into an action—depending on the circumstances. Planning and initiative, but also volatility, hesitation, and procrastination characterize this phase. The identification of stages has been shown to bear important implications for treatment. Interventions have been tailored to communicate stage-matched information to smokers who were in different motivational stages (Dijkstra et al. 1998).

3.2 The Health Action Process Approach

The *Health Action Process Approach* (HAPA model; Schwarzer 1992, 1999) suggests a distinction between preintentional motivation processes and postintentional volition processes. Within these two phases, different patterns of social-cognitive predictors may emerge (see Fig. 1).

In the initial *motivation phase*, a person develops an intention to act. In this phase, *risk perception* is merely seen as a distal antecedent within the motivation phase. It may include not only the perceived severity of possible health threats but also one's personal vulnerability to fall prey to them. Risk perception in itself is insufficient to enable a person to form an intention. Rather, it sets the stage for further elaboration of

thoughts about consequences and competencies. *Outcome expectancies* operate in concert with *perceived self-efficacy*, both of which contribute substantially to the development of an intention, whereas this pattern may change later in the postintentional stage (Schwarzer and Renner 2000).

After goal setting, people enter the postintentional stage or *self-regulation phase* in which they pursue their goal by planning the details, trying to act, investing effort and persistence, possibly fail, and finally recover or disengage. The guiding principle here is that each substage advances from the previous one and is, in addition, facilitated by perceived self-efficacy. Thus, at each point there are two predictors, namely the successful completion of the previous substage in conjunction with an optimistic sense of control over the next one. Risk perception and outcome expectancies are no longer of much influence when goals have been set and when it comes to predicting actual behaviors. Certainly, there may be various influential factors in the self-regulation phase, but according to the present knowledge, stage progression and self-beliefs appear as the most parsimonious and powerful set of predictors (Bandura 1997, Schwarzer and Renner 2000).

This health behavior change model is regarded as a heuristic to better understand the complex mechanisms that operate when people become motivated to change, when they adopt and maintain a habit, and when they attempt to resist temptations and recover from setbacks. It applies to all health-compromising and health-enhancing behaviors and could be extended to behavior change in general (see Conner and Norman 1996 and Weinstein 1993 for other models).

4. Future Directions

Health self-regulation encompasses a broad range of cognitions and behaviors. Research is needed to identify the key variables and the essential mechanisms that operate in this process. Certainly, different perspectives lead to an emphasis on different variables and mechanisms. A social psychological perspective would underscore social influence and the social context of self-regulation. A life-span developmental perspective would accentuate personal history, life events, as well as gains and losses. The present health psychology view pertains to rather short time periods that are typically involved in coping with addictive behaviors and in thorny cessation attempts. In the last two decades, social-cognitive prediction models had been preferred to explain the mechanisms of health self-regulation. These approaches are being replaced now by stage models that imply a cycling and recycling process of individuals across two or more stages. This advance in the field has led to some difficulties and controversies. One refers to the demand for parsimony because variables become easily inflated. For example,

the Transtheoretical model includes five stages and 10 processes. According to the present view, it is suggested to reduce the motivation stage to only three predictors and the self-regulation stage to only two components: stage progression and self-beliefs appear as the most parsimonious set of components. Further elaborations, however, are required and could benefit from other fields, in particular from relapse prevention theory (Marlatt et al. 1995) and self-regulation theories (Baltes 1997, Baumeister and Heatherton 1996, Brandtstädter and Greve 1994, Carver and Scheier 1998, Karoly 1993).

See also: Adolescent Health and Health Behaviors; Control Beliefs: Health Perspectives; Health Behavior: Psychosocial Theories; Health Behaviors; Health Behaviors, Assessment of; Self-efficacy; Self-efficacy and Health; Self-regulation in Adulthood; Smoking Prevention and Smoking Cessation

Bibliography

- Baltes P B 1997 On the incomplete architecture of human ontogeny: Selection, optimization, and compensation as foundation of development theory. *American Psychologist* **52**: 366–80
- Bandura A 1997 *Self-efficacy: The Exercise of Control*. Freeman, New York
- Baumeister R F, Heatherton T F 1996 Self-regulation failure: An overview. *Psychology Inquiry* **7**(1): 1–15
- Brandtstädter J, Greve W 1994 The aging self: Stabilizing and protective processes. *Developmental Review* **14**: 52–80
- Carver C S, Scheier M F 1998 *On the Self-regulation of Behavior*. Cambridge University Press, New York
- Conner M, Norman P (eds.) 1996 *Predicting Health Behavior: Research and Practice with Social Cognition Models*. Open University Press, Buckingham, UK
- DiClemente C C, Prochaska J O 1982 Self-change and therapy change of smoking behavior: A comparison of processes of change in cessation and maintenance. *Addictive Behaviors* **7**: 133–42
- Dijkstra A, De Vries H, Roijackers J, van Breukelen G 1998 Tailored interventions to communicate stage-matched information to smokers in different motivational stages. *Journal of Consulting and Clinical Psychology* **66**: 549–57
- Gollwitzer P M 1999 Implementation intentions: Strong effects of simple plans. *American Psychologist* **54**: 493–503
- Karoly P 1993 Mechanisms of self-regulation: A system view. *Annual Review of Psychology* **44**: 23–52
- Marlatt G A, Baer J S, Quigley L A 1995 Self-efficacy and addictive behavior. In: Bandura A (ed.) *Self-efficacy in Changing Societies*. Cambridge University Press, New York, pp. 289–315
- Schwarzer R 1992 Self-efficacy in the adoption and maintenance of health behaviors: Theoretical approaches and a new model. In: Schwarzer R (ed.) *Self-efficacy: Thought Control of Action*. Hemisphere, Washington, DC, pp. 217–42
- Schwarzer R 1999 Self-regulatory processes in the adoption and maintenance of health behaviors. The role of optimism, goals, and threats. *Journal of Health Psychology* **4**: 115–27

Schwarzer R, Renner B 2000 Social-cognitive predictors of health behavior: Action self-efficacy and coping self-efficacy. *Health Psychology* **19**(5): 487–95

Weinstein N D 1993 Testing four competing theories of health-protective behavior. *Health Psychology* **12**: 324–33

R. Schwarzer

Health Surveys

Information collected by health surveys is used by public health officials to plan, conduct, and evaluate programs; by elected officials to inform policies and legislation; and by researchers to understand the health status of the population, the determinants of health, and the health care system. While such surveys are sponsored and carried out by many private entities, core information is generally obtained through relatively large, multipurpose, government-sponsored data collection systems; these will be the focus here, although most of the issues discussed pertain to all health surveys.

The World Health Organization definition of health as a state of physical, mental, and social well-being and not merely the absence of disease or infirmity provides the starting point for health surveys. The multidimensionality of this definition requires different survey methodologies. Since health is both a biologic and a social construct, health surveys must reflect this duality. Major topics to be covered include aspects of survey design, conceptualization and measurement of health status, determinants of health, summary measures of health, health care utilization and health behaviors. The effect of recent technological advances on data collection methods are discussed as are the growing opportunities for data linkage and the future challenges of providing access to data while protecting confidentiality and privacy.

In the late 1990s, 9.2 percent of the United States population reported they were in fair or poor health; 14 percent of children under 18 years of age had no health insurance coverage; 79 percent of children 19–35 months old were up to date on immunizations; 25 percent of adults were smoking cigarettes; and 24 percent were hypertensive (National Center for Health Statistics, NCHS 1999). Such information is used by public health officials to plan, conduct and evaluate programs; by elected officials to inform policies and legislation; and by researchers to understand the health status of the population, the determinants of health and the health care system. Information on the health of a population can be obtained through a variety of data collection approaches; health surveys represent one mechanism.

While such surveys are sponsored and carried out by many private entities, core information is generally

obtained through relatively large, multipurpose, government-sponsored data collection systems; these will be the focus here, although most of the issues discussed pertain to all health surveys. These surveys are used to profile the health of the population, to track demographic shifts and trends, and to monitor change in health and health care. In addition, the data are also used to investigate epidemiologic relationships between risk factors and health outcomes and to provide information for prevention, evaluation, research, planning, and policy. Many nations include the conduct of general purpose health surveys in their statistical systems. Table 1 provides a list of countries and the title and year of their most recent general health survey (CDC, NCHS 1999). It is tempting to directly compare the results of these surveys, but cross-national cultural differences and variation in the nature of health care systems as well as in survey designs must be taken into account when making international comparisons.

1. Definition and Operationalization of Health

The World Health Organization definition of health as a state of physical, mental, and social well-being and not merely the absence of disease or infirmity provides the starting point for health surveys (WHO 1948). However, because this definition is broad, complex, and multidimensional, it is difficult for any single survey to collect all the needed information. Depending on the specific objectives of a particular survey, there will be variation in the universes from which the samples are drawn and in the health and health-related information obtained. Since health is both a biologic and a social construct, health surveys must reflect this duality. In fact, these inter-relationships between the biological and social aspects of health are often best approached through health surveys, and advances in technology have made it increasingly possible to incorporate direct measures of biologic aspects of health into more traditional surveys.

2. General Design Issues

Health surveys can either obtain information directly from individuals (the subject or an informant) or can rely on administrative records. A wide range of sampling methodologies, sampling frames, and data collection strategies are used depending on survey objectives and the characteristics of the area in which the survey is done. Countries that have universal health care with centralized administrative systems can utilize designs that would be inappropriate in countries with less well-developed or decentralized health care systems. The simplest sampling scheme is a simple random sample of a population obtained from a complete and up-to-date list such as from a census of

Table 1
National population-based health surveys

Country	Title and year
Australia	Australia Health Survey, 1983 National Health Survey, 1995
Austria	Special Health Survey, 1999
Canada	Canada Health Survey, 1978–79 National Population Health Survey, 1996–97
China (People's Republic of)	Health Services Household Survey of China, 1998
Czech Republic	Health Interview Survey, 1996
Denmark	The Danish Health and Morbidity Survey, 1994
Estonia	National Health Interview Survey, 1996
Finland	The Finnish National Health Survey, 1995 Mini-Finland Health Survey, 1977–81
France	France Health Survey, 1991–92 Health and Social Protection Survey, 1988
Germany	National Health Examination and Interview Survey, October 1997–March 1999
Hungary	Health Interview Survey, 1984
Iceland	Multipurpose Survey (surveys on health promotion), 1996
Ireland	The National Health and Lifestyle Survey, 1998
Italy	Survey of Health Conditions of the People and the Use of Health Services, 1994
Japan	Comprehensive Survey of Living Conditions of People on Health and Welfare, Triennial since 1986
Korea (Republic of)	National Health Interview Survey, 1983
Latvia	Living Conditions Survey, 1994
Netherlands	National Health Interview Survey, Continuous since 1981
New Zealand	New Zealand Health Survey, 1996–7
Norway	Health Interview Survey, 1998
Poland	Family Health Survey, 1987 Health Interview Survey, 1996
Portugal	National Health Survey, 1998–9
Singapore	National Health Survey, 1998
Spain	National Health Survey, 1993
Sweden	Survey of Living Conditions, Continuous since 1975
Switzerland	Swiss Health Survey (SOMIPOPS), 1981–2 Swiss Health Survey, 1997
Ukraine	National Health Survey, Annual since 1986
United Kingdom	Health Survey for England, Continuous since 1991 Scottish Health Survey, Triennial since 1995
United States	National Health Interview Survey (NHIS), Continuous since 1957 National Health and Nutrition Examination Survey (NHANES), Continuous since 1999

Source: National Center for Health Statistics, International Health Data Reference 1999.

the population or a centralized health care system. In the case of surveys of persons, if such lists do not exist, area sampling can be used with clustering to reduce costs. In order to be able to report information on subpopulations of interest as defined by geography and/or demographic, socioeconomic or health characteristics, designs often include oversampling of these groups. It is often necessary to incorporate screening into designs to achieve this objective. Information can be obtained either by mail, by phone, by in-person interview, or by some combination of modes depending on the ability of the researchers to contact the population using these means. The latter two modes can also incorporate computer-assisted technologies, and the use of the Internet for the administration of health surveys is being investigated.

In general, obtaining information directly from the survey subject provides more reliable and valid health data. However, this is not universally true and depends on the nature of the information sought and the characteristics of the subject (Moore 1988). For example, while accurate information on risk behaviors among adolescents can probably only be obtained from the adolescent in a private environment, adolescents are not good reporters of health care utilization or household income. Even more problematic is the situation where a health condition does not permit a subject to respond for him- or herself. Since eliminating the subject from the survey would seriously bias the results, a proxy respondent is often used.

While necessary, the use of proxy respondents does introduce a source of error into health surveys, particularly longitudinal surveys. Obtaining information on the relationship of the proxy to the subject as well as the conditions leading to the use of a proxy can reduce this error. In other cases, subject reporting is fraught with a great deal of measurement error. For example, information on the characteristics of an insurance plan is usually better obtained from the insurer rather than the person holding the insurance. Finally, reliable assessment of specific types of morbidity, risk factors, and mortality requires direct measurement through physical examination or access to administrative records from health care providers.

Many health surveys limit their population of interest. For example, it is common for health surveys to only include the noninstitutionalized population. The dramatic differences in the living conditions of the institutionalized and noninstitutionalized populations make it difficult to design survey methodologies that would apply in all situations. The military population is also often excluded from general health surveys. The scope of the universe needs to be clearly defined, especially if populations that differ in their health status are omitted—such as persons residing in nursing homes.

When obtaining information directly from the population is inefficient or would result in either poor data quality or high costs or both, an alternative is to

sample from existing records that were created for other purposes. Information on health care utilization for specific conditions is more easily obtained from hospital or provider records than from a population sample given the relative rarity of the phenomena of interest. The vital statistics system can also be considered a survey of administrative/legal records for the purpose of describing the health of the population. Administrative systems, in addition to providing basic health information, are often used as sampling frames for population based surveys. They are very attractive for this purpose as they eliminate the need to identify members of the population who have the characteristics of interest.

3. Content

Information collected on health surveys can be divided into three general types: (a) health status; (b) determinants or correlates of health including health behaviors; and (c) health care utilization. In addition, information is also collected on demographic and socioeconomic factors that affect health. For illustrative purposes, examples will be provided from two major health surveys conducted in the United States by the National Center for Health Statistics, CDC: The National Health Interview Survey (NHIS) (<http://www.cdc.gov/nchs/nhis>), and the National Health and Nutrition Examination Survey (NHANES) (<http://www.cdc.gov/nhanes>). Many statistical agencies work closely with counterparts in other nations in the design of their health surveys, thus, many observations regarding the NHIS and NHANES apply to similar surveys in other countries. Attempts to standardize various parts of health surveys have increased in recent years.

The NHIS is a principal source of information of the health of the civilian noninstitutionalized US population. Conducted annually since 1957, the NHIS currently collects information through computer assisted in-person interviews from approximately 40,000 households covering 100,000 people. Conducted about once a decade, NHANES is based on detailed physical examinations in addition to in-depth interviews. A nationally representative probability sample of the US civilian, noninstitutionalized population is interviewed at home and then invited to visit a Mobile Examination Center (MEC) for a standardized physical examination of about four hours, including a physician's exam, physical measurements, blood and urine tests, and nutritional assessments.

4. Overall Health Status

As noted above, health status is a multidimensional concept, requiring multiple indicators and multiple methodologies for adequate measurement. Several

different indicators of health status are usually included in health surveys, including single summarizing measures; questions relating to disease incidence and prevalence; and questions relating to functioning (physical, cognitive, emotional, and social). Health status measures also vary by whether they are based on objective information obtained from standardized examinations or medical records or from information obtained from the individual or a proxy. To be most effective, individual health surveys should capture a variety of aspects of health status so as to provide a more comprehensive and complete assessment of health status than would be possible from any single strategy. Measures of health status should also be constructed to be useful in epidemiologic analyses of risk factors as well as for monitoring trends.

The subject's self-assessment of his/her health status as excellent, very good, good, fair, or poor is a popular summary indicator of health status. This measure has been shown to be highly correlated with other measures of health status and is predictive of mortality and admission to long-term care facilities (Idler and Angel 1990). However, this measure can be problematic when used to monitor change over time. The means by which individuals evaluate various aspects of health have been shown to be affected by contextual parameters. Implicit in self-perceived health status is the individual's evaluation of his/her health status against some unstated standard. Societal norms act to define the standard, but the norms and the standards change in response to a variety of conditions. The reporting of general limitation of activity has also been shown to be affected by changes in such things as the criteria used by government agencies in determining eligibility for disability benefits (Wilson and Druny 1984). Thus, observed changes in the indicator may not reflect changes in underlying health status.

5. Medical Conditions

A variety of approaches can be used to measure disease incidence and prevalence including: diagnosis by a health care provider; symptomatology (if an appropriate symptom battery exists); or medication use. Additional information of interest related to conditions includes date of onset, the use of health care services, and the impact of the condition on the individual's ability to function. Severity of condition is often inferred from the type of medications or health care services used or the nature of the impact of the condition. In examination surveys, diagnostic testing adds to the information available on the prevalence of conditions. Most important, direct physical exams allow for the measurement of previously undiagnosed conditions.

A major issue in the collection of condition data is the number of conditions to be considered. Condition checklists can be used to obtain information on a great

number of conditions, but in most cases, only a minimal amount of information on any given condition is collected. In addition, long checklists are difficult to administer and result in reporting errors. However, the use of comprehensive lists does allow for the ascertainment of co-morbidities. An alternative is to obtain more detailed information on a select group of conditions. Criteria such as the prevalence of the condition or the potential impact for the individual and the health care system are used to select the target condition. In the NHIS, for example, information is obtained for adults on heart disease, diabetes, cancer, pulmonary diseases, depression, and arthritis.

Some surveys, such as the NHIS (until 1997 when a major questionnaire redesign was fielded), limit their focus to conditions that result in the receipt of medical care or in activity restriction, thus placing more emphasis on the impact of the condition than on the fact that a diagnosis had been made or symptoms reported. This provides a somewhat more concrete referent for the questions being asked which should result in improved data validity and reliability. However, limiting the definition of health problems to those that result in behavior such as seeking care or restricting activity is more appropriate when ill health is primarily the result of acute conditions where onset is clear and there is a closer association between the condition and the behavior. As a result of the 'epidemiologic transition,' acute conditions have given way to chronic conditions as the major sources of morbidity. Chronic conditions are more difficult to diagnose, and longer periods are spent in the disease state both prior to and after diagnosis. As a result, the collection of information on symptoms takes on greater importance when investigating chronic conditions that may or may not have been diagnosed. Diagnostic examination or testing is necessary for some important conditions (e.g., hypertension, cancer) which may be asymptomatic in their early or less severe stages.

Requiring that health care use or activity limitation be present in identifying conditions also makes it difficult to investigate social correlates of health status since the definition incorporates a social dimension. Access to sick leave or affordable medical care may be limited by social or economic factors. Persons who do not demonstrate the required behaviors because of these social factors will not be considered as having a health condition.

Examination surveys like the NHANES provide a complementary mechanism for obtaining information on conditions. Self-reported information, similar to that collected on the NHIS on self-perceived health status, disease symptomatology, physician diagnosis, and level of functioning, is obtained from in-home or MEC interviews. In addition, the results of the physical exam, and of the blood and other tests, can be used to objectively measure health status. For example, information on diagnosed diabetes is

obtained from the interview, and information on undiagnosed diabetes is obtained from blood tests. Information on functional limitations related to diabetes (e.g., vision and mobility problems) coupled with similar information available from the other components of the NHANES allows for a detailed characterization of an individual's health status. The information is also used for the investigation of risk factors for disease and for monitoring of functional outcomes.

Information on conditions can also be obtained from surveys of health care providers (hospitals, private physicians, hospital outpatient departments, emergency departments, ambulatory surgery centers, and long-term care). Sampling frames of providers must first be developed. Random samples of discharges, stays, or visits are then selected from the records of a random sample of providers. Information on conditions is routinely found on the medical records in addition to other descriptive information concerning the patient and the encounter. Such survey designs are generally used to obtain information on health care utilization and the health care system, but the information available can be used to study conditions that are associated with medical care.

Health care or provider surveys offer an independent assessment of disease prevalence by approaching the problem from the provider side rather than the patient side. Health care surveys are particularly important since they obtain information on rare conditions that would not be picked up in population based surveys. They also provide larger sample sizes than would be obtained in surveys such as the NHIS and the NHANES. While information is collected on all conditions for which health care is obtained, the nature of the information available is limited to that which would routinely be obtained from records. In addition, the event (e.g., the discharge), rather than the person, is the unit of analysis. For example, while it is possible to estimate the number of hospital stays associated with hip fracture, it is not possible to estimate the number of persons with hip fracture even if one could assume that all hip fractures were admitted to the hospital. An enhancement to health care surveys would be to follow patients after the health care encounter.

Cause of death statistics obtained from a vital statistics system also provide essential information on medical conditions. While not error free, a great deal of effort is put forth to standardize the collection and coding of cause of death (<http://www.cdc.gov/nchs/icd9>).

6. Functioning

An important component of health status concerns functioning—physical, cognitive, emotional, and

social. While conditions relate more to the biologic aspects of health, functioning moves more into the social realm. Information on functional status can be ascertained through objective tests but is most often obtained through direct reporting by the individual or a proxy. Information on functional status encompasses a wide range of activities from the very basic such as raising an arm to the more complex such as working and going to school. Obtaining information on functioning is complicated by the fact that performance is affected not only by physiologic abilities but by the environment as well. A person's ability to walk may be impaired by a condition that leads to weakened muscles, but with the use of braces or a wheel chair along with appropriate accommodations in the physical environment (e.g., ramps), that individual is mobile and able to carry out appropriate social roles. For some purposes, information is needed on functional ability without the use of assistance of any type. In other cases, functioning ability as measured by usual performance is important. In any case, questions must be specific concerning how performance should be measured. A large number of question batteries have been developed to measure function (McDowell and Newell 1996). Some use only a few questions, focus on the more complex social roles, and do not differentiate between different aspects of functioning (physical, cognitive, emotional, or behavioral). Others are quite lengthy and obtain detailed information on a wide range of activities both with and without assistance.

In the NHIS, information is obtained on limitation of activity (i.e., the subject's ability to perform age-/sex-appropriate roles such as working, going to school, keeping house), activities of daily living (e.g., bathing, dressing, toileting, eating, and transferring), and physical functioning (e.g., walking, bending, and standing). Information is also obtained on the conditions causing the limitation. A subset of the NHIS functioning questions is provided in Table 2.

As for self-perceived health status, measures of functioning are affected by contextual parameters as well as unstated standards against which respondents compare their own functioning. The less explicit the referent standard, the more likely that the response will be affected by external factors. For example, no explicit standard is provided when asking subjects if they are limited in the kind or amount of work they can do. Responses to this question will be affected by the kind of work that the individual thinks he or she should be doing as defined by their own expectations and by the requirements set by government agencies such as disability programs as well as by more objective states of health. This is less of a problem if one is interested in monitoring the impact of health on work. However, it is much more of a problem if the aim is to understand and monitor a more objective measure of health.

Many health surveys are limited to the noninstitutionalized population and therefore do not include

Table 2

Examples of questions on functioning from the National Health Interview Survey

Family members under five years of age	Are/Is {fill in name of child} limited in the kind or amount of play activities he/she/they can do because of a physical, mental, or emotional problem? Is {subject name} able to take part at all in the usual kinds of play activities done by most children {subject name}'s age? Do any of the children under 18 in this family, {fill in names of children under age of 18} receive Special Education or Early Intervention Services? Has a doctor or health professional ever told you that {subject name} had Mental Retardation or any other developmental delay? Has a doctor or health professional ever told you that {subject name} had Attention Deficit Hyperactivity Disorder (ADHD), Attention Deficit Disorder (ADD), Mental Retardation, or any other developmental delay.
Family member between the ages of 4 and 17	Do any of the children under 18 in this family {fill in names of children under age of 18} receive Special Education or Early Intervention Services?
Family member older than 17 years of age	Do you now have any health problem that requires you to use special equipment, such as a cane, a wheelchair, a special bed, or a special telephone? By yourself, and without any special equipment, how difficult is it for you to walk a quarter of a mile—about 3 city blocks, walk up 10 steps without resting, stand or be on your feet for about 2 hours, sit for about 2 hours, stoop, bed, kneel, or reach up over your head? (each asked separately) By yourself, and without using any special equipment, how difficult is it for you to use your fingers to grasp or handle small objects, lift or carry something as heavy as 10 pounds such as a full bag of groceries, push or pull large objects like a living room chair? (each asked separately) By yourself, and without using any special equipment, how difficult is it for you to go out to things like shopping, movies, or sporting events, participate in social activities as visiting friends, attending clubs and meetings, going to parties, do things to relax at home or for leisure (reading, watching TV, sewing, listening to music ...)? (each asked separately) Because of a physical, mental, or emotional problem, {do/does} {you/anyone in the family} need the help of other persons with personal care needs, such as eating, bathing, dressing, or getting around inside of the home?

Table 2
(Continued)

{Do/Does} {you/subject name} need the help of other persons with bathing or showering, dressing, eating, getting in or out of bed or chairs, using the toilet, including getting to the toilet, and getting around inside of the home? (each asked separately) Because of a physical, mental, or emotional problem, {do/does} {you/anyone in the family} need the help of other person in handling routine needs, such as everyday household chores, doing necessary business, shopping, or getting around for other purposes? Does a physical, mental, or emotional problem now keep {you/any of these family members} from working at a job or business? {Are/Is} {you/anyone in the family} limited in any way because of difficulty remembering or because {you/they} experience periods of confusion? What condition or health problem causes you to have difficulty with activities/these activities? How long {have/has} {you/subject name} had this condition?

Source: Centers for Disease Prevention, National Center for Health Statistics. 2000 NHIS Family Section, Sample Adult, and Sample Child Questionnaires.

those segments of the population in poorest health or with the highest level of functional limitations, such as those in long-term care settings. Special studies of these populations can be carried out and there should be some attempt to combine results from the non-institutionalized and long-term care populations (<http://www.cdc.gov/nchs/about/major/nnhsd/nnhsd.htm>, <http://www.cds.duke.edu/text/nltcs.html>). Changes in the delivery of long-term care with the expansion of transitional establishments have made this an even more important issue for survey designers.

7. Determinants of Health

Health surveys generally include measures of risk factors, health behaviors, and nonhealth determinants or correlates of health such as socioeconomic status. The range of measures that can be included is wide and varies by survey. Basic demographic and socioeconomic variables are usually included (i.e., age, race/ethnicity, education, income, urbanicity, region). Tobacco use, alcohol use, diet, and physical exercise are common health behaviors. In order to identify causal relationships between determinants and outcomes, longitudinal designs are needed. The value of health surveys would be increased by the inclusion of

a broader array of potential determinants of health (biologic, psychological, and social).

8. Summary Measures of Health

There has been considerable interest in the development of summary measures of health in order to more easily characterize and monitor the health status of a population and for use in cost-benefit and cost-effectiveness analyses. There is also increasing interest in developing basic health status indicators comparable to the economic indicators currently in use. These indicators would be used to monitor the effects of public policy and to identify areas where interventions are needed. An overall health measure is very appealing in this context. Many health surveys include the information necessary to calculate the summary measures that have been developed. One such approach is to use a measure of morbidity or health status to transform life expectancy into healthy life expectancy. Examples include Disability Free Life Years, Healthy Life Expectancy, and Years of Healthy Life. Constructs such as these combine measures of the different aspects of health into a single number using a conceptual model that views health as a continuum. The health states along the continuum are assigned numbers that represent the values that either society as a whole or individuals place on that health status. Various methods have been used to determine

the values of the different health states. These measures are used to modify duration of life (Erickson et al. 1989).

9. Health Care Utilization

The use of the health care system is a major dimension of health. Information on utilization and related characteristics such as health insurance are collected from person-based surveys and through surveys of administrative records. In person-based surveys, respondents are asked to recall the number of contacts they have had with various health care providers and the nature of the provider and the contact. Information on the receipt of particular services, such as immunizations or mammography, is also obtained. This is often difficult for respondents. Research has shown that reporting of doctor visits drops off considerably when the recall period extends beyond two weeks (Edwards et al. 1996). To address this recall problem, some surveys such as the Medical Expenditure Panel Survey (MEPS) ask respondents to keep detailed diaries of their medical care contacts, obtain the name of the provider, and then contact the providers (with the subject's permission) to verify the self-reported information (Monheit et al. 1999). Information on utilization can also be obtained from surveys of providers, usually providing more accurate information on the nature of the contact, including the reason for the contact and the services provided. Surveys of providers can also yield valuable information on the structure of the health care system itself and how the characteristics of the system affect access to care and the cost of care.

10. Longitudinal Designs

Most health surveys use a cross-sectional design which provides data for the calculation of population averages and is most appropriate for monitoring changes in health over time. A cross-sectional design is not appropriate for investigating causal patterns or for documenting transition probabilities from one health state to another where change at the individual level needs to be measured. A solution would be longitudinal studies which are expensive and more complicated to field. One alternative is to transform cross-sectional studies into longitudinal studies by linking information from mortality records to the survey data. The United States has developed a National Death Index which allows investigators to track the vital status of their study cohorts by matching identifiers to a complete listing of death certificates. Using a probabilistic matching algorithm, information is provided on the fact, date, and cause of death (<http://www.cdc.gov/nchs/r&d/ndi>). Cross-sectional surveys also can form the baseline cohorts for future active longitudinal follow-ups (US DHHS 1987). These

longitudinal activities present extensive opportunities for understanding the relationship between risk factors and disease outcome as well as the natural history of disease. Generation and analysis of such longitudinal data is greatly facilitated in countries (e.g., Sweden) where various governmental information systems and surveys can be linked.

11. Technological Advances

Advances in technology have affected the conduct of all surveys. Computer-assisted telephone and personal interviewing have allowed for increasing levels of complexity of survey administration. Errors associated with keying and coding have been replaced by errors associated with programming. The use of these technologies has resulted in the need for longer lead times prior to fielding a study but has reduced the time from the end of data collection to the availability of results.

The increase in electronic as opposed to paper files has greatly expanded the opportunities for linking data from various sources. Ecological or contextual information about the geographic area in which survey respondents live can now be linked to the survey data. This has the advantage of allowing researchers to address different levels of analysis. Survey data can also be augmented with administrative data from mortality files and from health care providers. In addition to reducing survey costs, this often leads to data of improved quality.

12. Methodological Research and Analytic Techniques

The quality of health survey data has been greatly increased by methodological research on survey techniques (Bradburn and Sudman 1979, Fowler 1995, Lessler and Kalsbeek 1992). A major advancement has been the development and application of cognitive methods for questionnaire construction (Sirken et al. 1999). Drawing from the field of cognitive psychology, cognitive research laboratories have been established and can investigate how respondents perceive and answer questions. Another fruitful area of research involves studying the behavior of both interviewers and respondents within the social context of the interview situation (Cannel et al. 1977). A more structured approach to the errors found in health surveys will enhance data quality. The results of this research will lead to improved design that minimizes error but will also lead to ways to adjust for known errors in the analysis and interpretation of results (Jabine 1987).

Advances have also been made in the development of appropriate analytic tools for use with health survey data. These include methods to account for such issues as survey nonresponse and complex survey design. As

it is possible to collect data of increasing complexity, it will be necessary to develop the analytic approaches appropriate for the collection methods used (Korn and Graubard 1999, Biemer et al. 1991, Groves and Couper 1998, Rubin 1987).

13. Privacy and Confidentiality

For ethical reasons and in order to obtain high response rates and valid information, most health surveys closely guard the information provided by respondents. In some cases, the requirement to protect confidentiality is legislatively mandated. This is particularly important for health surveys, given the personal nature of the information collected. Confidentiality can also be protected by not releasing data that could identify a respondent. Most survey sponsors subject files that are to be released for public use to rigorous disclosure review. The risk for inadvertent disclosure has increased in recent years. It is no longer necessary to have access to large mainframe computers to utilize survey data, much of which are provided on the Internet or on CD-ROMs. In addition, databases not related to the survey data, but which can be used to identify individuals in the survey data files, are more available and more easily accessible.

It is the potential linking of survey data to these external databases that increases the risk of disclosure, and the risk increases with the amount of information that is available. The ability to link external data to survey responses presents an analytic breakthrough as the utility of survey data can be greatly increased through the appropriate linkages, however, this ability also greatly increases the risk of disclosure especially when the linkages are done in an uncontrolled and inappropriate way. The increased sensitivity to issues of privacy in many countries, especially as related to health care, is affecting how confidential data are being protected. There has been a decrease in the amount of data that can be released as public use files. Other mechanisms are being developed so that access to data can be maximized while protecting confidentiality. The use of special use agreements, licensing and research data centers are examples of these approaches (Journal of Official Statistics 1998).

See also: Health Research, Qualitative; Public Health; Public Health as a Social Science; Quality of Life: Assessment in Health Settings; Statistical Systems: Health

Bibliography

Biemer P P, Groves R M, Hyberg L, Mathiowetz N, Sudman S (eds.) 1991 *Measurement Errors in Surveys*. Wiley, New York

- Bradburn N M, Sudman S 1979 *Improving Interview Method and Questionnaire Design*. Jossey-Bass, San Francisco
- Cannell C F, Monquis K H, Launent A 1977 A summary of studies of interviewing methodology. *Vital and Health Statistics* (Series 1, No. 69)
- Edwards W S, Winn D M, Collins J G 1996 Evaluation of 2-week doctor visit reporting in the National Health Interview Survey. National Center for Health Statistics. *Vital and Health Statistics* (Series 2, No 122)
- Erickson P, Kendall E A, Anderson J P, Kaplan R M 1989 Using composite health status measures to assess the nation's health. *Medical Care* 27(3): 566–77
- Fowler F J 1995 *Improving Survey Questions*. Sage, Thousand Oaks, CA
- Groves R M, Couper M P 1998 *Nonresponse in Household Interview Surveys*. Wiley, New York
- Idler E L, Angel R J 1990 Self-rated health and mortality in the NHANES I. Epidemiologic follow-up study. *American Journal of Public Health* 80(4): 446–52
- Jabine T B 1987 Reporting chronic conditions in the National Health Interview Survey: A review of tendencies from evaluation studies and methodological tests. National Center for Health Statistics. *Vital and Health Statistics* (Series 2, #105)
- Journal of Official Statistics 1998 Disclosure limitation methods for protecting the confidentiality of statistical data. *Journal of Official Statistics* 14: 337–573 (Special issue)
- Korn E L, Graubard B I 1999 *Analysis of Health Surveys*. Wiley, New York
- Lessler J T, Kalsbeek W T 1992 *Nonsampling Error in Surveys*. Wiley, New York
- McDowell I, Newell C 1996 *Measuring Health: A Guide to Rating Scales and Questionnaires*. Oxford University Press, Oxford, UK
- Monheit A C, Wilson R, Arnett R H (eds.) 1999 *Informing American Health Care Policy: The Dynamics of Medical Expenditure and Insurance Surveys, 1977–1996*. Jossey-Bass, San Francisco
- Moore J 1988 Self/proxy response status and survey response quality: A review of the literature. *Journal of Official Statistics* 4(2): 155–72
- National Center for Health Statistics 1999 *Health United States*. National Center for Health Statistics, Hyattsville, MD
- National Center for Health Statistics 2000 *International Health Data Reference Guide 1999*. National Center for Health Statistics, Hyattsville, MD
- Rubin D B 1987 *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York
- Sirken M G, Herrmann D J, Schecter S, Tourangeau R (eds.) 1999 *Cognition and Survey Research*. Wiley, New York
- Stewart A L, Ware J E (eds.) 1992 *Measuring Functioning and Well-being: The Medical Outcomes Study Approach*. Duke University Press, Durham, NC
- United States of America, Department of Health and Human Services 1987 Publication No. (PHS) 92-1303, Series 1, No. 27. Plan and Operation of the NHANES I Epidemiologic Followup Study
- Wilson R W, Druny F 1984 Interpreting trends in illness and disability: Health statistics and health status. *Annual Review and Public Health* 5: 83–106
- World Health Organization 1948 *World Health Organization Constitution in Basic Documents*. World Health Organization, Geneva

J. H. Madans

Published by Elsevier Ltd.

6627

Hebb, Donald Olding (1904–85)

Donald Hebb was responsible in no small part for the profound change in psychological outlook that emerged in the 1950s. His influence on the field of cognitive neuroscience is still apparent.

1. Education and Early Life

Hebb was born, the eldest of four siblings, on July 22, 1904 in Chester, Nova Scotia. Both his parents were physicians. He was educated at home until the age of eight by his mother, using Montessori methods. He then attended the local school, which he hated.

Hebb's ambition as a young man was to become a novelist. He graduated from Dalhousie University (Halifax, Nova Scotia) in 1925 with a BA in English and philosophy. While awaiting inspiration for the 'Great Canadian Novel' he returned to his village school as a teacher. In search of experience he spent the next summer as an itinerant farm worker in western Canada and tried, unsuccessfully, to get a job as a deckhand in Vancouver.

In 1927 Hebb was back East, teaching at a Montreal high-school. Becoming interested in Sigmund Freud's ideas, he applied to McGill University to study psychology part time. The next year he was appointed principal of an elementary school and enthusiastically set about instituting and testing new teaching methods. In 1930 he was bedridden for a year with a tubercular hip, during which time he conceived a theory of how spinal reflexes might be learned *in utero*. This formed the basis of an MA thesis, now lost. Hebb (1980) later described it as nonsense.

In 1932 Hebb began doctoral research, still part-time, at McGill. The experiments, which involved conditioning dogs, were supervised by Leonid Andreyev, a former student of Pavlov, but Hebb soon became disillusioned with the technique and when his first wife was killed in a motor accident he abandoned the research.

Hebb enjoyed teaching and administration and was tempted to make a career of being a headmaster. He finally chose to continue his studies, however, and in 1934 he was accepted as a student at the University of Chicago by Karl Lashley, whose book, published a few years earlier (Lashley 1929), had established him as the pre-eminent physiological psychologist of the period.

Hebb spent a year in Chicago where, in addition to his close association with Lashley, he attended lectures by Wolfgang Köhler, Paul Weiss, C. J. Herrick, and the statistician L. L. Thurstone. Then Lashley accepted the offer of a professorship at Harvard and Hebb accompanied him. He was granted a Harvard PhD in 1936 for research on the effects of early experience on vision in the rat.

2. Montreal Neurological Institute

After a post-doctoral year of research with Lashley, Hebb was awarded a two year Fellowship to study mental changes resulting from brain operations at the newly established Montreal Neurological Institute (MNI). The operations were mostly frontal and temporal lobectomies performed by Wilder Penfield to combat epileptic seizures. Hebb made what is probably the first observation of a visual deficit after human right-temporal lobe removal. To his surprise he found no measurable effect of frontal lobectomy on standard measures of intelligence (Hebb and Penfield 1940). Attempts to understand this result lay at the foundation of much of Hebb's later work.

3. Primate Research with Lashley

After his MNI Fellowship, Hebb obtained a post as Lecturer at Queen's University (Kingston, Ontario). There he devised a variable rat-maze so that he could continue his brain lesion research using rats. In 1942 Lashley offered him a five year Fellowship to study effects of brain lesions in chimpanzees at the Yerkes Laboratories of Primate Biology, Florida. Partly because of his heavy teaching load at Queens, he accepted it.

The project was to study the effects of brain lesions on intelligence and emotion in chimpanzees. Lashley was to work on intelligence tests, Hebb on tests of emotion, but no lesions were made before 1947 when Hebb's Fellowship ended. Hebb did, however, assemble a good stock of chimpanzee stories that contributed to the great popularity of his subsequent lectures, and in the intellectually stimulating surroundings he wrote his most influential work, *The Organization of Behavior*, which was published in 1949.

4. McGill Once More

After World War II, the McGill psychology department had dwindled to one part-time child psychologist, so in 1947 Hebb was one of four psychologists who were recruited to rebuild the department. The next year he was appointed chairman.

Having made a highly successful foray into the world of science, Hebb returned to teaching and administration with enthusiastic vigor. He taught the introductory course to overflow classes for the next 25 years, and wrote three editions of a highly respected textbook. He also conducted a seminar for all incoming graduate students, which gave him the opportunity to coach them in the presentation of succinct and closely timed talks, as well as to instill a thorough understanding of a range of theoretical issues. To test some predictions of his theory, he supervised research on early learning in rats and dogs, and effects of sensory deprivation on human thought processes.

In later years Hebb received many honors; he was President of both the Canadian (1952) and the American (1960) Psychological Associations. He was elected a Fellow of the Royal Society of London (1966), a Foreign Associate of the National Academy of Sciences (USA), and an Honorary Member of the American Academy of Arts and Sciences. In 1970 he was appointed Chancellor of McGill University. In 1974 Hebb retired to his family home in Chester Basin and became active in the psychology department at Dalhousie University, where he was appointed Honorary Professor. He published his last monograph, *Essay on Mind* in 1980 and died in Halifax on the 20th of August, 1985 of heart failure during hip-replacement surgery.

5. Hebb's Contributions to Neuroscience

During his years as a graduate student Hebb learned the hard way what was wrong with contemporary psychology. At McGill, as is evident from the topic of his Masters thesis, he flirted with the extreme empiricism then rampant in North America. Influenced by Boris Babkin and Leonid Andreyev, students of Pavlov, he passed through a Pavlovian phase. When he moved to Chicago he entered the Gestalt enclave of Wolfgang Köhler and Karl Lashley; and at Harvard he encountered the historian, Boring, who taught him what psychology had been like before the Behaviorists. Sooner or later, for reasons that are cogently presented in the first few chapters of *The Organization of Behavior* (Hebb 1949), he turned against all the established schools of psychology.

Because of his research on cortical lesions and his neural theories, Hebb is often labeled as a physiological psychologist, but his interests were really much broader. *The Organization of Behavior* was an important catalyst, aiding the revival of cognitive psychology, and although he was a zealous proponent of the idea that the brain is the organ of mind, Hebb's interest in neurophysiology took second place to his interest in thinking and other aspects of behavior. In his hands neurologizing was a tool to restore respectability to the study of mental functioning in psychology.

As a psychologist, Hebb was frustrated by the behaviorists' refusal to acknowledge such important processes as attention, expectancy, and thought, lest they should be accused of vitalism. Hence his eagerness to show that those mental processes could be performed by an entirely material brain. He based his neural model as far as possible on established physiological principles, but neuroscience was on the brink of a revolution and many of those principles were soon to be revised.

For example, Eccles, whose defense of the electrical theory of synaptic transmission Hebb had embraced, recanted and demonstrated excitatory and inhibitory

chemical synapses. Furthermore, the behavioral importance of subcortical structures such as the reticular arousal system, the amygdala and the hippocampus became clearer, and improved staining techniques revealed a much more complex visual system than was known to Hebb. The fact that these revelations had almost no effect on the impact of Hebb's book shows to what extent the psychological implications of his neurologizing outweighed the underlying physiology. By proposing a material basis for concepts, Hebb demolished barriers to the study of cognitive processes that the behaviorists had erected a generation earlier. In his Presidential address to the Experimental Psychology Division of the American Psychological Association in 1954, Hebb introduced the reticular activating system into his model to extend his explanation of emotion, but apart from that he did little to revise the theory in the light of new neuroscientific data.

His most famous contribution to neuroscience was not based on neurophysiological research, it was inferred from behavior. Most physiologists thought that learning synapses were strengthened by use, but Hebb along with other psychologists realized that association involved at least two activities. Thus he postulated that learning changes at synapses require simultaneous firing of both presynaptic and postsynaptic neurons. Synapses with properties similar to those postulated were subsequently identified experimentally and are now known as Hebb synapses.

Hebb held strong views about the 'nature-nurture' question. After his early gyrations he adopted the unassailable position that both inheritance and experience are totally necessary for intelligent behavior, using the analogy of area as the product of length and breadth. Of course, one may still be interested in whether an area is square, long and narrow, or short and fat. In spite of Hebb's bipartisanship, his concept of the development of vision was long in the 'nurture' dimension. Hebb's visual cortex bore a strong resemblance to Locke's *tabula rasa*.

Thus, Hebb assumed that neurons are randomly connected in the newborn visual cortex. With experience, synapses are strengthened between neurons activated simultaneously by sensory input, producing groups of connected neurons (*cell assemblies*) that then represent the input. Hebb further postulated that a complex figure elicits a series of eye movements, thereby activating a succession of associated cell assemblies that Hebb called a *phase sequence*.

Anatomical diagrams of the cortex suggested to Hebb that cell assemblies consist of networks of closed loops. This led him to speculate that once a cell assembly is activated impulses continue to circulate around the loops for some time (reverberation), constituting a short-term memory. In this way Hebb explained the persistence of an idea or mental image in the absence of the corresponding object. Today, short-term memory may be explained more plausibly by

persistent chemical traces, but Hebb never abandoned his reverberation theory, which still dominates the thinking of many neural model builders and neurophysiologists who study the visual system.

According to Hebb, the impulses reverberating around the cortical loops of an activated cell assembly produce synaptic changes that bind the neurons together more permanently. This explains long-term memory. Ideas arouse other ideas by association, thus cell assemblies must acquire associations with each other whenever they are active at about the same time. Hebb never satisfactorily explained how, after linking in this way, they avoid merging to become bigger cell assemblies.

The capacity to explain intention, attention and other mental processes was an important goal of Hebb's theorizing but one that, in *The Organization of Behavior*, he was content to achieve in principle rather than by rigorous neural modeling. Short-term expectancy and attention were explained in terms of the phase sequence getting ahead of the sensory input. Thus a previously experienced sequence of stimuli sets up a central process that facilitates cell assemblies in advance of the sensory input they represent. Hebb postulated higher order central processes that facilitate more extended sequences. In much of his writing Hebb treats the cell assembly as equivalent to an idea; its neural basis is tailored to fit behavioral observation.

6. Summary

The Organization of Behavior is now a classic. Half a century after it was written it is still frequently cited. Many psychologists regard it as the herald of an era of cognitive research that blossomed in the 1950s. Hebb's ideas about the representation of concepts and their associations with each other are the common currency of almost all present-day psychologists, even those who prefer to ignore the brain, but devotees of neural networks and neurophysiologists studying sensory systems are those who most often refer explicitly to his theory.

After Hebb left Lashley's 'think tank' in Florida and took up his duties as chairman of the McGill psychology department, most of his energies turned to building and running a strong department there, including teaching. Although Hebb took every opportunity to promote neurologizing in psychology, he showed little further inclination to keep pace with the rapid advances in neuroscience or to make any major changes to his neural theory.

Apart from his own characteristically stylish and witty account (Hebb 1980a), there have been several more recent commentaries that complement this entry for those who wish to learn more about Hebb's life and work. These include nine essays in his honor, written by former students and edited by Jusczyk and Klein (1980), and a delightful chapter by Stephen Glickman (1996) which accurately portrays the man and his

achievements. Fentress (1999) and Klein (1999) edited a special issue of the *Canadian Journal of Experimental Psychology* that contains a further collection of papers on Hebb and on research inspired by his ideas. Orbach, a student of both Hebb and Lashley, published a book (1998) in which the contributions of his two mentors are discussed and compared.

See also: Behaviorism; Brain Stimulation Reward; Cognitive Neuropsychology, Methodology of; Cognitive Psychology: History; Cognitive Psychology: Overview; Learning and Memory, Neural Basis of; Lesion and Behavior Approaches in Neuroscience; Linear Algebra for Neural Networks; Long-term Depression (Cerebellum); Long-term Depression (Hippocampus); Long-term Potentiation and Depression (Cortex); Long-term Potentiation (Hippocampus); Memory: Synaptic Mechanisms; Neural Development: Mechanisms and Models; Neural Networks, Statistical Physics of; Neuropsychological Functioning, Assessment of; Synaptic Efficacy, Regulation of

Bibliography

- Fentress J C 1999 The organization of behavior revisited. *Canadian Journal of Experimental Psychology* **53**: 8–20
- Glickman S E 1996 Donald Olding Hebb: Returning the nervous system to psychology. In: Kimble G A, Boneau C, Wertheimer M (eds.) *Portraits of Pioneers in Psychology*. Erlbaum Press, Mahwah, NJ, Vol. 2, pp. 227–44
- Hebb D O 1949 *The Organization of Behavior: a Neuropsychological Theory*. Wiley, New York, pp. 1–335
- Hebb D O 1955 Drives and the C.N.S. (conceptual nervous system). *Psychology Review* **62**: 243–54
- Hebb D O 1980 *Essay on Mind*. Erlbaum Press, Hillsdale, NJ
- Hebb D O 1980a D. O. Hebb. In: Lindzey G (ed.) *A History of Psychology in Autobiography*. Freeman, San Francisco, CA, Vol. 7, 273–303
- Hebb D O, Penfield W 1940 Human behavior after extensive bilateral removal from the frontal lobes. *Archives of Neurology and Psychiatry* **44**: 421–38
- Jusczyk P W, Klein R M (eds.) 1980 *The Nature of Thought. Essays in Honor of D. O. Hebb*. Erlbaum Press, Hillsdale, NJ
- Klein R M 1999 The Hebb legacy. *Canadian Journal of Experimental Psychology* **53**: 1–7
- Lashley K S 1929 *Brain Mechanisms and Intelligence*. University of Chicago Press, Chicago, pp. 1–186
- Orbach J 1998 *The Neuropsychological Theories of Lashley and Hebb*. University Press of America, Lanham, MD

P. M. Milner

Hegel and the Social Sciences

G. W. F. Hegel (1770–1831), born in Stuttgart, achieved his full fame in Berlin between 1818 and 1831 as a professor of philosophy. Much of his fame rests

on the posthumously edited and published lectures on the philosophies of art, religion, and history, along with his epochal book of 1807, the *Phenomenology of Spirit*. However, despite Hegel's immense influence on modern thought, he has remained one of the most controversial and difficult to interpret philosophers of the modern period. It is nonetheless widely agreed that Hegelian-inspired theories of society rest on certain fundamental theses: first, they give fundamental priority to the 'totality' of social life over more individualist accounts; second, such theories claim that we can only understand the kinds of demands social life makes in terms of its historical embeddedness; and, finally, such theories claim that there can be no value-free approach to history and society. There is, however, not nearly as much consensus as to what each of these general claims meant for Hegel or about what follows from them. In some interpretations, Hegel has been taken as an ominous figure, a kind of dark warning for modern academics to stay away from metaphysics; in other interpretations, Hegel is taken as one of the original theorists of European modernity and as an ongoing source of insight and inspiration.

Although part of the difficulty of interpreting Hegel has to do with his notoriously obscure style and his sometimes ferociously idiosyncratic jargon, much of the difficulty has to do with how one is to take his statements about the role of mentality (*Geist*, perhaps better rendered in English somewhat artificially as 'mindedness') in nature, society, and history (and often one's interpretation of that depends how one understands Hegel's ultimate stance toward religion). This is also tied into Hegel's general holism, the thesis that the politics, art, religion, and economy of a society are so intimately related to each other that not just any concatenation of such practices and institutions is compatible with any others, and that the point of such practices and institutions is to provide the vehicles through which a way of life can collectively reflect on and become self-consciously aware of what ultimately matters to it. In Hegel's own lifetime and afterwards, people took his holism more or less in one of two ways: either as a grandly metaphysical statement about an absolute reality (called *Geist*, mindedness) that assumes different forms in history and which progressively reveals itself to itself; or as a more normative account of agency that does not invoke that kind of metaphysical entity acting behind the scenes to explain things.

To understand what is at stake in that debate, one must at least have some idea of the rather heated atmosphere present in German philosophy at the time in which Hegel came to prominence.

By the end of the eighteenth century, Kant's critical philosophy had changed the very terms of debate in Germany, and while certainly not everybody became a Kantian, almost all aspiring philosophers in Germany felt that they had to come to terms with Kantian

thought. Kant had, according to his own description of things, essentially demolished the longstanding feud between empiricists—those who believed that science was only a sophisticated and methodical purification of our more ordinary responses to the givens of experience—and rationalists—those who thought that Galilean and Newtonian physics showed that the mind was capable of mathematically articulating the structure of the universe independent of experience. Kant argued that there were no givens of experience at all, that our conscious experience of the world always involved an active and spontaneous 'taking-up' of that experience (a rationalist theme), although our activities of 'taking-up' experience depended for that experience for its empirical content (an empiricist theme). He also argued that in the moral realm, we are completely self-legislating; there are no givens within human nature or in the cosmos at large that can determine what may count as a moral principle for us. The reality of this radical human spontaneity and freedom was secured, so Kant had argued, by distinguishing between the way in which we necessarily had to describe and order our experience of the world and the way that the world existed in itself independent of any description we might give it. We could never know, in Kant's famous phrase, the way things were in themselves, only the way in which they necessarily appear to us. Thus although we had to appear to ourselves as standing under deterministic natural causal laws, we also had to postulate for practical reasons the necessity of our freedom as we were 'in ourselves,' even though we could never theoretically demonstrate the existence or lack of existence of such radical freedom.

For Hegel's generation, who had come to believe, however indistinctly, that they were leading unprecedented lives and for whom the French Revolution was the defining event, Kant's philosophy of freedom and spontaneity had an electrifying effect. Kant's immediate successors were also convinced that the outstanding problems with his own solution called for an effort to 'complete' the project he had begun. This took the forms of attempting to establish Kantianism on better foundations, or of focusing on the way in which Kant had (or should have) rejected all forms of 'givenness' in epistemology and morality, or of showing how Kant's own dualisms were to be reconciled by showing them to be only mutually conflicting 'appearances' of some deeper underlying 'absolute' reality that itself was only to be apprehended through some non-discursive act of intellectual intuition. The latter turn was famously carried out by Hegel's old friend, F. W. J. Schelling, who crafted an entire theory of nature and mind as modes of appearance of the 'absolute,' or as the way the absolute gradually realizes itself in history. (In part, much of the debate about Hegel's philosophy has since centered on just how 'Schellingian' or 'non-Schellingian' his thought really was.)

In much of his writing prior to the 1807 *Phenomenology of Spirit*, Hegel had worried about what shape a 'modern' way of life would take and in particular what role religion, especially a reinterpreted form of Christianity, might play in such a modern way of life. Throughout his development, however, he held fast to the post-Kantian notion that the modern, post-revolutionary world was something qualitatively new and demanded a 'new sensibility' appropriate to itself. To that end, Hegel's 1807 *Phenomenology* was intended to demonstrate what would be required of 'we moderns' if we to conceive of our lives without any secure, 'given' foundations. In his *Phenomenology*, Hegel sarcastically characterized Schelling's notion of the Absolute as 'the night in which all cows are black,' signaling his rejection of the idea of an absolute reality that is beyond all conventional descriptions and differentiations and which can only be grasped in some non-discursive mode. So Hegel argued, rejecting all 'givens' and affirming Kantian notions of spontaneity implied there could be no inherently privileged viewpoint without at the same time also implying that we were stuck with any kind of pure relativism about viewpoints.

In an early chapter in the *Phenomenology* (drawing on some themes taken from Fichte and Schelling), Hegel constructed an imaginary encounter between two agents, each thinking of himself as seeking only to satisfy the desires given to him by organic life. To the extent that each agent legitimates his own judgments about what is valid for him only in light of locating himself within the totality of 'life' and its demands, he demands of the other that the other recognize his point of view as authoritative, that he may do what he pleases to satisfy his own desires, even to use the other to get what he wants. In making that demand, though, each finds that it is not 'life' itself that is setting the terms of judgment but his own self-conceived project for his life, of his bestowing on *his* desires a prior claim to satisfaction over the desires of others. Each thus demands 'recognition' from the other of the authoritativeness of his own point of view—not merely that he has pressing desires but that he determines which desires and whose desires have a legitimacy to satisfaction. Those reciprocal and incompatible demands necessitate a struggle for such recognition, and to the extent that one of the parties decides that his own self-conception is more important than life itself and is willing to risk his life to secure the authoritativeness of his self-conception, it becomes a struggle to the death.

If in such a struggle for recognition, one of the combatants out of fear for his life submits to the other's demand for recognition, he then enters into the relationship of servitude to that other, who now becomes the 'master.' The vassal then comes to judge the validity of his own desires, tastes, and plans in terms of how they fit into the master's point of view. Or, as Hegel put it, the vassal's own *self-relation* is now mediated through an 'other.' Through the discipline of

his work for the master, the 'vassal' comes to distinguish more clearly between his own subjective point of view and a more impersonal, normative point of view represented by the master. The master's point of view represents the 'totality' in terms of which the vassal must orient himself. However, as only a particular point of view whose sole claim to authority has to do with its ability to force itself on the vassal, it cannot ultimately sustain any normative allegiance to itself. As the vassal reflects on his status, he comes to understand the sheer contingency of the master's hold on him, and the master comes to understand that the recognition he requires from the vassal is only compelled, not the free recognition he requires.

Self-consciousness in Hegel's scheme thus only comes about through struggles over mutual recognition. To be self-conscious in Hegelian terms is thus not to have some inner sensation of oneself or merely to be consciously alive; it is rather to locate oneself and others (and similarly be located by others) in a complex 'social space' of norms, commitments, entitlements, and the like, and for both parties to act in light of that mutual self-locating and being located. This conception of self-consciousness as emerging only in mutual recognition and of its being the result of a struggle turned out to be one of Hegel's most enduring and widely interpreted legacies. Although others had often made claims about the inherently social nature of man, Hegel was the first to make this claim about the very structure of reasoning and agency itself. Each agent needs the other not to help him satisfy some given desire but to *be* an agent at all. Agency is thus itself a reciprocally conferred social *normative status*, not a metaphysical *fact* about people: to be an agent is to be the kind of organic creature that binds himself to and is held by others to epistemic, practical, even religious and aesthetic *responsibilities*. We become agents, almost paradoxically, by treating each other as agents.

The rest of the *Phenomenology* presented a historical account of how various modes of sociality—of complex reciprocal holding of self and others to basic authoritative norms—have failed to be able to sustain themselves and became thereby unlivable, compelling their participants to come to recognize that they had over time failed to sustain their own status as agents and (without intending it) had come to require of themselves a different form of collective life. In such transitions, what had seemed like an 'absolute' point of view had turned out to be incoherently unlivable or to be only the point of view of some particular group, and its capacity to sustain allegiance to itself unraveled. In those terms, Hegel reinterpreted not only the great transition from the ancient to the modern world but also certain particular historical transitions such as that from ancient Greece to Rome or from the aristocratic society of the *ancien régime* to the moralistic society of postrevolutionary Europe. Each such transition came to be understood as part of the

development of reason itself, of what counted as a basic norm and therefore what counted as agency itself. This was in Hegel's account not the development of a latent capacity somehow already present within us but which was only gradually actualizing itself nor was it the result of any kind of natural law compelling such historical movement, as it were, from the outside; it had to do with the inherent difficulties involved in binding each other to certain types of norms and in the way in which we came to be required to hold ourselves to different norms—that is, in *being* certain types of agents—and in the specific ways being those types of agents had failed and quite unintentionally resulted in different collective ways of being agents.

In his own lifetime and until today, people have tended to read Hegel's philosophy of history in the *Phenomenology* as a kind of story about how some appropriately grand absolute reality (identified with mindedness itself, *Geist*) gradually develops itself and actualizes its own potentiality by using various world-historical individuals and nations as a way to follow its own destined path of maturation. However, Hegel himself argued that the sense of destiny in history had nothing to do with laws of history or any such grand metaphysical scheme but with the ways in which human agents can understand the kind of agents they concretely *are* only in light of the kinds of options that their history has opened up for them. The realization that that there is no grander show behind the curtain of appearance pulling the strings for us, either in the form of historical laws (metaphysical or material) or in the form of gradual actualizations of longstanding latent potentialities, is itself distinctly modern and, as an *unsurpassable* form of self-understanding, as marking perhaps *the* decisive event in human time, is itself *absolute*. It also represents, so Hegel thought, the human community coming to a more adequate collective self-consciousness about the kinds of norms to which it had committed itself and to which it had to keep faith; and, quite famously, Hegel held that art, religion, and philosophy are to be understood as historically embedded social practices whose underlying point has to do with the manner in which they embody a way of life's concrete reflection on who they have come to be and on the matters that have come to be authoritatively valuable for them.

The different ways of taking Hegel's thought also apply to the different interpretations given of his mature work in social and political philosophy, the 1820 *Elements of the Philosophy of Right*. Although the book was unfairly criticized both during and after Hegel's lifetime as a mere 'apology' for Prussian absolutism, it was actually an attempt on Hegel's part to articulate the rational form of the kind of reformed, modern European state that modern reformers like Baron von Stein, and later, Prince von Hardenberg, had tried to establish in Prussia. The core idea of the book is that what counts as 'right' in general is what is necessary for the realization of freedom.

Hegel thought of freedom, though, not as the exercise of some special form of causality (as Kant had done) but as a kind of 'self-relation,' as having one's actions come from oneself (always though in terms of such self-relations being mediated by relation to like-minded others); and, quite radically, Hegel held that for 'we moderns,' freedom involves being held to norms that are rationally redeemable. The opposite of such freedom is to be pushed around by factors or considerations not really our own but stemming from something else that we cannot rationally redeem (for example, brute desires or mere social conventions). In Hegel's almost paradoxical way of putting it, one can be an agent only by being recognized as an agent; the conditions of agency are the social conditions under which such agents can reciprocally hold themselves and each other to being certain kinds of agents.

For agents to be free, they must be able to practically reason about their activities, and that requires that they have some conception of some 'good' that they are seeking to actualize. Hegel wished to argue that all genuine practical reasoning has as its major premise some statement about what is ultimately good and best, and when an individual acts rationally, he acts on the basis of some (again, perhaps sometimes implicit) deliberation about what is necessary for him to achieve what is good and best for himself.

For 'we moderns,' that good must be that of 'freedom,' but stated so abstractly, it seems to offer little guidance for deliberation on what would be required to achieve such 'freedom.' To that end, Hegel argued, in the modern world, the realization of freedom must be articulated into three more determinate spheres, which he characterized as 'abstract right,' 'morality,' and 'ethical life' (*Sittlichkeit*). Each of them embodies a way in which very specific institutions and practices underwrite and sustain the ways in which our freedom is actualized in that each of them provides individuals with more concrete, specific first premises about 'the good' (freedom) on the basis of which they may then rationally deliberate what they are required to do.

If we are to have any concrete first principles for moral reasoning, therefore, we must grasp them not as specifications of some 'master rule' but as elements of a social *practice*, ways in which we prereflectively learn to orient and move ourselves around in the social world. Hegel called this sphere, 'ethical life,' *Sittlichkeit*. 'Moral' individuals exercising their 'abstract rights' to life, liberty, and property require a location in these kinds of social practices, which embody within themselves determinate conceptions of what is 'ultimately best,' namely, as the way in which individuals exercise their rights, manage their moral obligations, and come to be 'at home' in the social world by virtue of acquiring a kind of 'ethical virtuosity' in being brought up and socialized in these practices. *Sittlichkeit* thus provides us with determinate principles and a kind of practice-oriented ethical 'know-how,'

and such institutions and practices form what Hegel called 'concrete universals.'

There are three such institutionalizations of *Sittlichkeit* in the modern world, each serving to give individuals a concrete specification of these final goods upon which they can then rationally deliberate. These are the modern family, civil society, and the constitutional state. Together they form a social 'whole' in terms of which individuals orient themselves and which reconciles them to modern social life, gives them good grounds for believing that modern life really is, although imperfect and finite, nonetheless for the best.

Since the modern state appeals to neither God nor natural law for its legitimacy, it must appeal to some sense of what a 'people' collectively establish as rational. This drives political philosophy into a philosophy of history, which, so Hegel argued, demonstrates that from the vague intimations in the ancient eastern states that 'one' (the emperor, the pharaoh) is free, and from the Greco-Roman political conceptions that 'some' (aristocratic males) are free, history has culminated in the modern world's recognition of the principle that 'all' are free. This, Hegel argued, is the *meaning* of world history, and Europe has been its penultimate staging ground. European culture, with its inherent sense of 'negativity,' self-doubt, and skeptical questioning, has by virtue of the kind of culture it is, propelled itself progressively to abandon those criteria, whose lack of rationality made itself manifest until in the modern period it has come to recognize, at least implicitly, that freedom has always been its collective goal, and that such freedom can only be vouched for in an institutional setup much like the one outlined in the *Philosophy of Right*. Freedom itself is thus reinterpreted as a kind of social norm, a historical *achievement*, not the better realization of a potency that, for example, the Greeks also possessed but failed to actualize. (In a completely empty sense of 'potentiality,' such capacities were, of course, always empirically present exactly in the sense that, for example, empirical science or abstract art has always been a human possibility. The kind of 'potency' that is being realized is more like that involved in a narrative than it is like an orthodox Aristotelian realization of a potentiality.)

After his death, in the rather tense atmosphere in Berlin in the 1830s and 1840s, there was a lot of pressure to interpret Hegel in either conservative or progressive terms. The 'July Monarchy' in France and the subsequent revolutionary upheavals in other parts of Europe (along with social unrest throughout Germany) made the ruling authorities anxious while raising the hopes of the reformers. There was a line of thought, especially popular among liberals, to the effect that a kind of historical law was at work that mandated countries to pass through a revolutionary upheaval, establishment of a republic, years of war, and then roughly fifty years later to settle into a constitutional monarchy (first in England, later in

France); on that view, England represented the end-point of modernity, France was almost there, and progress seemed to call for the establishment of constitutionalism in Germany by revolution if necessary; many of Hegel's own students enthusiastically embraced this view of the 'law' of history (despite Hegel's own rejection of such a view). His school thus split famously into 'right' and 'left' Hegelians. The 'right' Hegelians made Hegel into more of a Schellingian and had him espousing a view of history that basically endorsed the (non-constitutional) Prussian monarchy as metaphysically destined, as the only form of social organization compatible with the necessary structure of the cosmos. The 'left' Hegelians, led first by Hegel's friend and follower, Eduard Gans, tended to stress the elements of work and recognition as accomplishing the necessary historical tasks. In the hands of later 'left' Hegelians, however, such as Ludwig Feuerbach, the question began to be raised as to whether it was necessary to have any kind of 'absolute' at all. If history was a matter of labor and recognition, and people's labors were always socially and therefore materially mediated, then why could Hegelianism not be turned into something more like sociology, or into some form of empirical social theory?

The young Karl Marx, coming to maturity in this left-Hegelian atmosphere, transformed Hegel into a materialist thinker. (He was helped in this endeavor by the deservedly forgotten H. M. Chalybäus's characterization of Hegel's thought as following the strict scheme of thesis/antithesis/synthesis, a bowdlerization that has been applied ever since.) Eschewing all talk of the 'absolute' and of 'mindedness,' Marx read into Hegel a kind of evolutionary doctrine of the gradual realization of human capacities that were themselves historically mediated and were themselves the achievements of labor and struggle. The material conditions of life—both its organic demands for sustenance and its social demands for the organization of labor and distribution of wealth—set the terms in which people interact with each other. Hegel's doctrine of the 'struggle for recognition' thus became shorn almost completely of its epistemological bearings and became a doctrine of materialist social reality and history. In the writings of the mature Marx, this was complemented by a theory of how human consciousness is itself socially mediated and is therefore dependent on the ways in which that society is structured, particularly in terms of the relevant forces of production in society and the patterns of ownership of the means of production. Famously, Marx also transformed Hegel's philosophy of history into a set of dialectical laws to the effect that revolutions come about when the forces of production and the relations of production come into irrevocable conflict with each other. Hegel had argued that history was not best understood as a matter of divine providence, natural law, or even as the consequence of the decisions of

great men, but rather as a kind of broad scale drama of people collectively trying to come to terms with the people they had come to be in terms of the norms to which they had collectively bound themselves; Marx transformed that into a broad scale story about the ways in which human societies are organized around key material matters such as the production of the necessities for organic and social life, and he held that these follow a law-like pattern in a manner analogous to natural events.

Like Hegel, though, Marx thought that one had to understand social and political phenomena as a whole, even if he ascribed more or less independent importance to the facts about ownership of the means of production; and like Hegel, he put great emphasis on how a given social formation understands itself—what it basically values and counts as authoritative for itself, even though there too he ascribed a more or less independent value to the economic organization of society in determining how such communities arrive at this kind of collective self-consciousness. Marx's own ideas about the historical destiny of the proletariat, of course, went far beyond anything to be found in Hegel's thought. (Later 'Western' Marxists such as Georg Lukacs tried to synthesize Marx's and Hegel's alternative understandings of historical development.)

Simultaneous with Marx's materialistic development of Hegel's thought, there were ongoing 'liberal' developments of Hegel's views. Lorenz von Stein (whose impact on Marx is difficult to prove but who developed certain key 'Marxist' ideas before Marx had developed them himself) developed a kind of liberal post-Hegelian theory of society and history in the 1840s and 1850s through a series of historical writings on the social and political developments in France from 1789 to 1848. Stein argued that 'society' (*Gesellschaft*) had come to assume a prominent role in modernity such that one could best understand the way in which a people comes to reflect on itself not in terms of a 'people' or a 'nation' but in terms of the way in which the 'social' has in modernity separated itself from the 'familial,' the 'political,' and even from 'religion' to make itself central to modern life. The basic norm of freedom had historically generated a sense of individualism that, together with the emerging industrial revolution, made 'social' matters (and, more specifically, economic matters) absolutely central to modern forms of self-understanding. In particular, notions of 'possession' and 'property' assumed all the more importance in the self-understanding of such modern individuals; in 1842 Stein argued that a new social class, the 'proletariat,' had thus emerged within modern 'society,' and the tensions emerging between propertied owners and unpropertied proletarians was to be the central social issue of modern times; since the propertied classes and families would quite naturally seek to take over state power for themselves, it was crucial to construct Hegel's notion of the political state around an idea of a socially oriented state that would

secure the fair distribution of the basic goods of society for all its members. (Like Hegel, Stein also thought constitutional monarchy was consistent with such a state.) Retaining the Hegelian emphasis on the 'totality' and the ways in which collectivities understand themselves, Stein nonetheless transformed Hegelian idealism into a kind of quasi-empirical sociology and sociological theory of history; the dynamics of social change and contemporary social structures were to be understood historically in terms of the different modes of self-understanding at play in them; the crucial role of the economy in modern self-understandings was not a timeless natural law (as with Marx) but the result of the kind of self-understandings that modern individualism and industrial society had given rise to. (The next step, although Stein himself did not take it, was to transform Hegel's notion of social practices of family, society, and state into a theory of social *roles*. Hegel's own conception of agency as itself a reciprocally conferred normative status opened the way to understand social action in terms of the more theatrical notion of role-playing, since for Hegel there is no self 'behind' or 'beyond' the deeds that express the norms to which we mutually bind ourselves.) Hegel's own more philosophical worries about how, if all knowledge is socially embedded, the thinker could be in a position to know that or to truthfully assert it were also absent in Stein's transformation of Hegelianism into sociology.

In sum: for Hegel, mindedness, *Geist*, formed a sphere distinct from nature without itself being any kind of non-natural thing; indeed, *Geist* was, for Hegel, no *thing* at all but instead a socially mediated normative status. Hegel's view, moreover, toward the incipient social sciences of his own time was, on the whole, positive; he seemed to think, for example, that Adam Smith and the other Scots had made economics into a science. There is no reason for a Hegelian to deny that there are many interesting causal facts or correlations about societies that can be discovered and established only by empirical investigation. (Thus, the post-Hegelian distinction between the 'understanding' appropriate to the human sciences vs. the 'explanation' appropriate to the natural sciences would be viewed by Hegelians as yet another inadequately dualistic approach.) However, Hegel's key idealist notion—that we best understand societies by understanding how they collectively interpret themselves and the kinds of authoritative values to which they bind themselves and with how those acts of mutual binding must be understood historically—remains forever at odds with all doctrines that posit that some set of basic, natural facts fully determine how we are to understand the course of human events (such as the overall material geography of a place, as the vastly influential historian, Ferdinand Braudel, has argued). Such natural facts, while clearly forming the context in which such collective attempts at self-interpretation occur, cannot fully determine those interpretations. To understand

those interpretations, we must understand why that way of life came to regard those specific norms and not others as binding for them, and to understand that, we must understand how such commitments emerged as necessary in light of the way in which past attempts at collective self-definition had proved themselves to be unlivable. Underlying that is the 'Idea,' as Hegel put it, of freedom, of humanity realizing its destiny in a way of life whose members could be at home in the social world they create, although of course never under the conditions chosen by themselves. For Hegel, all social science only made sense as a smaller part of a larger story about modernity's coming to terms with the radical sociality of our own mindedness and with our very modern awareness of modernity's decisive rupture in human time.

See also: Cognitive Theory: ACT; Emotion in Cognition; Evolution of Cognition: An Adaptationist Perspective; Functionalism, History of; Hegel, Georg Wilhelm Friedrich (1770–1831); Human Cognition, Evolution of; Marr, David (1945–80); Phenomenology in Human Science; Phenomenology: Philosophical Aspects; Situated Cognition: Origins

Bibliography

- Adorno T W, Horkheimer M 1991 *Dialectic of Enlightenment* [trans. Cumming J]. Continuum, New York
- Hegel G W F 1956 *Philosophy of History* [trans. Sibree J]. Dover Publications, New York
- Hegel G W F 1971 *Werke in zwanzig Bänden*. [eds. Moldenhauer E, Michel K M]. Suhrkamp Verlag, Frankfurt, Germany
- Hegel G W F 1991 *Elements of the Philosophy of Right* [trans. Nisbet H B; ed. Wood A W]. Cambridge University Press, Cambridge, UK
- Honneth A 1995 *The Struggle for Recognition: The Moral Grammar of Social Conflicts* [trans. Anderson J]. Polity Press, Cambridge, MA
- Lukacs G 1971 *History and Class Consciousness: Studies in Marxist Dialectics* [trans. Livingstone R]. MIT Press, Cambridge, MA
- Marcuse H 1960 *Reason and Revolution: Hegel and the Rise of Social Theory*, 2nd edn. Beacon Press, Boston
- Marx K 1959–1974 *Werke*. Dietz, Berlin
- Pinkard T P 1994 *Hegel's Phenomenology: The Sociality of Reason*. Cambridge University Press, Cambridge, UK
- Pinkard T P 2000 *Hegel: A Biography*. Cambridge University Press, Cambridge, UK
- Pippin R B 1997 *Idealism as Modernism: Hegelian Variations*. Cambridge University Press, Cambridge, UK
- Stein L von 1921 *Geschichte der Sozialen Bewegung in Frankreich von 1789 bis auf unsere Tage*. Drei Masken Verlag, Munich, Germany
- Stein L von 1971 *Proletariat und Gesellschaft* [ed. Hahn M]. Wilhelm Fink, Munich, Germany

T. Pinkard

Hegel, Georg Wilhelm Friedrich (1770–1831)

1. Biographical Data

Hegel, Georg Wilhelm Friedrich the philosopher was born on August 27, 1770, in Stuttgart (Germany), and died on November 14, 1831, in Berlin. In 1788, Hegel, son of a medium-grade civil servant, entered the Protestant Seminary in Tübingen, where he won the friendship of his roommates Hölderlin and Schelling. He held a ducal scholarship for studying Theology and Philosophy at the University of Tübingen. After graduating in 1793, he earned a living as a private tutor in patrician families, first in Bern, Switzerland (1793–1796), and then in Frankfurt (1797–1800), where he regularly met Hölderlin. At the beginning of 1801, Hegel joined Schelling in Jena and qualified for lecturing with his habilitation thesis *On the Orbits of the Planets*. While at Jena (1801–1806), Hegel published his *The Difference Between Fichte's and Schelling's System of Philosophy* (1801). In the years 1802 and 1803, his essays *The Relationship of Skepticism to Philosophy*, *Faith and Knowledge*, and *On the Scientific Ways of Dealing with Natural Law* came out in the *Critical Journal of Philosophy*, which he copublished with Schelling. The major work of this period was, however, the *Phenomenology of Spirit* (1807), which was intended to be the first part of his system of philosophical science.

After the Battle of Jena, Hegel, having no income, took over as editor of a newspaper in Bamberg in 1807 and one year later he became principal of a secondary school in Nuremberg. There, Hegel married Marie von Tucher with whom he had two sons. During the Nuremberg period, Hegel published his *Science of Logic* in three volumes (1812, 1813, and 1816), the first of which, the *Doctrine of Being*, reappeared in heavily revised edition in 1832. In 1816, Hegel was finally granted the privilege of earning a secured income as a full professor in Heidelberg. In 1818, owing to the personal intervention of Minister Altenstein, he became Fichte's successor in Berlin.

From 1816 on, Hegel published—apart from several minor writings—the *Encyclopedia of Philosophical Science* (1817 in Heidelberg, and in two greatly enlarged editions 1827 and 1830) as well as his *Elements on the Philosophy of Right* (1821). His lectures on Esthetics, the History of Philosophy, and the Philosophy of History and Religion were not only attended by students but also by members of the Berlin society hungry for education. Until the mid-1940s, Philosophy in Germany could not be exercised without taking a position on Hegel's philosophy. Moreover, his philosophy influenced the formation of those special sciences, which Dilthey later called '*Geisteswissenschaften*.' Soon after Hegel's unexpected death, his school split into different wings as a consequence of

philosophical disputes on religion. In 1837, David Friedrich Strauss classified this splitting into the left (e.g., Karl Marx), the right (e.g., Friedrich Wilhelm Hinrichs), and the central (e.g., Karl Rosenkranz).

2. Aspects of his Philosophy

2.1 Interests of the Young Hegel

The young Hegel studied the ancient authors, the literature of the Enlightenment and above all Kant's critical philosophy (see *Kant, Immanuel (1724–1804)*), Rousseau's works on cultural criticism and theory of politics (see *Rousseau, Jean-Jacques (1712–78)*) as well as the dispute on Spinoza, which had been initiated by F. H. Jacobi. Furthermore, he enthusiastically paid close attention to the course of the French Revolution. At that time, he mainly worked on three questions. Firstly, how must a national religious education be conceived for it to strengthen the motivation for ethical and political action? This question had strongly suggested itself after Rousseau had failed to come to a convincing result concerning the question: which type of religion can be legitimated from the political point of view of a rational community? In his works on religion and ethics, Kant had then shown that a conflict is avoidable between religious and ethical duties on the one hand and religious and political duties on the other. Therefore, what mattered was to translate the idea of such a union of all those duties into an educational program.

Hegel, Hoelderlin and Schelling devoted themselves to finding a corresponding way of life, which they called 'invisible church.' Related to these interests is the second, historical question: what explains the fact that Christian faith developed into a 'positive,' authoritarian religion, which became a tool for despotism to misuse? It was the strategic behavior of the scientific community that gave rise to the third question. In the mid-1990s, the theological Orthodoxy of Tübingen began to misuse Kantian philosophy in an attempt to justify anti-liberal church dogmas in a new manner. This led to the question: how can Kant's critical philosophy be protected argumentatively from such a dogmatic misuse? From his penetrating critique, first of Kant's doctrine of postulates, emerged a system of philosophy in the end whose position may be called 'speculative idealism.' But it did not replace his original interests. Indeed, this system was to provide for their answers.

2.2 Phenomenology of Spirit

The *Phenomenology of Spirit* (PS) exhibits the 'way of despair' which 'natural consciousness' has to go in

order to justify its knowledge under the conditions set forth by the conscious contrast constitutive of itself: its own subjectivity—made into a first principle—in opposition to the objectivity not reducible to it. On the way, it becomes clear that the knowledge intended by various shapes of consciousness does not satisfy its own, immanent truth criteria. For example, it is regarded as impossible to guarantee the intended immediate knowledge of perceptual data by means of indexical references alone ('sense-certainty'). This self-examination is the first step of the procedure. In a second step, 'we,' who contemplate consciousness philosophically, have to identify this negative result—taken positively—with a new shape of consciousness. The self-examination is not complete until the contrast of subject and object has been overcome and the contents of the consciousness fulfill its truth criteria. Hegel calls this ultimate shape of consciousness 'absolute knowing,' which is at the same time an undeveloped knowing of the Absolute.

Apart from the much discussed chapter 'Lordship and Bondage'—which exhibits how two people self-consciously struggling for recognition end up developing the specific shape of intersubjectivity of lordship and bondage—the chapters on practical reason, spirit, and religion of the PS should be of some interest to the social scientist. The chapter on *Reason* exhibits that any effort of self-actualization—i.e., gaining pleasure, making inner laws universally valid, and spreading virtue—must necessarily fail. Even a self-conscious practical reason cannot reach its goal of self-actualization if it merely advances the concealed Good of the world instead of rearranging the world in the first place.

The chapter on *Spirit* contains the doctrine of the basic convictions of a society concerning the purpose of society or its central elements. The internal problems of occidental societies—exhibited in their historical order—transformed them into one another, so that—in them—the Spirit reached new basic convictions about itself. The relevant problems, however, are not those which an objective observer of a later stage in history can discover, but those which the members of the societies in question could experience themselves.

In the chapter on *Religion*, Hegel carries out a similar analysis for the religions in history. As Max Weber after him, Hegel believed that the religions above all—*qua* their internally developed convictions—promoted a process of rationalization in world history (see *Weber, Max (1864–1920)*), a process which Hegel comprehended as an advance in the consciousness of freedom. For him, Christianity was the religion, which first gave effect to the principle of subjectivity and thereby pushed through the consciousness that all men as men are free. At the end of the PS it is supposedly possible to expose a concept of cognition based on the historical experiences which natural consciousness has made in the course of

history of occidental societies and religions. This concept, which Hegel called ‘absolute knowledge,’ has the process of these experiences as its contents and the type of rules concerning the transformation of the shapes of consciousness as its form. It is the purpose of this knowledge to promote humanity, which, in turn, Hegel determines to be the process ‘to press onward to agreement with others’ at the end of the preface to the PS. The knowledge arrived at in the PS, however, so far only concerns the historical genesis of the present as an epoch which has to bring out humanity. In order to satisfy this demand, knowledge of what emphatically *is* has to be acquired, too. (Preface to the Philosophy of Right.) Since the PS has shown that certain fundamental concepts are the central constituents of each shape of consciousness, Hegel conceives of a doctrine of such ‘determinations of thinking’ in his *Science of Logic*: It has to examine whether these determinations are true for themselves.

2.3 Science of Logic

To logically unfold the absolute knowledge developed in the PS is what Hegel declared to be ‘proper metaphysics.’ It is its task to settle an adequate concept of philosophical recognition and thereby of itself (‘absolute Idea’). To fulfill this task it primarily criticizes the aims at knowledge of pre-Kantian metaphysics and Kant’s transcendental philosophy. In the *Science of Logic* (SL), which is this metaphysics, the traditional concepts of Metaphysics and Logic incapable of defining a concept of the Absolute are improved in an iterated process: abstract understanding begins with determining the traditional concepts in a way that makes them as distinct as possible from other concepts and hence suitable to define the Absolute. In a second step, dialectical reasoning demonstrates that no concept, however well determined by understanding, can be distinguished even from its opposite, and demands to skeptically suspend any judgment in the face of this negative result. But then, speculative reason identifies this negative result with a new term of the history of philosophy, which the understanding tries to determine again (see Sect. 2.4). Since the results of the first two steps are included in the new concept, the concepts of the SL become richer in content in the process. Once dialectical reason cannot find an opposite concept anymore, the process reaches its end. Hegel calls the ultimate concept ‘absolute idea,’ which is at the same time the concept of the philosophical recognition carried out in the SL (Fulda 1991).

In the final step, this concept is arrived at by overcoming the dualism of our theoretical epistemic attitudes (‘Idea of Truth’) and our practical ones (‘Idea of the Good’). In these attitudes of finite recognition—paradigmatically conceived of by Kant as theoretical and practical reason—we regard the

world as being a well-structured touchstone of our knowledge, but also as insignificant and worthless in the face of the Good which lies only in the will of the perceiver. The concept of the absolute idea is intended to deliver us from this conflict between our references to the world. But in order not to give rise to a new conflict between the finite knowledge (prescientific and of the special sciences) and philosophical recognition, Hegel required translation rules mediating between both types of attitudes (see Sect. 2.4) and obliged philosophical recognition of the Real to demonstrate the correspondence of its results to ‘empirical appearances’ (see Sect. 2.4). According to Hegel, a philosophy of the Real also depends on integrating new expressions from other sciences to denote the developed concepts. To that extent recognizing the Real is underdetermined by the SL, and a philosophy of the Real as well as Logic itself depends on the status quo of human knowledge.

2.4 Philosophy of the Real

The *Encyclopedia of the Philosophical Science in Outline* was meant to instruct the students who attended Hegel’s lectures, and therefore only contains the rudiments and basic concepts of all philosophical sciences. The logic outlined above makes up the first part of the *Encyclopedia*, the second and third parts being the philosophy of Nature and Spirit. It is primarily through these two parts that we have access to Hegel’s philosophy of the Real.

2.4.1 Philosophy of Nature. In his philosophy of Nature Hegel tries to answer the question of how we can overcome the duality of our theoretical and practical interests concerning Nature and thereby satisfy our needs without turning Nature—and thereby aspects of ourselves—into the Other per se, which would merely become controlled by technology. This philosophy looks for a third way between the modern objectification and quantification of phenomena in the natural sciences on the one hand and a romantic subjectification of Nature on the other. In doing so it does not, however, regard itself as competing with the natural sciences, but rather tries—on the basis of the results of scientific research—to conceive of Nature as an idea in which we can finally recognize ourselves.

2.4.2 Philosophy of Spirit. In so far as Hegel gains his concept of Spirit from his doctrine of living natural beings, which contrasts the sphere of Spirit from that of Nature, he tries to overcome Descartes’ dualism between *res cogitans* and *res extensa* in his whole

Philosophy of Spirit, without, however, reducing the phenomena of one sphere to those of the other.

In particular, Hegel teaches in his *Doctrine of the Subjective Spirit* how the subjective spirit of individuals rises step-by-step from its embodiment (Anthropology) and its object-boundedness (Phenomenology) towards higher cognitive and volutative competencies which culminate in the knowledge of and will to its own freedom (Psychology). In the first part, the *Anthropology*, the phenomena to be comprehended are e.g., dreaming, habits, aging, gestures, and mental disorders, while the *Phenomenology* treats the different levels of consciousness of objects. Sensory consciousness, perceptual consciousness, and comprehending consciousness are all subsumed under the heading ‘consciousness,’ since these forms of consciousness do not imply any knowledge the subject has of itself. Under the heading ‘self-consciousness’ are reviewed those forms into which some knowledge of the subject of consciousness of itself necessarily enters (desire, social approval). Finally, ‘Reason’ is the consciousness of objects, in which it is comprehended that knowledge of general objects is only possible when the subject understands that the functions of his subjectivity constitute these objects in the first place. In conclusion, the higher cognitive performances (intuition, recollection, imagination, memory, and thinking) and volutative ones (practical emotions, instincts, arbitrary acts, and striving for happiness) are discussed in *Psychology*. Within the ‘free spirit,’ the highest form of the subjective spirit, the attitude is reached and comprehended in which an individual knows of itself that it is only free and can only realize its freedom when it has overcome the dualism of interests which are merely of a theoretical or practical nature.

The *Doctrine of the Objective Spirit*, which Hegel has also worked out as *Elements of the Philosophy of Right* (PR), comprehends the objective institutions in themselves, which men who know of their freedom give themselves—right in a narrower sense (property, contract, punishment of felonies), moral norms of acting individuals who strive towards the individual or collective Good (*Moralität*), and the forms of social life in family, civil society, state, communities of states, and world history (*Sittlichkeit*). In order to accomplish this project, Hegel did not try to set up as large a set of consistent principles as possible, but instead conceived of concepts adequate to the subjects in question. The adequacy of the concepts is to be achieved by penetrating the fundamental conceptual material of effective norms and existing convictions of common sense and thus correcting the concepts.

Hegel’s analysis of civil society has been especially influential (see *Marx, Karl (1818–89)*). According to this analysis, modern society tends to accumulate great wealth in the hands of a few, alienate workers in the working processes from their products and from themselves, press large parts of the population below

the subsistence level, undergo times of over- and underproduction, form global markets, and colonize other countries (Elements of Philosophy of Right (PR)). But at the same time, civil society implements institutions to administer justice, provide for the common welfare, and unite people in trade associations and municipalities. These institutions of society and others of the state make comprehensible how civil societies can develop and stabilize themselves in spite of their inherent tendency to crises. In his doctrine of state, Hegel tried to comprehend institutions which represent the general interests of a rational community and enforce them in the face of the special needs of estates and individuals of the civil society. So the state becomes the authority which alone can guarantee the ‘concrete freedom’ of the individual of self-determination.

According to Hegel, world history, the doctrine of which makes up the final part of the PR, is constituted by the state. The principle organizing world history is the ‘world-spirit,’ the process of the developing consciousness that all men are free. In this history, several states have made—to Hegel’s mind—decisive contributions to this progress at certain times. In these Periods, Hegel thought, they have the right—and indeed, the power—to politically and organizationally establish and expand a new principle. France was an example of Hegel’s time: a country which had carried out a progress in the consciousness of freedom in its Revolution and had expanded this new order through the Napoleonic Wars in great parts of Europe. Since the doctrine of the objective Spirit did not reach the highest form conceivable of freedom, Hegel’s philosophy of spirit does not end with it, but rather with the doctrine of the absolute spirit.

Hegel’s *Doctrine of the Absolute Spirit* is a doctrine of the three leading forms of culture in which men have at all times tried to understand what is significantly true and who they are within the whole of actuality that depends on such a ‘Truth.’ According to Hegel, these forms are art, religion, and philosophy. However, they have not always had the same importance since the time they came into being. Each form rather has its particular period in which it becomes the model form to exhibit the Truth. Hegel believed that for the ancient Greeks, art was the medium through which they best understood themselves and were able to express what they cherished most. For the Christians, religion was the best form for men to represent their relation to the Truth and the Truth itself. Both forms, however, have the drawback that their means of explication are at most somewhat discursive—or, indeed, not at all. It is the purpose of philosophy as a science to conceive the content of Christianity’s true representations in an argumentative discourse and thereby to gain insight into the deepest reasons for which men come to agreement with each other and afford a stable basis for their life with each other. Therefore, Sect. 2.4 closes with a concept of phil-

osophy, which is also to justify in retrospect the recognition carried out in the system of the philosophical sciences.

3. Some Lines of Reception in the Social Sciences

Karl Marx is the most prominent recipient of Hegel's philosophy in the domain of the social sciences; therefore, see *Marx, Karl (1818–89)*.

3.1 Lorenz von Stein (1815–1890)

Lorenz von Stein was one of the most politically influential Hegelians among the theoreticians of society in the nineteenth century. With his diverse writings on modern society, he influenced central economic terms of Marx' theory (e.g., surplus value, and accumulation) on the one hand, and considerably influenced Bismarck's social policy on the other. In the sciences, however, he was almost forgotten when Positivism became dominant. It was not until the 1920s that he was studied by theorists of public and constitutional law (Rudolf Smend and Carl Schmitt), a discussion that was continued in the German debates on the principle of the welfare state after the second World War (Stein 1972). Stein took up the question posed in Hegel's doctrine of '*Sittlichkeit*': how must a state be structured for the principles of the French Revolution to be realized considering the conditions of the industrializing society? He regarded the political constellation of his times as an unavoidable antagonism between the civil society based on the satisfaction of individual needs and the state aiming at general freedom. To his mind, this antagonism leads the propertied to try to occupy all state functions through unequal elections, in order to prevent the socioeconomical rise of the nonpropertied. Therefore, Stein thinks it to be the duty of a legitimate state and his institutions (monarch, officials, army) to implement social reforms resulting in a rise of the non-propertied to the economical level necessary for them to participate in the process of forming a public will. In addition to this, the state should counteract the inequalities of the civil society by introducing universal and equal suffrage and guaranteeing basic rights. To this extent, Stein's doctrine of the state—like Hegel's—aims at more than just the guarantees of the liberal constitutional state, because it does not confine itself to the task of protecting the freedom of the individual from the state, but instead goes further to protecting the political freedom of the individual against the inequalities of the civil society through the state.

3.2 St. Louis Hegelians

This American philosophical school was founded by William Torrey Harris (1835–1909) and Hans Conrad

Brokmeyer (1828–1906). After the Civil War, this school was institutionalized by the foundation of the *St. Louis Philosophical Society* (1866) and the *Journal of Speculative Philosophy* (1867), the first philosophical journal in the USA. In contrast to the Transcendentalists at Yale and many of the German intellectuals, questions concerning religion did not play an important role for the St. Louis Hegelians. Instead, questions concerning social justice in a modern society were central for them. In addition to this, another one of Hegel's basic interests became important again, on which, however, Hegel had only remarked in passing in his systematic works: how must a pedagogy be in order to make the individuals growing up able to competently participate in forming modern society? Besides being influenced by Hegel, Denton Snider (1841–1925) and Thomas Davidson (1840–1900), two St. Louis Hegelians, were also influenced by the pedagogy of the German Friedrich Froebel (1782–1852) when they realized several educational projects: the *Communal University* in Chicago, the *Chicago Kindergarten College*, the *Goethe School* in Milwaukee, and the *Breadwinner's College* in New York City, which was devoted to the education of the working class. In 1889, Harris became the United States Commissioner of Education. Brokmeyer played a key role in the Constitutional Convention of Missouri in 1875, which guaranteed the right to education for all aged six to 20. Besides these important roles of the St. Louis Hegelians in the formation of the modern educational system of the USA, the *Journal of Speculative Philosophy* offered a platform for the publications of important American philosophers such as John Dewey, William James, Charles S. Peirce, and Josiah Royce. Beyond offering this platform, they successfully animated the study of Hegel's philosophy (Flower and Murphy 1977).

3.3 Wilhelm Wundt (1832–1920)

Wundt is commonly regarded as the founding father of Psychology, which established itself as a science around the beginning of the twentieth century. In 1879, Wundt founded the first psychological laboratory of the world in Leipzig, Germany, where he mainly studied sensations and feelings by employing experimental methods. This foundation also became influential outside Germany due to the many visits of foreigners—especially American students and scientists (see *Hall, Granville Stanley (1844–1924)*). Although Wundt often mocks at the methodology of Hegel's philosophy, he explicitly praises Hegel's achievement of 'granting the right to exist to all 'things in themselves,' whether they are to count as metaphysical borderline concepts or as practical postulates' (Wundt 1911, p. 736). Without this basis, it would have been impossible for Wundt to pursue Psychology as an empirical special science, because the soul would have been withdrawn from the applic-

ability of methods borrowed from the natural sciences, as long as it were conceived of as an independent substance removable from the body (Descartes) or were postulated as such on practical grounds (Kant). Hegel, on the other hand, 'has regarded the individual soul as the immediate relationship of the experiences of the soul, as a piece of that infinite reality of the world spirit, which still draws only on its very own actual reality to obtain all that it means' (Wundt 1911, p. 736). So Wundt arrives at the conclusion that Hegel 'had also declared the liberty of Psychology when it proclaimed for all that all mental coming into being—and, therefore, all happening in the soul as well—is actuality, is immediate experienced reality, and that essence and appearance of the Spirit are one and the same, and only mean something different in so far as we comprehend the essence as the relationship of the appearances which is correctly recognized' (Wundt 1911, p. 737; own translations) (see for more *Wundt, Wilhelm Maximilian (1832–1920)*).

3.4 George Herbert Mead (1863–1931)

The posthumously published book *Philosophy of the Present* (1959) primarily promoted an interpretation of Mead's thinking that considers the last phase of it as on the way towards a speculative metaphysics of Hegel's kind. This is surely exaggerated. It is nevertheless possible to prove that Mead named Hegel as an inspirer of his own thoughts throughout his working period. Hegel's philosophy especially stimulated the scientific program of the young Mead. His early outline of a program of philosophical disciplines puts forward two main tasks of Psychology, one of which was inspired by Hegel's PS: Psychology is to give 'an analysis of the objective world in terms of the consciousness of the objective individual' (Mead 1900, p. 12). In pursuing this task, Hegel—according to Mead's view—e.g., proved that the organization of different characters, which constitute 'social objects,' comes from 'our mutual relations' (Mead 1903, p. 106). In contrast to the explaining and describing Psychology, Logic is a normative science (Mead 1903, p. 107), into which 'the general theory of the intelligent act as a whole would fall,' and which is as such of Hegel's kind (Mead 1900, p. 2). However, apart from this praise, Mead also continued to criticize Hegel's doctrines throughout his life: e.g., he falsely subsumed Hegel under the philosophy of romanticism, which, to his mind, only discussed the rationalization of bygone attitudes—quite contrary to the orientation towards the future of the American society and philosophy. Nevertheless, Mead praises Hegel's philosophy of history as '[t]he most grandiose attempt to combine the three great motives of speculation (that of Greek contemplation, that of the church's plan of salvation, and that of experimental science)' (Mead 1964, p. 505) (see *Mead, George Herbert (1863–1931)*).

3.5 Juergen Habermas (1929–)

In his essay *Can Complex Societies Form a Rational Identity?* (Habermas 1976), which Habermas chose as the basis of his speech on the occasion of accepting the Hegel Award of the city of Stuttgart, Germany in 1974, he explicitly classifies his proposal as standing in the tradition of Hegel's and Marx' thinking (p. 118). Although his interpretation of Hegel's answer to the question posed in the title of his essay is affirmative in many points, Habermas mainly criticizes—firstly—that for Hegel, 'only the idea of freedom' is to be conceived as a constituent of modern consciousness. In addition to this, 'the unrestricted objectifying thinking and a radical orientation towards the future' are also constitutive for it (p. 105). Therefore, Hegel did not succeed in abolishing the threefold division between the modern subjectivity and external nature, society, and inner nature (p. 102) in *all* respects. Habermas objects, second, that modern society did not gain its rational identity in the constitutional state; and he objects, third, that philosophy is unable to exhibit this identity as rational for all individuals—as religion has done in the past (p. 106 ff.). It is Habermas' opinion instead that modern societies could gain their rational identity in 'the consciousness of the general and equal opportunities of participation in leaning processes of forming values and norms' (p. 119) by realizing communicative structures 'by themselves,' in whose 'autonomy' alone a new collective identity can crystallize 'today' (p. 120; own translations). Critical remembrance of tradition, science, philosophy, and art propose rival identity projections (p. 121). However, Habermas seems to have partially taken back his criticism of Hegel with his concept of 'constitutional patriotism' (Habermas 1990). In addition to this, his interpretation of the rational function of Right to deliver moral discourse (Habermas 1996) can be understood as a continuation of Hegel's thoughts on the relationship of '*Moralität*' and '*Sittlichkeit*.'

See also: Civil Society, Concept and History of; Civil Society/Public Sphere, History of the Concept; Dialectical Thinking, Psychology of; Hall, Granville Stanley (1844–1924); Kant, Immanuel (1724–1804); Marx, Karl (1818–89); Mead, George Herbert (1863–1931); Phenomenology: Philosophical Aspects; Rousseau, Jean-Jacques (1712–78); Weber, Max (1864–1920); Wundt, Wilhelm Maximilian (1832–1920)

Bibliography

Flower E, Murphy M G 1977 The absolute immigrates to America: The St. Louis Hegelians. In: *A History of Philosophy in America*. Putnam, New York, Vol. 2

- Fulda H F 1991 Philosophisches Denken in einer spekulativen Metaphysik. In: Pätzold D (ed.) *Hegels Transformationen der Metaphysik*. Dinter, Cologne, Germany
- Habermas J 1976 Können komplexe Gesellschaften eine vernünftige Identität ausbilden? In: *Zur Rekonstruktion des Historischen Materialismus*. Suhrkamp, Frankfurt a.M, Germany
- Habermas J 1990 *Die nachholende Revolution*. Suhrkamp, Frankfurt a.M, Germany
- Habermas J 1996 *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Polity Press, Cambridge, UK
- Hegel 1968 ff Gesammelte Werke. Hrsg. In: *Verbindung mit der Deutschen Forschungsgemeinschaft von der Nordrhein-Westfälischen Akademie der Wissenschaften*. Felix Meiner, Hamburg, Germany
- Hegel 1969a *Science of Logic*. Translated by Johnston W H (int. al.). Allen and Unwin, London
- Hegel 1969b ff *Werke*. In: Moldenhauer E, Michel K M (eds.) *20 Bände und ein Registerband*. Suhrkamp, Frankfurt a.M, Germany
- Hegel 1979 *Phenomenology of Spirit*. Translated by Miller A V and introduced by Findlay J N. Clarendon Press, Oxford, UK
- Hegel 1983 ff *Vorlesungen: Ausgewählte Nachschriften und Manuskripte*. Felix Meiner, Hamburg, Germany
- Hegel 1988 *Aesthetics. Lectures on Fine Art*. Translated by Knox T M. Clarendon Press, Oxford, UK
- Hegel 1990 *Encyclopedia of the Philosophical Science in Outline*. Behler E (ed.). Continuum, New York
- Hegel 1991 *Elements on the Philosophy of Right*. In: Wood A W (ed.) Cambridge University Press, Cambridge, UK
- Hegel 1995a *Lectures on the History of Philosophy*. Translated by Haldane E S. University of Nebraska Press, Lincoln, NE
- Hegel 1995b ff *Lectures on the Philosophy of Religion*. Hodgson P C (ed.). University of California Press, Berkeley, CA
- Mead G H 1900 Suggestions toward a theory of the philosophical disciplines. *Philosophical Review* 9: 1–17
- Mead G H 1903 The definition of the psychical. *Decennial Publications of the University of Chicago*. First Series, Chicago, Vol. III, 77–112
- Mead G H 1959 In: Murphy A E (ed.) *The Philosophy of the Present*. Open Court Publishers, La Salle
- Mead G H 1964 *Philosophy of the Act*. Morris C W (ed.). The University of Chicago Press, Chicago
- Stein V L 1972 *Gesellschaft-Staat-Recht*. Forstthoff E (ed.). Propyläen Verlag, Frankfurt a.M, Germany
- Wundt W 1911 *Grundzüge der physiologischen Psychologie*. Leipzig [*Principles of Physiological Psychology*. Titchener E B (ed.). Sonnenschein and Macmillan, London 1904]

M. Wildenauer

Hegemony: Anthropological Aspects

1. Introduction

By the time anthropologists discovered Antonio Gramsci's idea of hegemony in the late nineteen eighties, it had been reworked by others so that it

barely resembled Gramsci's. Although anthropologists cite Gramsci's *Prison Notebooks* (1971) when they refer to hegemony, most are inspired by revisionist interpretations of hegemony.

2. Hegemony

Hegemony is the central idea among many by which Antonio Gramsci established a humanistic, neo-Marxist approach to revolutionary change. Instead of subordinating the superstructure of ideas to the force of the economic base, Gramsci empowered the influence of ideas over Marxism's economic determinism. At the core of Gramsci's methodology is the dialectical relationship between hegemony (consent–support) and domination (coercion–force). His primary theoretical concern was how to resolve this contradiction, for in political practice one is always tempered by the other. He tried to reach resolution in part by redefining hegemony.

Lenin defined hegemony as domination. Gramsci redefined it as an *intellectual and moral leadership* directed by contradictory political and cultural agents and organizations he called *organic* and *traditional* intellectuals. Organic intellectuals represented the interests of working classes. One of their tasks was to make the proletariat aware of capitalist domination and exploitation. Traditional intellectuals represented the interests of the bourgeoisie. One of their tasks was to quell counter-hegemonic resistance to capitalism.

More than other Marxist thinkers, Gramsci recognized the significance of culture in revolutionary practice. Gramsci defined culture as 'the exercise of thought, the acquisition of general ideas *the habit of connecting cause and effect* ... enlivened by organization' (Kurtz 1996a, 1996b). Culture was central to Gramsci's political strategy. He argued that culture was a political product of intellectuals, for another of their tasks was to develop the people's culture through education and ideological indoctrination. Traditional intellectuals worked to convince people of the virtues of capitalism and obtain their consent to capitalist rule. Organic intellectuals worked to help people develop an alternative to a capitalist culture. Among other strategies, Gramsci suggested how organic intellectuals might use popular cultural forms and signs to marshal working class revolutionary energies, depose capitalism, and construct a unitary proletarian culture.

3. Revisionists

Of the many works that explore Gramsci's thinking (Kurtz 1996a), anthropologists rely largely on those of Williams (1977) and Laclau and Mouffe (1985).

Williams is concerned with the relationship between culture and hegemony. He argues that hegemony goes beyond culture, is synonymous with culture, serves as a form of 'cultural domination,' and, as a structure of cultural forms, provides a repository for traditional values. But, by equating hegemony with tradition, Williams divests it of Gramsci's political specificity.

Laclau and Mouffe are concerned with the heterogeneous social organization of contemporary capitalist nations. They equate hegemony with ideology grounded in discourses by which oppressed categories in capitalist societies—women, gays, ethnic categories, and the like—can be made aware of their oppression. For Laclau and Mouffe hegemony (read ideology) provides the 'organic cement' of political formations that will transform western capitalism into a 'radical socialist democracy.'

These authors appropriated Gramsci's idea of hegemony and made themselves its authority, yet little that they say corresponds to Gramsci's idea of hegemony. They also ignore the role of intellectuals in political practice. Still, the beauty and bane of their interpretations are the extent to which hegemony becomes amenable to diverse interpretations by anthropologists, but sometimes at the cost of the political intent with which Gramsci imbued it.

4. *Anthropology and Hegemony*

Anthropologists were attracted to the idea of hegemony for several reasons. It provided a neo-Marxist approach that complemented the idealist orientation among anthropologists who emphasize the power of ideas over material forces; it provided a new way to explore political relationships; it enabled reconsideration of old ideas, such as culture. Above all, it was thought provoking. Consequently there is little consistency regarding how anthropologists use it.

Most often anthropologists use hegemony as a trendy buzz-word that dilutes the concept's power. The idea of hegemony becomes more complicated when anthropologists use it seriously to analyze political relations. Only rarely do anthropologists use Gramsci's definition of hegemony to explore relations of domination and subordination. Instead, most anthropologists cite Gramsci, if only in passing, as the source of the idea of hegemony. Then they relegate his idea of hegemony to a secondary status and make it a point of departure to use or develop other interpretations of hegemony (Kaplan and Kelly 1994, Kurtz 1996a).

As a tool for analysis, anthropologists most often equate hegemony with a political force in society which usually emanates from those in power. Some anthropologists use revisionist ideas of hegemony to focus on other of Gramsci's ideas, such as *common sense* (peoples' awareness of their condition) (*Amer-*

ican Ethnologist 1993). However they approach hegemony, anthropologists infer that a hegemony of the proletariat to counter domination is rare (*American Ethnologist* 1993, Kurtz 1996b).

When anthropologists explore the relationship between hegemony and culture, it is Williams' ideas of hegemony and culture, not Gramsci's, that informs these explorations. Others use Laclau and Mouffe's idea of hegemony as ideology in conjunction with Williams' ideas of hegemony as culture and tradition to rework Gramsci's ideas. In these instances anthropologists become the source and authority of new ideas of hegemony. Almost no anthropologist places either hegemony and domination or traditional and organic intellectuals in the dialectical context that Gramsci thought to be important in hegemonic encounters.

This diversity of application of hegemony by anthropologists can be explained by their frequent argument that Gramsci did not define hegemony clearly. It is true that Gramsci's *Prison Notebooks* are fragmentary and open to interpretation. But anthropologists also do not read the *Notebooks'* informative footnotes, Gramsci's preprison writings, or the authoritative exegeses of his work. As a result they privilege others' interpretations of hegemony. Regardless how anthropologists define and locate hegemony theoretically, it is most applicable to two related concerns: the consequences of state and colonial policies and political relations of domination and subordination.

4.1. *State-Colonial Policies*

Only a few anthropologists apply Gramsci's formulation of hegemony to explore the impact of colonial policies. Carstens (1991) used Gramsci's idea of hegemony to account for how unequal relations between dialectically opposed intellectuals, British missionaries and Okanagan Indian chiefs, changed Okanagan culture and relegated the Indians to an inferior economic status in contemporary Canada. Kurtz (1996b) extrapolated from Gramsci's idea of hegemony to develop the idea of *hegemonic cultivation* to explain why political leaders strive to instill an ideology of work in their political communities and why resistance to these efforts is rare.

More commonly, analyses rely on Williams' idea of hegemony as cultural domination. Fox (1989) used Williams' idea of hegemony as culture to explore how cultural meanings arose in colonial India and compelled colonizer and colonized to conform to certain behaviors. Contrary to Williams, hegemonic agents are central to Fox's idea of hegemony. He demonstrates how intellectuals continually experiment with existing cultural meanings which both enable and constrain resistance to colonial rule. But hegemony is not a source of revolutionary change for Fox. Instead,

as Williams' cultural domination, hegemony continues to subordinate behaviors to the meanings it encodes.

In analyzing cultural confrontations in colonial South Africa, the Comaroffs (1991) emphasize Williams' ideas of hegemony as tradition, and Laclau and Mouffe's idea that discourse has the power to transform indigenous culture. For the Comaroffs, hegemony represents a tradition which, through discourses, acquires an ideology that helped to reorganize indigenous South African societies in accord with colonial rule. Ideology, not hegemony, provided the force of change. As tradition, hegemony had power only to shape actions, perceptions, and changes to the extent that it provided the source of new ideological challenges to existing traditions. As a process, hegemony (read tradition) is reproduced by its interplay with ideology into a functionalist 'hegemonic order' that provides a repository from which cultural forms may be created through the discursive power of ideology.

4.2 Domination and Subordination

Anthropologists who analyze the impact of state and colonial policies on anthropological subjects do not ignore the relationship between resistance and domination. Domination and subordination are implicated in almost all application of hegemony. But anthropologists who emphasize these relations explore them largely in post colonial and authoritarian societies, such as Sri Lanka and Bolivia, through ideas of hegemony that are largely of their own making. They agree only that hegemony is always a process.

Among these explorations hegemony is perceived as a *synthesizing principle* for welding class interests into a coherent world view (McDonogh 1991). In another context hegemony is a *social process* embedded in discourses aimed at instilling an ideology of nationalism (Brow 1988). The persistence of patron-client relations is maintained by hegemony identified as a *force* that maintains a social structure through the cultural shaping of experience (*American Ethnologist* 1993). Hegemony, conceived as a *historical moment*, explains how a ruling class shapes perceptions of ideas, such as female beauty, to legitimate their rule over the working class (*American Ethnologist* 1993). Some anthropologists load the idea of hegemony with an array of theoretical considerations. In one instance, hegemony functions as the source of *education* for Gramsci's idea of 'common sense' (Woost 1993). In another it represents an ongoing *formative process* that may promote resistance to domination (*American Ethnologist* 1993).

Most anthropologists who consider the role of agents in these relations emphasize the role of elites, and argue that proletarian agents are not very effective in developing counter-hegemonies. Even those few anthropologists who emphasize the role of working

class intellectuals in confrontations with elites conclude that proletarian intellectuals have little impact on changing the sources of rule and domination. As Laclau and Mouffe suggest, this is because the proletariat is too alienated and submerged under a dominant culture and rule. A logical consequence of this tendency is to privilege hegemony as tradition and render it apolitical.

Although anthropologists use hegemony as understood by others than Gramsci, and often in more convoluted fashion, they tend to agree on what could only be disheartening to Gramsci: the hegemonic impact of colonial and some contemporary governments has sufficiently restructured the culture and society of subordinate peoples that counter-hegemonies to domination are difficult to attain. Hegemony, however conceived, seems to accommodate best the interests of the elites.

5. The Future of Hegemony in Anthropology

The concern of anthropologists with the idea of hegemony peaked in the early 1990s. Now hegemony is firmly ensconced in anthropological discourse and theoretical orientations concerned with culture, structuralism, a poststructuralism *cum* postmodernism, even a revised historical materialism that objectifies ideas as material forces. Regardless how anthropologists conceive hegemony, it holds a privileged place among a cadre of non-materialist, neo-Marxist anthropologists and provides them a humanistic methodology to explore novel cultural and political relations. The collapse of the Western European and, more recently, Soviet empires and the energy of the expanding world capitalist order have produced global disjunctions in social, cultural, economic and political relations. To explain these epiphenomena, anthropologists are likely to continue to develop creatively the explanatory power of the idea of hegemony.

See also: Civil Society/Public Sphere, History of the Concept; Class Consciousness; Gramsci, Antonio (1891–1937); Hegemony: Cultural; Marxism in Contemporary Sociology; Marxist Social Thought, History of

Bibliography

- American Ethnologist* 1993 20 (1)
- Brow J 1988 In pursuit of hegemony: representations of authority and justice in a Sri Lankan Village. *American Ethnologist* 15: 311–27
- Carstens P 1991 *The Queen's People*. University of Toronto Press, Toronto
- Comaroff J, Comaroff J L 1991 *Of Revelation and Revolution*. University of Chicago Press, Chicago
- Fox R G 1989 *Gandhian Utopia*. Beacon Press, Boston

- Gramsci A 1971 *Selections from the Prison Notebooks*. Lawrence and Wishart, London
- Kaplan M, Kelly J D 1994 Rethinking resistance: Dialogics of 'disaffection' in colonial Fiji. *American Ethnologist* **21**: 123–51
- Kurtz D 1996a Hegemony and anthropology: Gramsci, exegeses, reinterpretations. *Critique of Anthropology* **16**: 103–35
- Kurtz D 1996b Hegemonic culture and work in state formations. In: Claessen H J M, Oosten H G (eds.) *Ideology and the Formation of Early States*. Brill, Leiden, The Netherlands
- Laclau E, Mouffe C 1985 *Hegemony and Socialist Strategy*. Verso, London
- McDonogh G W 1991 Culture and categorization in a turn-of-the-century Barcelona elite. *Cultural Anthropology* **6**: 323–45
- Williams R 1977 *Marxism and Literature*. Oxford University Press, Oxford
- Woost M D 1993 Nationalizing the local past in Sri Lanka: Histories of nation and development in a Sinhalese village. *American Ethnologist* **20**: 502–21

D. V. Kurtz

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

Hegemony: Cultural

The concept of 'hegemony' has provided one of the most influential theoretical frameworks in contemporary cultural and media studies, and Marxist intellectual work in particular. While imprisoned by Mussolini's Fascist regime in the 1920s and 1930s, the Italian Marxist Antonio Gramsci developed the concept of hegemony to address the question of why the working class failed to make a communist revolution in the advanced capitalist countries despite Marx's prophecy. By 'hegemony', Gramsci (1971) intended to conceptualize the pivotal role of cultural persuasion in acquiring, exercising, maintaining, or challenging power without resorting to physical coercion. Stated briefly, the concept of hegemony refers to a historical process in which a dominant group exercises 'moral and intellectual leadership' throughout society by winning the voluntary 'consent' of popular masses.

1. The Concept of Hegemony

The concept of hegemony can be more effectively understood when it is compared to Gramsci's notion of 'domination,' as each term represents a different kind of power. Whereas domination rests on a dominant group's coercion of the less powerful, hegemony relies on a dominant group's ability to procure subordinate groups' consent to the existing social order. Domination belongs to the state, which Gramsci called 'political society,' and is equipped with instruments of coercion such as the police and armed forces. In contrast, hegemony takes place in the realm

of private institutions—including those outside of the spheres of economic production and the state, such as families, churches, trade unions, and the media—which Gramsci called 'civil society.' Gramsci attributed the political stability and obedience of the working class in Western European countries to the hegemony built into the vigorous civil societies of those countries. Consent and coercion always coexist, but typically, dominant groups in contemporary capitalist societies rule through hegemony unless severe social and political crises require directly coercive interventions.

1.1 The Concepts of Consensus and Historical Bloc

The essence of the concept of hegemony lies in the idea of subordinates' consent. According to Gramsci, the consent of subordinates is constructed when a dominant group's particular world view is accepted by the subordinate groups as their world view, and thus the dominant group's interests appear to be the interests of the society at large. As a result, the dominant group is seen as entitled to its ethical and intellectual leadership and arises as a hegemonic group in a particular historical period. Moral and political authority legitimated by popular consent characterizes hegemony as power from below, distinguished from power wielded by coercion or by manipulation from the top down.

However, the paradox of subordinates' consent to their subordination in the long run raises suspicions about the nature of popular consent (Lears 1985). Indeed, some critics of the concept of hegemony have questioned how subordinates can give consent to what will adversely affect their interests unless they are self-deceived in false consciousness. The very idea of 'false consciousness' threatens to undermine the concept of hegemony in at least two ways. First, if one accepts the idea of false consciousness as the basis of popular consent, subordinates' consent is the outcome of sheer manipulation, and Gramsci might be accused of elitism for his implicit assumption that the popular masses are dupes. Alternatively, if one rejects the idea of false consciousness by arguing that people know the purpose of their actions and consistently base their actions on rationally determined self-interest, then the consensus may represent genuine or true consensus. Then, the characteristic of hegemony as dominance over subordinates would be invalidated, and their actual grievance or resistance may slip from view.

Both Jackson Lears (1985) and Stuart Hall (1996a) responded to this dilemma by rejecting the binary opposition of false and true consciousness in understanding the nature of consent. They instead embraced Gramsci's idea of 'contradictory consciousness' grounded in 'common sense.' According to Gramsci, common sense as everybody's spontaneously arrived

1.2 Hegemony Cultural Resistance and Social Change

ff

ff

ff

ff

ff

ff

ff

ff

ff

ff

a priori

ff

2. Applications of the Concept of Hegemony in Cultural and Media Studies

1.3 The Implications for Theories of Culture

2.1 The Analysis of Journalism

British Cultural Studies

ff

Media and Social Movements

2.2 The Analysis of Texts

ff

ff

ff

ff

ff

2.3 Studies of Subculture and Popular Culture

ff

British Cultural Studies

British Cultural Studies Audiences

Audiences

ff

2.4 *The Analysis of Audience*

teleno ela

identities (Martín-Barbero 1993) (see *Soap Opera/Telenovela*).

Gramsci's concept of hegemony has been continually reinterpreted and applied to different areas of cultural and media studies, depending on the immediate social and cultural questions of theorists in specific social and historical contexts. However, all research relying on the concept of hegemony shares a common pursuit in criticizing culture as a source of control, and, simultaneously, in envisioning it as a source of resistance.

Bibliography

- Condit C M 1994 Hegemony in a mass-mediated society. *Critical Studies in Mass Communication* 11(3): 205–30
- García Canclini N 1995 *Hybrid Cultures: Strategies for Entering and Leaving Modernity*. University of Minnesota Press, Minneapolis, MN
- Gaventa J 1980 *Power and Powerless: Quiescence and Rebellion in an Appalachian Valley*. University of Illinois Press, Urbana, IL
- Gitlin T 1980 *The Whole World Is Watching*. University of California Press, Berkeley, CA
- Gramsci A 1971 *Selections from the Prison Notebooks*. Lawrence & Wishart, London
- Hall S 1980 Encoding/decoding. In: Hall S, Hobson D, Lowe A, Willis P (eds.) *Culture, Media, Language*. Hutchinson, London pp. 128–38
- Hall S 1996a The problem of ideology: Marxism without guarantees. In: Morley D, Chen K (eds.) *Stuart Hall*. Routledge, London pp. 25–46
- Hall S 1996b Gramsci's relevance for the study of race and ethnicity. In: Morley D, Chen K (eds.) *Stuart Hall*. Routledge, London pp. 411–40
- Hall S, Critcher C, Jefferson T, Clark J, Roberts B 1978 *Policing the Crisis*. Macmillan, London
- Hall S, Jefferson T (eds.) 1976 *Resistance Through Rituals*. Hutchinson, London
- Hallin D C 1989 *The Uncensored War*. University of California Press, Berkeley, CA
- Hallin D C 1994 *We Keep America on Top of the World*. Routledge, London
- Hebdige D 1979 *Subculture: The Meaning of Style*. Methuen, London
- Herman E S, Chomsky N 1988 *Manufacturing Consent*. Pantheon, New York
- Laclau E, Mouffe C 1985 *Hegemony and Socialist Strategy*. Verso, London
- Lears T J J 1985 The concept of cultural hegemony: Problems and possibilities. *American Historical Review* 90: 567–93
- Liebes T 1997 *Reporting the Arab–Israeli Conflict*. Routledge, London
- Lukes S 1974 *Power: A Radical View*. Macmillan, London
- Martín-Barbero J 1993 *Communication, Culture, and Hegemony: From the Media to Mediations*. Sage, London
- Morley D 1980 *The 'Nationwide' Audience*. British Film Institute, London
- Rose T 1994 *Black Noise: Rap Music and Black Culture in Contemporary America*. University Press of New England, Hanover, NH

Williams R 1977 *Marxism and Literature*. Oxford University Press, Oxford

Willis P E 1977 *Learning to Labor*. Columbia University Press, New York

S. Kim

Heider, Fritz (1896–1988)

Fritz Heider was born in Vienna on February 19, 1896, spent his childhood in Graz, and received a classical education at the official state Gymnasium. Unable to serve in the army during World War I because of the loss of vision in one eye due to an injury during his childhood, he entered the University of Graz (where ironically at one point he studied depth perception). He did not have a professional goal in mind, and in the European tradition attended a wide range of lectures open to all University students. His interests eventually centered on the study of philosophy and psychology (in opposition to the pursuit of architecture desired by his more practical father).

As a student, Heider met Alexius Meinong, an influential figure in European philosophy, who suggested the topic of Heider's doctoral thesis with the obscure question: 'Why do we say we see the house, rather than saying that we see the sun, for the reflected rays of sunlight are the stimuli that actually strike the eye?' In discussing this problem, Heider distinguished between 'thing,' or the objective structure in the environment, and 'medium,' or that in the environment providing information about the object. This theme reappears in much of his work on person perception. He believed in a 'real' world, with the task of the observer being to come to know this reality. In Graz, Heider was also influenced by the psychologist Vittorio Benussi, one of the first people to publish experiments in the field of gestalt perception.

After finishing his dissertation, written during 1919, Heider worked for a year as an applied psychologist. But he soon became restless and moved to Berlin. At the suggestion of Bernussi, he made contact with the psychologists at the University of Berlin. An air of excitement permeated the Berlin Department of Psychology. The lectures in perception, especially those of Max Wertheimer, were very popular among students and intellectuals. There was a feeling that Gestalt principles would have an important influence on the development of psychology. Heider attended the lectures of Wertheimer, as well as those of Wolfgang Köhler. He found many Gestalt concepts, such as laws governing 'unit formation,' useful in his later work in person perception and interpersonal relationships. He also formed a lasting social and intellectual

friendship with Kurt Lewin, at that time a younger member of the Berlin faculty.

After Berlin came the *Wanderjahre*, a period when Heider spent a good deal of time meandering and reading philosophy, especially Spinoza and Nietzsche, as well as psychology and literature. Then, in 1927, he became an assistant to Wilhelm Stern at the University of Hamburg. In 1930, supported by a recommendation from Stern, Heider took leave from Hamburg to accept what he thought of as a one-year position with the Gestalt psychologist Kurt Koffka, who was on a research fellowship at Smith College in Massachusetts, USA. Heider's principle work was at the Clark School for the Deaf, where Grace Moore was also a member of Koffka's research group. They were married shortly thereafter and Heider remained in the USA.

Heider settled at Smith College as an Assistant Professor until 1947, where he combined research at the school for the deaf, teaching, and raising three sons. In 1947, he accepted an offer from the University of Kansas as Professor of Psychology. Roger Barker, the new Chairperson of the Psychology Department, had brought with him a group of people like himself who had been closely associated with Kurt Lewin. Then, with the aid of Beatrice Wright, Heider completed his long-awaited book, *The Psychology of Interpersonal Relations* (1958). A number of awards followed, including the Distinguished Scientific Contribution Award from the American Psychological Association. Fritz Heider died in Lawrence, Kansas, on January 2, 1988 at the age of 92. He was a gentle man with a sly twinkle in his eye, shy and introspective, with a love of art and literature, and a wide-ranging intellectual curiosity.

1. Scientific Contributions

Unlike other influential psychologists, Heider published few articles, did little experimental work, had few students, and was removed from the mainstream of American psychology. He apparently sought neither power nor fame. Nonetheless, his book *The Psychology of Interpersonal Relations* (1958), published when he was 62 and nearing retirement, had enormous impact. In this work, Heider applied the laws of the perception of physical objects and events developed by Gestalt psychologists to the perception of persons and their actions. Synthesizing object and person perception is one of Heider's important scientific accomplishments.

1.1 Balance Theory

Two central and associated thematic statements were most fully articulated in *The Psychology of Interpersonal Relations* (1958). One of these relates to the

concept of balance. Balance theory as formulated by Heider should be considered within the context of Gestalt views regarding the perception of objects. Gestaltists contended that there is a tendency toward simplicity in the perceptual system such that simple structures are preferred. For example, faces are perceived as symmetrical although this is not exactly the case, and even when a circular form is not quite closed, it is nonetheless perceived as a circle. Furthermore, parts of a system become organized and the groupings form a particular structure in an uncontrollable and inevitable manner. For example, two similar objects in close proximity to one another are joined to form a unit, as in the following example: ** ##. In the perceptual field, organisms impose particular arrangements.

Heider reasoned that object configurations, including social objects, could be classified into those that are in a balanced or in an imbalanced state. Balanced states are preferred to imbalanced states because they form 'good' figures or simple structures. For example, in a two-person system, if a likes b and b likes a, then the system is in balance. But if a likes b while b does not like a, then there is disequilibrium—imbalance exists and there is a force to change in the direction of balance. This can be brought about by changing attitudes toward the other, misperceiving the sentiments of the other, changing the attitudes of the other, and the like.

Configurations with three entities, such as those involving two persons and an object, are more complex. If, for example, a likes b as well as object x (e.g., the elected president), then if b does not like x, the situation is again imbalanced and a force will be created to bring the system into a state of equilibrium by, for example, a change in one's attitude toward b or toward x. In Heider's characteristic manner of integrating everyday observations, literary examples, and scientific analysis, he noted that many literary tragedies make use of balance principles (e.g., Romeo loves Juliet, Juliet loves Romeo, Juliet loves and obeys her parents, but her parents do not accept Romeo). An example of 'latent' imbalance, which contributed to the dramatic tension in Ibsen's *Wild Duck*, occurs when a scrupulous husband loves his wife who, unbeknownst to him, has engaged in a dishonest act for his benefit. Here we can see the relations between object perception, unit formation, the preference for particular arrangements, and social behavior.

Balance principles have been widely applied in the study of interpersonal relations, attitude change, social cognition, and other areas of research. They also provided a new approach in motivation inasmuch as motivation to change was derived from the configuration of cognitive elements, rather than from viscerogenic need states, which were involved in most studies of motivation. In accord with Gestalt principles, the whole pattern of the structure determines its motivational effects.

1.2 Causal Attributions

Causal attributions, or beliefs regarding the causes of events, were the second major focus in *The Psychology of Interpersonal Relations*. Attribution theory was even more impactful than Heider's balance ideas, and became the dominant theme in social psychology for nearly fifteen years, between 1970–1985.

Heider postulated that people are motivated to understand and to master their environment—understanding is adaptive and instrumental to future behavior, and also there is a basic curiosity and desire 'to know.' Understanding and mastery require knowledge of the causes of events.

Pursuing the distinction in his thesis between thing versus medium, Heider stated that we are often only in contact with immediate facts or raw data (the 'medium') but we search for the underlying core processes or dispositional properties to explain these facts (the causal 'things'). These perceived causes are generally enduring aspects of the world. Covariation information regarding the presence and absence of the inferred cause and the effect are important sources of causal knowledge. But when explaining the actions of persons, Heider reasoned that we often ascribe their behavior to some stable dispositional quality or trait, rather than to situational factors. The underestimation of the situation as a perceived cause of the behavior of others, and over-attribution to the person, subsequently was labeled 'the fundamental attribution error.'

Although the topics of balance and attribution have been treated separately in the psychological literature, Heider saw them as intertwined. He reasoned that balance principles guide causal attributions. Thus, we infer that a good act was done by a good person, and ascribe success more than failure to the self (the 'hedonic bias'), for in both of these examples the inferences result in balanced states or simple structures.

1.3 Scientific Method

The sources of Heider's ideas were largely observational and literary: he often relied upon 'thought experiments' in which interpersonal situations were imagined and altered in a systematic manner (e.g., assume Romeo does not like Juliet, or that Juliet does not care about her parents). The psychological consequences of these new cognitive configurations were then inferred. This method may seem atavistic given the movement of psychology toward an experimental science, but it proved fruitful for Heider.

Heider also made use of a second methodology: Examining common speech in order to capture *a priori* logical relations. Concepts such as belonging, benefit, harm, and sentiment were subject to his logical analysis. This was particularly the case regarding

perceptions concerning the causes of success and failure, where he articulated the relations between can and try as causal determinants of achievement outcomes. Because of his mental experiments (rather than manipulation of variables and responses by other), and the use of everyday language in his theoretical analysis, Heider is often labeled a 'naïve' psychologist. This characterization is not entirely correct inasmuch as his theoretical analyses embrace abstract principles that transcend any one concrete situation, providing laws that a layperson could not articulate.

2. Contemporary Relevance

As already indicated, attribution theory dominated social psychology in the 1970s. However, its influence did not end there. It remains a presence in social psychology and has made deep inroads as well in the fields of clinical psychology, motivation and emotion, cognitive psychology, educational psychology, and personality. Furthermore, it is almost impossible to discuss interpersonal relationships without making some reference to the concepts of balance or attribution. When assessing the contemporary impact of Heider, it must be recognized that this influence is now carried through more contemporary figures in social psychology who interpreted and elaborated Heider's ideas, while adding their own insights. Among the figures one generation removed from Heider are Edward Jones, Harold Kelley, and Bernard Weiner, three among the authors of the book: *Attribution: Perceiving the Causes of Behavior* (1972). More recently, other researchers two generations removed from Heider have, in turn, replaced these figures as leaders in their field.

What follows is a sampling of some of the attribution-related research areas outside of social psychology that can be traced to the foundation laid by Heider. Rather than assess the 'truth' of the various hypotheses that have been suggested, these are presented to show the presence of research activity stimulated by Heider's ideas.

2.1 Clinical Psychology

There is substantial research examining whether causal ascriptions are precursors to depression. It has been contended that when the causes of aversive outcomes are perceived as unchanging, unchangeable, and/or implicate the self (i.e., there is learned helplessness), then depression may follow. Furthermore, perceptions of the causes of dysfunctional symptoms by others may contribute to the course of the illness. Critical and hostile emotional expressions, spawned by causal beliefs that the other is responsible for the illness, affect the recidivism rates of schizophrenics as well as depressives.

2.2 Aggression and Prejudice

A number of different forms of aggression have been subject to attributional analyses. Some types of child abuse, wife battering, and marital hostility have been traced to the aggressor ascribing intentional misbehavior to the other (e.g., ‘My child refuses to obey me;’ ‘She purposely made me a bad meal’). In general, aggressive individuals tend to over-attribute intentionally to others when experiencing a negative outcome. Prejudice also is a product of (or may give rise to) different attributions for in-group and out-group members. For example, aggression by an out-group member may be perceived as controllable by others and intentional, thus producing further out-group derogation.

2.3 Coping

Following a negative life stressor, such as rape, unwanted pregnancy, or illness, there often is a search for causality (‘Why me?’). Some individuals ascribe the aversive event or outcome to their character (‘I am too impulsive’), whereas others place blame on a particular behavior (‘I was at the wrong place at the wrong time’). Those causal beliefs elicit different coping mechanisms, with the more adaptive behavioral blame resulting in active coping, whereas the less adaptive avoidance coping is in part evoked by attributions to properties of the self that are unchanging.

2.4 Motivation and Emotion

A large literature documents that attributions for achievement success and failure have important consequences on subsequent striving. Guided by the can/try distinction articulated by Heider, it has been found that attributions for failure to low aptitude are particularly debilitating, in contrast to ascriptions to insufficient effort, which promote action. In a similar manner, the belief that ability is changeable is more adaptive than the belief that one’s ability level is fixed. In addition, reactions of others are guided by beliefs about the causes of failure. Non-attainment of a goal because of lack of aptitude elicits sympathy and help from others, whereas failure due to low effort gives rise to anger and neglect.

A great deal of research under the rubric of appraisal approaches to emotion has contended that how one feels is determined by what one thinks. As intimated above, beliefs about the causes of events comprise an important aspect of this literature. For example, guilt is often preceded by the belief that one ‘could and should have done otherwise’ whereas shame often follows from the causal inference that the causes of

personal failure mark one as different from others and cannot be changed. Anger, sympathy, regret, and gratitude also are typically preceded by specific causal beliefs.

2.5 Cognitive Psychology and Psycholinguistics

Many researchers are examining the antecedents of causal beliefs, ranging from individual differences in cognitive preferences to the role of covariation information and the probabilistic co-occurrence of antecedents and consequences. Not all research attempting to understand the contents and process of causal inferences can be traced back to Heider. Nonetheless, some of this research is directly related to his thinking. Linguists have pointed out that verb forms also implicate causality. For example, the phrase ‘I admire him’ results in causality residing in the admired person, whereas ‘I hit him’ places causality in the person committing the act. Some people interested in gender issues have pointed out that journalists may influence lay causal beliefs when they state: ‘The flirtatious woman was raped by him’ as opposed to ‘He raped the flirtatious woman.’ In sum, it is evident from this very select review that, like his ideas about perceived interpersonal relations represented in balance theory, the understanding of perceptions of causality, represented in attribution theory, had a broad influence in psychology. Fritz Heider, an armchair psychologist with knowledge of philosophy and literature, and armed with theoretical acuity, has left a flourishing psychological legacy.

See also: Attributional Processes: Psychological; Authoritarian Personality: History of the Concept; Causation (Theories and Models): Conceptions in the Social Sciences; Emotion: History of the Concept; Khler, Wolfgang (1887–1967); Lewin, Kurt (1890–1947); Psychology: Historical and Cultural Perspectives; Social Psychology; Social Psychology, Theories of

Bibliography

- Heider F 1927 Ding und Medium. *Symposium, Philosophische Zeitschrift für Forschung und Ausprache* 1: 109–57
- Heider F 1944 Social perception and phenomenal causality. *Psychological Review* 51: 358–74
- Heider F 1946 Attitudes and cognitive organization. *Journal of Psychology* 21: 107–12
- Heider F 1958 *The Psychology of Interpersonal Relations*. Wiley, New York
- Heider F 1976 A conversation with Fritz Heider. In: Harvey J H, Ickes W J, Kidd R F (eds.) *New Directions in Attribution Research*. Erlbaum Hillsdale, NJ, Vol. 1, pp. 3–21
- Heider F 1983 *The Life of a Psychologist: An Autobiography*. University of Kansas Press, Lawrence, KS

- Heider F 1988 *The Notebooks of Fritz Heider*. M. Benesh-Weiner (ed.) Psychologie Verlags Union, Munchen-Weinheim, Germany, Vols. 1–6
- Jones E E, Kanouse D E, Kelley H H, Nisbett R E, Valins S, Weiner B 1972 *Attribution: Perceiving the Causes of Behavior*. General Learning Press, Morristown, NJ
- Weiner B 1995 *Judgments of Responsibility: A Foundation for a Theory of Social Conduct*. Guilford, New York

B. Weiner

Helmholtz, Hermann Ludwig Ferdinand von (1821–94)

1. Life

Hermann Ludwig Ferdinand von Helmholtz was born in Potsdam, Germany on 31 August, 1821. His father taught humanistic subjects at the local gymnasium, and encouraged Hermann's keen interests in music and the arts generally, in enjoying nature, and in philosophy. At the gymnasium Helmholtz distinguished himself in every subject of study. Although his great intellectual love was physics, his father persuaded him that there was little call for professional physicists and that, moreover, he could not afford to send Hermann to university. Helmholtz thus enrolled (in 1838) at the Königlich-medizinisch-chirurgisches Friedrich-Wilhelms-Institut in Berlin, the Prussian military medical school. The school provided full support to its students in exchange for eight years of military service.

After graduating in 1842 and serving a year as an intern in Berlin's Charité hospital, Helmholtz served nearly five years (1843–8) in Potsdam as an army surgeon. The position afforded him much free time, and he spent it publishing papers in physiology and conducting research in physics. In 1848–9, he served as Johannes Müller's assistant and as an anatomy instructor in Berlin, before being released early from his remaining military obligations in order to begin an academic career as Extraordinary Professor of Physiology at the University of Königsberg.

While at Königsberg, Helmholtz invented the ophthalmoscope (1850). It soon revolutionized ophthalmology by allowing doctors to examine the living retina, and thus diagnose the eye far more thoroughly and help treat it for disease. It also led to Helmholtz's becoming a full professor, boosting his reputation among the Prussian educational authorities, and giving him a European-wide reputation. In 1855, he transferred to the University of Bonn, where he spent the next three years as Professor of Anatomy and Physiology. In 1858, he accepted Baden's offer to

become Professor of Physiology at the University of Heidelberg. A year after arriving there, his wife, Olga, died, leaving him a widower with two young children to raise. He soon married Anna von Mohl, with whom he had three children and who helped introduce him to the world of German high society. The university built him a new institute of physiology; he served as the university's prorector in 1862–3, and, more generally, became one of Heidelberg's stellar figures during its golden era.

With the establishment of the German Reich in 1871, the Prussian authorities were determined to reacquire Helmholtz as part of their effort to turn the University of Berlin into the centerpiece of German academic life. Helmholtz felt he could not refuse the call to Berlin, where he now became Professor of Physics and Director of the Institute of Physics. Prussia built him an entirely new physics building, the largest such in Germany; his institute became a mecca for German and foreign students alike. In going to Berlin, Helmholtz also left physiology and devoted himself almost exclusively to physics. In Berlin, Helmholtz became an even more prominent figure in German academic and intellectual life. Since the early 1850s, he had regularly been called upon to give popular lectures on science, which he continued to do throughout Germany and in Britain. His home became one of Berlin's leading *salons*. He regularly served the Prussian and Reich governments as a representative at national and international scientific meetings. In 1883, the German emperor ennobled him for his services. Four years later, he was appointed the founding President of the Physikalisch-Technische Reichsanstalt (PTR), a nonacademic government institute devoted to research in physics and technology, including the setting of physical standards and testing. He spent the last seven years of his life directing the PTR and pursuing theoretical physics. He died in Berlin on 8 September, 1894.

2. Scientific Contributions Pertinent to the Social Sciences

Helmholtz's contributions to science concentrated above all on physics and physiology. He produced over 150 scientific papers, which were reproduced in his three-volume *Wissenschaftliche Abhandlungen* (1882–95); a three-part *Handbuch der physiologischen Optik* (1856–67) that analyzed and synthesized the results of physiological optics; a pathbreaking musicological treatise, *Die Lehre von den Tonempfindungen, als physiologische Grundlage der Theorie der Musik* (1863); and a two-volume collection of some twenty-five popular and philosophical essays (*Vorträge und Reden*, 5th edn., 1903) on various aspects of science. The hallmarks of Helmholtz's efforts in science included a gift for generalizing and synthesizing results;

using the methods or results of one discipline to advance another; showing the importance of instrumentation and measurement for advancing science; turning philosophical problems into scientifically addressable questions; analyzing the epistemological foundations of science; relating science and art to one another; and popularizing science. Although Helmholtz was not a social scientist, several highly selective aspects of his work in physics and physiology merit attention for their pertinence to the social sciences, above all psychology.

2.1 *The Conservation of Energy and Socioeconomic Thought*

In the mid-1840s, Helmholtz and several other scientists sought to establish an ‘organic physics,’ i.e., to explain life processes exclusively on the basis of the laws of physics and chemistry, and thereby rid physiology of the vague notion of ‘life force.’ Helmholtz centered his early scientific efforts on the problem of heat and its mechanical equivalent. In 1847, he published his epoch-making conception of the principle of conservation of force. Although he was only one of several scientists in the 1840s who conceived of the idea of force conservation in physical or physiological systems, his presentation offered by far the most systematic, mathematical, and general account, thereby helping to unite various disparate branches of physics. The work was part of his larger program of providing a mechanical interpretation of nature, organic as well as inorganic. Nonetheless, Helmholtz never maintained that human beings, whom he believed to possess free will, could be understood solely in material and mechanical terms.

During the 1850s and 1860s several physicists transformed Helmholtz’s principle of force conservation into the first law of thermodynamics. This new energy-based physics and the associated concept of ‘work’ generally inspired a number of economic and social thinkers in the second half of the century. Thanks in good measure to Helmholtz’s work, physical scientists and social thinkers alike now unequivocally recognized the impossibility of perpetual motion. Social and economic thinkers began to see both the individual human body and socioeconomic systems in general in terms of individual or societal energy production, consumption, transformation, and output. Helmholtz helped give economists a greater sense of the importance of the conservation and general circulation of energy. His work in thermodynamics also helped deepen understanding of the human body as a mechanical system, a human motor. The new thermodynamic viewpoint helped give new meaning to the terms ‘work,’ ‘labor,’ and ‘value’ (see Rabinbach 1992, Mirowski 1989).

2.2 *Nerve–Impulse Measurements*

Helmholtz’s work on the propagation velocity of the nerve impulse spoke directly to the interests of the emerging set of experimental psychologists. In 1849–50, Helmholtz conducted a series of experiments on the propagation velocity in frogs, demonstrating that the impulse was propagated at a finite, and measurable, velocity. He and others saw that reflex nervous action could also be measured in humans, and that there was a small, noninstantaneous time interval between human thought and bodily response.

As a result, the mind–body problem took on a new dimension, and a vigorous program of stimulus–response studies in physiological psychology began. By implication, Helmholtz’s work provided a measure of a subject’s reaction time to sensory stimulation, e.g., to visual or auditory cues, to voluntary movements, and to the association of ideas. Wilhelm Wundt and others built on Helmholtz’s work by constructing methods for measuring the time intervals in thought processes. Helmholtz’s work, along with Gustav Theodor Fechner’s on psychophysical experiments, and Wundt’s mark an important milestone in the emergence of experimental psychology as a field. It meant that at least some psychological phenomena could be treated quantitatively and become the subject of laboratory research.

2.3 *Physiological Acoustics and Optics*

Beyond neurophysiology, Helmholtz’s importance to the emergence of experimental psychology lay generally in his work in physiological acoustics and optics, especially in his two great books, the *Handbuch* and the *Tonempfindungen*. These two books became the standard reference works for psychologists and others. Both had a similar structure: they first gave analyses of the physical aspects of their respective subjects, physiological optics and acoustics; then physiological analyses of the respective sensations of light and sound; and finally arguments for the importance of the psychological factor as mentally integrating sensations into experienced perceptions or sounds. In particular, Helmholtz believed that unconscious inferences transformed the physical and physiological sensations received by the bodily organs and nervous systems into perceptions. From birth onwards, the individual learned through empirical experience that certain types of behavior produced certain types of responses, which, in the course of time, the mind transformed into unconscious inferences.

Helmholtz’s work in physiological acoustics and optics began with his theory (borrowed from Müller) of specific fiber energies. Müller had maintained that the stimulation occurring in each type of sensory organ is responsible for a specific sensory experience; each type of sensory nervous system is qualitatively

different from the others. Helmholtz extended this theory for vision by claiming, as had Thomas Young before him, that there were three sets of optical nerve fibers, each of which provided its own specific nerve energy that corresponded to the primary colors. Similarly, he extended Müller's theory for hearing by maintaining there were several thousand specific auditory energies. He announced his so-called resonance theory of hearing in the *Tonempfindungen*, showing that the ear is a set of resonators that analyzes homogeneous sounds into their components. Although he thought that sensory processes of hearing and seeing were in good part mechanical ones, he did not think that sensations were copies, let alone mirror images, of the external world; rather, they were 'signs' of it. Although Helmholtz's theory of specific energies was modified or rejected by some, it came to constitute part of the foundations of experimental psychology.

2.4 The Nativist–Empiricist Controversy

Helmholtz was not only one of the two principal antagonists in the nativist–empiricist controversy over the proper foundations of vision research. He also coined the labels of these two fundamental positions and did much to shape the controversy between himself and Ewald Hering, and their respective followers. Although their differences were many, the two sides mainly disagreed on whether the eye possessed and required a mind to see. The controversy began in the 1860s and lasted until the 1920s, though in several ways it has continued down to the present.

In the *Handbuch*, Helmholtz labeled Hering and his followers 'nativists,' since they claimed that spatial perception was innate to newborns. By contrast, he called himself and his followers 'empiricists,' since they claimed that spatial perception was an acquired phenomenon, one that each individual gradually learns through daily experience and which requires the mediation of unconscious inferences. The two sides also disagreed about the receptor mechanisms affecting color vision, and about explanations for contrast phenomena, adaptation, optical illusions, eye movements, the constancy of color, retinal correspondence, and so on.

The controversy helped shape the intellectual substance and disciplinary formation of physiology and psychology. Experimentally minded psychologists like Carl Stumpf, Georg Elias Müller, and Franz Brentano opposed the empiricist view insofar as it employed unconscious inferences and so harkened back to psychology's roots in philosophy. On the other hand, nativism's stress on the biological foundation and explanation of perception threatened to replace psychology with physiology. Yet nativism also emphasized the role of phenomenological experience, and that appealed greatly to a number of experimental

psychologists, including Stumpf, Hermann Ebbinghaus, Oswald Külpe, Erich Jaensch, and David Katz.

3. Impact on Psychology and Beyond

Unlike Fechner and Wundt, Helmholtz had no desire to create a new, institutionally independent discipline of psychology. His contributions and overall importance to the discipline came largely (and indirectly) from his work in physics and physiology, and from his general prestige. His contributions to the 'New Psychology'—a movement that arose in Germany between 1860 and 1910 and that was associated with names like Wundt, Ebbinghaus, Külpe, and Stumpf—consisted largely in the observational and experimental study of, and speculation about, sensations and perceptions. The New Psychology in good measure replaced the older philosophical psychology, and in so doing adopted, adapted, or extended many of its methods and results from sensory physiology and psychophysics, not least those introduced by Helmholtz.

Helmholtz's work deeply influenced and must be seen against the emergence and course of the New Psychology. During the 1860s and 1870s, a number of psychologists aligned themselves with (sensory) physiology, and it was during these decades that Helmholtz had his greatest influence on the field. For example, Wundt, who was Helmholtz's assistant at Heidelberg, shared Helmholtz's empiricist theory of perception, and Helmholtz later drew on Wundt's work on vision for his *Handbuch* and supported him professionally. As advocates of psychology as an experimental discipline, they were allies. Yet there were also differences between the two: Wundt later rejected Helmholtz's mechanistic approach to sensation, the two men had a priority dispute concerning the theory of unconscious inferences, and Helmholtz criticized several of Wundt's views. The young Ebbinghaus was similarly influenced and supported by Helmholtz. Helmholtz's work on the measurement of the velocity of nerve impulse served as a starting point for Ebbinghaus's famous work on measuring memory. Helmholtz was one of Ebbinghaus's examiners for the right to teach at Berlin, where Ebbinghaus lectured in psychology (1880–94) and where Helmholtz supported him in other ways.

In 1890, when Ebbinghaus and Arthur König founded the *Zeitschrift für Psychologie und Physiologie der Sinnesorgane*, they included Helmholtz on their editorial board. The new journal played a leading role and constituted an important milestone in physiological psychology. Yet the discipline of psychology was itself changing: as it gradually separated from physiology, a number of its practitioners gradually abandoned its elementalist, sensory-based approach in favor of the phenomenalist approach of the

Gestalt psychologists (especially after 1920), who rejected Helmholtz's mechanistic approach to psychology and, like the nativists, his theory of perception. In short, from the 1890s psychology moved away from Helmholtz's views and gained a new sense of intellectual and institutional autonomy.

Helmholtz also had a strong impact on other young psychologists, two of whom warrant special mention. As a medical student and young doctor, Sigmund Freud was especially taken with the program of the 'organic physicists.' (He had once hoped to study with Helmholtz, though nothing came of this.) Freud sought to link neuropsychological to physiological phenomena, and his early psychological language is the language of physiology (e.g., that of 'energy'). He called Helmholtz 'one of my idols.'

Likewise the young William James had intended to study with Helmholtz but never did. He called Helmholtz 'immortal' and said that he was 'perhaps the first scientific genius now [in 1868] above ground.' Helmholtz's work and views on physiological psychology constituted an important source for James: as a young scientist, he assiduously studied Helmholtz's *Handbuch*, the *Tonempfindungen*, and the popular and philosophical essays. Helmholtz was one of the most-frequently-cited authorities in James's *Principles of Psychology*. Yet although James often cited him positively, especially as a source of empirical evidence, he rejected Helmholtz's notion of unconscious inference and his sensation-based theory of perception, and he found Helmholtz's mechanistic and mathematical predilections in science alien to his own. James wrote of the *Handbuch*: 'Can I find fault with a book which, on the whole, I imagine to be one of the four or five greatest monuments of human genius in the scientific line? If truth impels I must fain try, and take the risks. It seems to me that Helmholtz's genius moves most securely when it keeps close to particular facts.'

Beyond psychology, Helmholtz also affected several of the future leaders of the next generation of social science in ways that are perhaps impossible to state. The young Franz Boas had originally intended to become a physicist and to study with Helmholtz, but he soon turned away from physics and towards geography before settling on anthropology. (Helmholtz did, however, serve as one of Boas's examiners when he [successfully] sought to obtain the right to teach at the University of Berlin.) The young Wilhelm Dilthey was a frequenter visitor to the Helmholtz home in Berlin. Although Dilthey's holistic and dynamic theory of consciousness stood in sharp contrast to Helmholtz's sensationalist approach, he studied Helmholtz's writings carefully; considered Helmholtz's popular scientific essays as the best such set that had ever appeared in German; engaged him in friendly discussions of their philosophical differences; and graced his *Einleitung in die Geisteswissenschaften* (1883) with an opening quote from Helmholtz.

By the 1920s Helmholtz's substantive achievements and outlook had lost their attraction for some psychologists. Nonetheless, for many in Germany and abroad, Helmholtz still stood, as he did for Max Weber in 'Science as a Vocation' (1919), as a model scientific researcher. And as late as 1950, Edwin G. Boring, in the second edition of his monumental *A History of Experimental Psychology*, ranked Helmholtz, along with Darwin, James, and Freud, as one of the four most important figures in the history of psychology.

See also: Empiricism, History of; Experimentation in Psychology, History of; Fechnerian Psychophysics; Neuropsychological Functioning, Assessment of; Psychology: Historical and Cultural Perspectives; Psychology: Overview; Wundt, Wilhelm Maximilian (1832–1920)

Bibliography

- Boring E 1950 *A History of Experimental Psychology*, 2nd edn. Appleton Century Crofts, New York
- Cahan D 1994 Helmholtz and the civilizing power of science. In: Cahan D (ed.) *Hermann von Helmholtz and the Foundations of Nineteenth-Century Science*. University of California Press, Berkeley, CA, pp. 559–601
- Cahan D (ed.) 1994 *Hermann von Helmholtz and the Foundations of Nineteenth-Century Science*. University of California Press, Berkeley, CA
- Cahan D 1995 Helmholtz als Kulturträger. *PTB-Mitteilungen* 105(4): 249–54
- Helmholtz H von 1856–67 *Handbuch der physiologischen Optik*. Voss, Leipzig, Germany
- Helmholtz H von 1863 *Die Lehre von den Tonempfindungen, als physiologische Grundlagen für die Theorie der Musik*. Friedrich Vieweg und Sohn, Braunschweig, Germany
- Helmholtz H von 1882–95 *Wissenschaftliche Abhandlungen*, 3 Vols. Barth, Leipzig, Germany
- Helmholtz H von 1903 *Vorträge und Reden*, 5th edn., 2 Vols. Friedrich Vieweg und Sohn, Braunschweig, Germany
- Helmholtz H von 1995 *Science and Culture: Popular and Philosophical Essays* (Cahan D ed.). University of Chicago Press, Chicago
- Koenigsberger L 1902–3 *Hermann von Helmholtz*, 3 Vols. Friedrich Vieweg und Sohn, Braunschweig, Germany
- Mirowski P 1989 *More Heat Than Light. Economics as Social Physics: Physics as Nature as Economics*. Cambridge University Press, Cambridge, UK
- Rabinbach A 1992 *The Human Motor: Energy, Fatigue, and the Origins of Modernity*. University of California Press, Berkeley, CA
- Turner R S 1994 *In the Eye's Mind: Vision and the Helmholtz-Hering Controversy*. Princeton University Press, Princeton, NJ
- Weber M 1919 Science as a vocation. In: Gerth H H, Mills C W (eds.) *From Max Weber: Essays in Sociology*. Oxford University Press, New York pp. 129–56

D. Cahan

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

6657

Hempel, Carl Gustav (1905–97)

Carl Gustav Hempel was a leading figure of the logico-empiricist movement that dominated much of American philosophy in the second third of the twentieth century, and has been a principal source of present day analytical philosophy. In the spirit of logical empiricism Hempel aimed at a scientific philosophy with the principal objective to arrive at a purely logical methodology of the empirical sciences, including the social sciences. His most important contribution in this respect is what can be regarded the first theory of scientific explanation. This theory is referred to as ‘the Hempel–Oppenheim model,’ the ‘covering-law model’ or the ‘standard model of scientific explanation.’ In his later years Hempel gave up the idea of a purely logical (i.e., syntactical and/or semantically) ‘rational reconstruction’ or ‘explication’ of basic methodological concepts like explanation, confirmation, law of nature, probability, type, etc. Both the internal failure of this program and the influence of Thomas S. Kuhn caused Hempel to acknowledge for a satisfying reconstruction of methodological concept also ‘pragmatic’ aspects of science in the sense of social and historical circumstances of its development.

1. Biography

Hempel was born on 8th January, 1905, in Oranienburg (near Berlin, Germany). He studied mathematics, physics, and philosophy at the universities of Göttingen, Heidelberg, and Berlin from 1923 to 1929. After graduation (*Staatsexamen*) that qualified him to teach in secondary schools (*Gymnasien*) Hempel went to Vienna to study with Moritz Schlick and, particularly, Rudolf Carnap. Both of them, together with the mathematician Hans Hahn and the economist Otto Neurath and others, had founded in 1923 the ‘Vienna Circle’ (*Wiener Kreis*), that was to become the nucleus of logical empiricism. After his return to Berlin in 1930 Hempel entered teacher’s training and teaching and worked at the same time with Hans Reichenbach, head of the ‘Berlin Group’ of logical empiricism, on a philosophical dissertation on the concept of probability. After graduation in 1934 Hempel preferred to leave Nazi Germany and went to Brussels to assist Paul Oppenheim, a private scholar, in carrying out philosophical research that related principally to concept formation in empirical science. In 1937 to 1938 Hempel worked for a year as research associate with Carnap in Chicago, and again went back to Brussels until he found a post of instructor at City College, New York in 1939. The academic year 1940 to 1941 saw him at Queens College, New York, from where he went as associate professor to Yale in 1948. In 1955 Hempel became Stuart Professor of

Philosophy at Princeton. After retirement at Princeton (1973) he worked as university professor at Pittsburgh University from 1977 until 1985. He died at Princeton on 9 November, 1997.

2. Major Contributions

Hempel has given important contributions: (a) to the rational reconstruction of methodological concepts, particularly the concept of scientific explanation; (b) to the ‘universality thesis,’ i.e., the thesis that the social sciences do not, and should not, use methods that are essentially different from the methods of empirical science; and (c) to concept formation and taxonomy in empirical science.

2.1 Scientific Explanation

Hempel reconstructs scientific explanations of concrete events in nature, history, and society as arguments, i.e. as logical deductions containing as premises (*‘explanans’*) at least one (general) scientific law L_1 , as well as initial and boundary conditions C_r , and as conclusion (*‘explanandum’*) the statement E describing the event in question. We thus obtain the following model of ‘deductive-nomological (*D-N*) explanation’

$L_1, L_2, \dots, L_n$	The <i>explanans</i>
$C_1, C_2, \dots, C_r$	[entails logically]
E	the <i>explanandum</i>

In case the *explanandum* contains essentially a statistical law (e.g., for all x : the probability of G , if F is given, equals r and as boundary condition that F is given for some b) the explanatory argument does not deliver a deduction of the *explanandum* statement E but rather gives inductive support for E , i.e., the *explanans* shows that E was to be expected with a certain degree r of probability. In simplified form (one statistical law, one boundary condition) we thus obtain the following model of ‘inductive-statistical (*I-S*) explanation’:

$p(G, F) = r$	The <i>explanans</i>
$F(b)$	[lends inductive support of probability r to]
$G(b)$	the <i>explanandum</i>

In the case of *I-S* explanation, one must further assume that the *explanans* is maximally specified, i.e., the *explanans* must contain all available information that is possibly relevant for the explanation of the *explanandum* E . Or, in other words, the *explanandum* must contain all pertinent statistical laws and those special facts that can be connected with the E -event by statistical laws. *I-S* explanations thus relate to the body of scientific knowledge at a given time, i.e., they

become ‘epistemically relative.’ The postulate of maximum specificity avoids the ‘explanatory ambiguity,’ i.e., the possibility to give acceptable *I-S* explanations based on true premises for mutually exclusive *explananda*.

In case one deduces as *explanandum* a general or a statistical law of lesser generality than the one(s) in the *explanandum* one obtains the reduction of a law or theory in the first case (e.g., explaining the law of free fall by means of the law of gravitation), and a deductive-statistical or *D-S* explanation of a statistical law by a more comprehensive statistical law in the second.

Originally Hempel claimed the structural identity of explanation and prediction. Later he had to admit that, although every adequate explanation is (under pragmatically changed circumstances) also a prediction, not every adequate prediction yields also an adequate explanation, for symptoms are often adequate for a prediction, but not for the corresponding explanation. So is, for example, a sudden fall of the barometer (together with suitable laws) sufficient for predicting a thunderstorm, but it does not explain it, since the fall of the barometer is not the cause of the thunderstorm, but rather a symptom of its arrival.

2.2 Unity of Science

Hempel claims the methodological unity of empirical science. With respect to scientific explanation this amounts to a universality thesis of scientific explanation: there is no essential difference between explanations in the natural sciences and those in psychology, the social sciences, or history. Explanations in these fields aim to show that the event in question was to be expected on the basis of antecedent and boundary conditions and general laws. In history those general laws are usually taken from psychology, sociology, economics, the natural sciences, etc. Hempel leaves the question open whether genuine historical laws exist. Normally explanations in history and sociology fail to include an explicit statement of the laws they presuppose. This failure is due to the fact that those laws are part of folk psychology and seem to be tacitly taken for granted. Apart from that it is difficult to formulate in a sufficiently exact way the underlying general assumptions about e.g., the outbreak and course of revolutions. Furthermore, most of the regularities in the social sciences are statistical. Therefore one can only expect explanation sketches in these fields.

The method of empathic understanding that is often claimed to distinguish the social from the natural sciences is neither sufficient, nor necessary for explanations in these fields. It is not sufficient, because the individual act of understanding is only a heuristic device in order to suggest general psychological hypotheses that might subsequently serve as explana-

tory principles in the case under consideration. It is not necessary, because a historian might well be able to explain the deeds of a paranoid historic personality by reference to the principles of abnormal psychology without being able to arrive at an empathic understanding of such a personality.

2.3 Concept Formation in Empirical Science

Hempel’s work on concept formation in empirical science originates in his early cooperation with Paul Oppenheim and consists basically in the application of principles of logical empiricism to the formation of scientific concepts. Hempel distinguishes: (a) classificatory; (b) comparative; and (c) quantitative concepts. These distinctions normally characterize at the same time the historical development of a science from its initial descriptive ‘natural history’ stages to its mature theoretical stages. Classification divides a given set or class of objects (e.g., human beings) into subclasses (e.g., beings that suffer from certain mental disorders). Each of these subclasses is defined by means of a certain concept (e.g., schizophrenia) that represents the complex of characteristics essential for the membership in that subclass. Hempel speaks also of ‘classificatory types.’ The elements of these subclasses are those individuals to whom the respective concept applies (e.g., schizophrenics). The characteristics that form the defining concept are ascertainable fairly directly by observation. Classificatory concepts as well as scientific concepts in general have to fulfil two requirements. They have: (a) to be objective; and (b) to have systematic import. Objectivity is attained first by using operational criteria and (often only partial) operational definitions in a large sense that includes observation as operational and second by eliminating criteria with valuation overtones. Scientific concepts have systematic import if they lend themselves to the formulation of general laws or theoretical principles. Whereas classificatory concepts are a yes-or-no affair, i.e., a certain individual belongs to some subclass or it does not, comparative concepts admit of more or less with respect to a certain trait (e.g., *x* is warmer than *y* with respect to cold-hot). They lead to (quasi)linear orderings as soon as one includes the relation of coincidence (e.g., *x* is equally warm as *y*). Hempel calls the extremes of such orderings ‘ordering types.’ The next step would consist in giving these linear orderings a metric and thus arriving at quantitative concepts (e.g., temperature).

For Hempel the ‘ideal types’ that have become popular in the social sciences since Max Weber are neither classificatory nor ordering types. Ideal types are interpretative or explanatory schemas that contain a set of empirical generalizations that establish subjectively meaningful connections between different aspects of some kind of phenomenon, e.g., purely

rational economic behavior, a capitalist society, a religious sect. Accordingly ideal types must rather be reconstructed not as concepts but as theoretical systems that are intended to provide explanations and therefore must contain testable hypotheses. Despite their lack of clarity and precision there thus are also with respect to ideal types no essential methodological differences between psychology or the social sciences and the corresponding methods in the natural sciences.

3. Influence, Impact, and Current Significance

Hempel's ideas constitute the mature form of logico-empiricist thinking. His work on scientific explanation 'stands out as the benchmark to which all subsequent studies of this topic must be referred' (Salmon 1998, p. vii). His reconstruction of concept formation in empirical science has become a sort of standard for objective procedures of concept formation. It has greatly influenced subsequent editions of the *Diagnostic and Statistical Manual (DSM)* of the American Psychiatric Association (Houts 2001). His later work reflects the openness of logical empiricism to new developments in the philosophy of science (particularly T. S. Kuhn's work) that put a strong emphasis on the historical and sociological ('pragmatic') factors of the development of science (Wolters 2000).

See also: Empiricism, History of; Explanation: Conceptions in the Social Sciences; Ideal Type: Conceptions in the Social Sciences; Idealization, Abstraction, and Ideal Types; Logical Positivism and Logical Empiricism; Mathematical Models in Philosophy of Science

Bibliography

- Hempel C G, Oppenheim P 1936 *Der Typusbegriff im Lichte der Neuen Logik. Wissenschaftstheoretische Untersuchungen zur Konstitutionsforschung und Psychologie*. A. W. Sijthoff, Leiden, The Netherlands
- Hempel C G 1952 Fundamentals of concept formation in empirical science. *International Encyclopedia of Unified Science*. University of Chicago Press, Chicago, Vol. II, no. 7
- Hempel C G 1965 *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. The Free Press, New York
- Hempel C G 1967 *Philosophy of Natural Science*. Prentice-Hall, Englewood Cliffs, NJ
- Hempel C G 1979 Scientific rationality: Analytic vs. pragmatic perspectives. In: Geraets T F (ed.) *Rationality To-Day—La Rationalité Aujourd'hui*. University of Ottawa Press, Ottawa, ON, pp. 46–66
- Hempel C G 2000 *Selected Philosophical Essays*. Jeffrey R (ed.). Cambridge University Press, Cambridge, UK
- Houts A C 2001 The diagnostic and statistical manual's new white coat and circularity of plausible dysfunctions: Response to Wakefield. *Behaviour Research and Therapy* **38**: 315–45
- Salmon W C 1989 *Four Decades of Scientific Explanation*. University of Minnesota Press, Minneapolis, MN

Salmon W C 1998 *Causality and Explanation*. Oxford University Press, Oxford, UK

Wolters G 2000 Die pragmatische Vollendung des logischen Empirismus—In memoriam Carl Gustav Hempel (1905–1997). *Zeitschrift für allgemeine Wissenschaftstheorie* **31**: 205–42

G. Wolters

Henry, Louis (1911–91)

Henry was born on February 19, 1911 and was 35 years old when he joined the newly created French Institut National d'Études Démographiques (INED) in 1946; he retired in 1973 and died on December 30, 1991. Most of his work was published under the auspices of INED, either in the journal *Population*, or in the Institute's monograph series. He was one of several retired artillery officers hired by INED because of their rigorous mathematical education at the national *Ecole Polytechnique*; others were Jean Bourgeois-Pichat and Sully Lederman, who would also leave their mark in the field, although they had no formal training in the social or historical sciences. Henry soon joined forces with Paul Vincent in the study of fertility, and this became his area of specialization. He made an early contribution to its measurement by inventing a new index, the parity progression ratio (Henry 1953). This is defined as the proportion of women with a particular number of children, or parity, who have gone on to have another child, and is computed on the basis of the number of children ever born recorded at a particular age.

Henry intended to go beyond the mere measurement of phenomena, and undertook to decompose reproductive processes into their component factors, what later became known as the proximate determinants of fertility. He decided to study the 'physiological' level of fertility, that is, the fertility that would prevail among married women or couples in the absence of deliberate efforts to limit their number of births; this he called 'natural' fertility. In a contemporary population, however, contraception and abortion that are subject to human decisions are widespread, and it becomes difficult to investigate physiological factors. In his quest to find reliable data on populations where family limitation was not practiced, Henry studied records from the European past, at a time when the recording of vital events was complete and accurate. He first used a series of genealogies on the ruling classes of Geneva, going all the way to the sixteenth century (Henry 1956); later he analyzed the parish records for a village of Normandy, Crulai, that had been extensively compiled by a local historian, Etienne Gautier (Gautier and Henry 1958). The records for Geneva demonstrated the sharp contrast between the

earliest marriage cohorts, where marital fertility by age seemed to be independent from age at marriage, and the later cohorts where the behavior of couples was clearly modified by the duration of their marriage and the number of children they had attained. It could be shown conclusively that the bourgeoisie of Geneva had adopted family limitation as early as the second half of the seventeenth century. In contrast until the end of the eighteenth century, Crulai, the first parish analyzed by the new method of family reconstitution that Henry had invented, showed the features of a regime of natural fertility, where parity does not affect a woman's subsequent behavior.

In a manual written with Michel Fleury (Fleury and Henry 1956), he laid out—to the minutest details—the rules for the processing of parish data. These rules have been followed faithfully by demographic historians, not only in France but also in other European countries such as England, where they guided the work of the Cambridge Group for the History of Population and Social Structure, and in Canada. Henry would publish another textbook on historical demography (Henry 1967). Moreover, applying the vast experience accumulated in his early parish studies, he initiated a study of rural France, using a representative sample of French parishes during the eighteenth century. He analyzed this material by aggregative methods, and for a smaller sample of parishes used nominal information to reconstitute families. The project provided the first national series of estimates of fertility, mortality, and nuptiality for a Western European country in the eighteenth century. This body of work makes Louis Henry the uncontested father of historical demography as it is practiced today.

Mining the past for data on populations that did not consciously limit their family size was initially only a tool towards measuring the physiological components of human reproduction. Henry identified the factors to be considered as fecundability, the probability that a conception would result in a live birth; the delay in the resumption of ovulation after a delivery or nonsusceptible period, and total or partial sterility. He provided estimates of the value of these parameters on the basis of his own historical studies of population, but also of various accurate measurements made on other selected populations. His work was picked up by English and American demographers and biometricians working on fertility models, and among them by Mendel Sheps and Jane Menken who edited a collection of translations from Henry's most important articles (Henry 1972.) The seminal importance of Henry's work was acknowledged by the Population Association of America when it made him the first recipient of the Irene B. Taeuber Award for Excellence in Demographic Research in 1977. Just before his death in 1991, the International Union for the Scientific Study of Population had named him its first Laureate.

The concept of natural fertility is often considered as one of Henry's most significant contributions to population studies. He did not invent the term, but he defined it rigorously and provided means to identify natural fertility regimes. It refers specifically to the behavior of married couples, and is therefore not affected by variations in the proportions married and the age at marriage. Natural fertility is compatible with behavior that may affect the levels of natural fertility (frequency of intercourse, length of breast feeding) or even with attempts to space birth that are not depending on parity. Although the notion remains controversial, it has been used widely in the study of the fertility transition, both in the West and in today's developing countries.

See also: Demography, History of; Demography: Twentieth-century History; Fertility: Proximate Determinants; Historical Demography

Bibliography

- Fleury M, Henry L 1956 *Des registres paroissiaux à l'histoire de la population. Manuel de dépouillement et d'exploitation de l'état civil ancien*. INED, Paris
- Gautier E, Henry L 1958 *La population de Crulai, paroisse normande. Etude historique*. Presses universitaires de France, Paris
- Henry L 1953 *Fécondité des mariages. Nouvelle méthode de mesure*. Presses universitaires de France, Paris
- Henry L 1956 *Anciennes familles genevoises. Etude démographique, XVIe-XXe siècle*. Presses universitaires de France, Paris
- Henry L 1967 *Manuel de démographie historique*. Droz, Geneva, Switzerland
- Henry L 1972 *On the Measurement of Human Fertility. Selected Writings of Louis Henry*. Elsevier, New York

E. van de Walle

Hermeneutics, History of

The term 'hermeneutics' refers to the science, art, or technique of interpretation, paradigmatically of written texts but by extension of human actions and other social phenomena. In this extended sense, it has come to denote a variety of approaches in the social and behavioral sciences. The word refers to the Greek messenger-god Hermes, and reflection on the problems of interpretation and criteria for the truth, validity, or adequacy of interpretations, goes back to ancient Greek thought in the European tradition and a similar historical distance in the other world civilizations. We can consider this as the prehistory of hermeneutics, revived in the philological criticism of classical texts in the Renaissance, the interpretation

of Roman Law, the interpretation of the Bible in (especially Protestant) Christianity and the philosophical analysis of texts.

The systematization of hermeneutics occurred largely in German-speaking Europe in the course of the nineteenth century, although two earlier thinkers deserve particular mention. Johann Martin Chladenius (1710–59) published in 1742 a general account of interpretation which brought together traditional Aristotelian rhetoric and more recent Enlightenment themes, developing in particular the notion of point-of-view or perspective (*Sehe-Punkt*), crucial to later philosophy of history and social theory (Mueller-Vollmer 1985). Second, Giambattista Vico (1688–1744) formulated the basic principle that our knowledge of what we ourselves have made (individually or collectively) is different from what we have not made. The world of human society and culture is in some sense ‘our’ product.

This distinction, crucial to later hermeneutics and ‘hermeneutic’ social science, between what we know from the inside and what we know because we have learned about it is taken up in a rather more speculative form in Hegel’s differentiation between reason (*Vernunft*) and the understanding (*Verstand*), although the maker here is the world-spirit coming to recognize its own productions (including, ultimately, the world itself) and its learning processes as rational, in contrast to the essentially contingent states and relations found in nature and described by the mathematical and natural sciences. Something more like Vico’s idea returns again in the second half of the nineteenth century, when Hegel’s concept of objective mind is extracted from its surrounding developmental ‘grand narrative’ and treated more as a descriptive category.

It was Friedrich Schleiermacher (1768–1834), following Friedrich Schlegel, who consolidated hermeneutics in a systematic form, establishing the term understanding (*Verstehen*), which had been used by Chladenius and which has survived as standard usage in English-language social science discourse, and making it central to interpretation—understood as a more systematic activity. Interpretation here involves both the linguistic understanding of meaning and a psychological understanding of the author’s intention. Schleiermacher introduced the distinction between technical and more developed or speculative forms of interpretation and formulated the much-discussed principle that it should be possible to understand an author better than he understood himself. Schleiermacher’s contribution was made the central and culminating point of the account of ‘the rise of hermeneutics’ given by Wilhelm Dilthey (1833–1911), with whom hermeneutics becomes central to the self-definition of what he called the human sciences or *Geisteswissenschaften*. In these sciences, as Dilthey put it, the mental activity of humans and of some other animals, and its products, can be understood.

Dilthey and his contemporary, the philosopher of history J. G. Droysen (1808–84), developed what we would now call a research program for history and the other human sciences based on the distinctiveness of human psychic expressions and the understanding of those expressions. In a move that was to become a definitional feature of later interpretive social science, Dilthey, like Schleiermacher, emphasized the continuity between everyday understanding and more formal processes of interpretation. His distinction between the natural and human sciences was developed in large part in opposition to Comtean positivism, which had become influential, even in the German-speaking countries, by the middle of the nineteenth century. In a parallel, but more methodological formulation, two other neo-Kantian thinkers, Wilhelm Windelband and Heinrich Rickert, argued that the study of culture is essentially concerned with individual processes and relating them to shared human values, whereas the natural sciences are concerned with general laws about objects which are essentially remote from questions of value. We are interested, for example, in the French Revolution, not just as a member of a class of revolutions exhibiting certain common features (this would be, for Rickert, a natural-scientific mode of approaching it), but as a unique event embodying, and perhaps also violating, certain crucial human values.

This opposition between positivism and methodological dualism, and more particularly between causal explanation, analyzed in terms of universal regularities, and ‘understanding’ comes to structure the emergent human or social sciences, as the term ‘culture’ increasingly gives way to ‘society’ or ‘sociation’ (*Vergesellschaftung*). There is a fairly strong line of influence from Rickert to Max Weber, both directly and also through his friend Georg Simmel, who discusses our knowledge of the social world in terms that foreshadow social constructionist theory (see *Constructivism/Constructionism: Methodology*).

Max Weber’s personal development from economic and legal historian to sociologist continues the encounter between hermeneutics and the social sciences which had been a dominant feature of the late nineteenth century, and he also develops a middle position on the issue of methodological dualism, as he had earlier on the related opposition in economic theory (the *Methodenstreit*) between the pursuit of a systematic laws and a more interdisciplinary and historical approach. For Weber, as he put it in an early essay, ‘the course of human action and human expressions of every sort are open to an interpretation in terms of meaning which in the case of other objects would have an analogy only at the level of metaphysics’ (Weber 1975, pp. 217–8). He therefore later defines sociology, in the first sentence of his major work ‘Economy and Society,’ as a science which aims at an interpretative understanding of action in order thereby to understand its course and its effects.

Whether by this Weber means that explanatory understanding is itself a form of causal explanation, or merely complementary to it, the crucial point for him is that explanations of social phenomena must be both 'causally adequate' and 'meaningfully adequate.'

For Weber, then, our access to knowledge of the social world is importantly different from our knowledge of nature. However, it is not, in his view, any less objective. He heroically attempts to hold together Rickert's principle that our perspectives on cultural phenomena, and our knowledge of them, are shaped by values (based for Weber on ultimately ungroundable existential choices), with the idea that the social sciences can attain a bedrock of solid and 'value-free' knowledge which would have to be accepted, as he sometimes curiously puts it, 'even by a Chinese.' Weber insists that it is the intentions and purposes (subjective meanings) of human actors which define their actions and which therefore have to be understood by the historian or sociologist, but he moves rapidly to the construction of a system of ideal types of action-orientation less limited than that found in economic theory but still substantially dependent on it.

Weber's synthesis was pulled apart from both sides in the decades following his early death in 1920. At one pole, there was now a more stridently naturalistic and indeed reductionist variant of positivism: the logical empiricism of the Vienna Circle, in whose 'unified science' the statements of all sciences should be ultimately reducible to material-object language or to statements in physics; *Verstehen* was of no more importance, in Otto Neurath's vigorous formulation, than a good cup of coffee which sustains the social scientist.

From the other direction, Alfred Schutz initiated the tradition of social phenomenology, and what is often loosely called 'hermeneutics' in the context of the social sciences, with a book published in Vienna in 1932 with the title *The Meaningful Constitution of the Social World*. Schutz felt that the problem with Weber's ideal types was not that they were insufficiently scientific, but precisely the opposite: Weber was too quick to impose them on the phenomena he described, paying insufficient attention to their grounding in acts of typification performed by ordinary members of society. For Schutz, the social scientist is merely constructing second-order typifications based on those already carried out in the lifeworld. This theme was taken up by the phenomenologist Aron Gurwitsch in the USA, where Schutz had also settled, and by Peter Berger and Thomas Luckmann (1967), who finally put the term 'social construction' on the map, offering in the guise of a sociology of knowledge a paradigm more directly adapted to use in social research. By the time Schutz's first book was republished, in Germany in 1970 and in the USA in 1967, the way had been prepared by his own later work and by Berger and Luckmann (1967).

The middle decades of the twentieth century replay many of the debates which had dominated the second half of the nineteenth century. Droysen's complaint, in a letter of 1852, about the rise of what he called 'crass positivism,' and his crusade against it in a course which he taught from 1857 onwards, is echoed by antinaturalist social science just over a hundred years later. Social phenomenology also of course had affinities with the well-established minority North American tradition of symbolic interactionism, which also experienced a certain resurgence in the 1960s with, for example, the republication of the work of G. H. Mead and that of Herbert Blumer, and also the publication of various studies by Erving Goffman. But where the interactionists tended not to spend time on formal critiques of empiricism, Harold Garfinkel's ethnomethodology (Garfinkel 1967) was more aggressive. A former student of the structural functionalist Talcott Parsons, Garfinkel was led by his study of the deliberations of a trial jury to emphasize, like Schutz, the importance of practical reasoning in everyday situations. He showed that our drive to interpret leads us to impute a meaning even to random processes in the social world, as in one of his later experiments in which an interviewer playing the role of a counsellor replied to the clients' questions by a random 'yes' or 'no,' leading them into more and more contorted interpretative efforts. The production of meaning is at the same time the production of social order—Parsons' major concern. Unlike his former teacher, however, Garfinkel insisted that social actors are not simply bearers of their social roles ('cultural dopes'), but active subjects obliged to practice social analysis in order to function in everyday society.

Social phenomenology found a somewhat unexpected ally in a convergent move within analytic philosophy. Beginning on the margins of the Vienna Circle, Ludwig Wittgenstein came to abandon the simple conception of a picturing relation between propositions and the world, and was drawn into a more sensitive and holistic analysis of the practicalities of 'language-games' based on implicit rules and embedded in what he enigmatically called 'forms of life.' An important book by the Wittgensteinian philosopher Peter Winch (1958) drew the consequences for social theory, using Max Weber, as Schutz had done, as one of the foils for his argument. For Winch, knowing a society meant learning the way it is conceptualized by its members. He thus revived the central principle of nineteenth century German historicism, according to which every age must be understood in its own terms. Winch directly identified himself with the German idealist tradition by insisting that social relations are 'like' logical relations between propositions (1958, p. 126) as well as, more concretely, with an ethnographic field-work approach.

Hermeneutic theory itself also took a new turn with the 'philosophical hermeneutics' of Hans-Georg Gadamer (1900–), whose *Truth and Method*, published in

1960 and translated into English in 1975, insists, in opposition to historicist hermeneutics, on the practical dimension of interpretation, conceived in Heidegger's sense of an 'encounter' between the 'horizon' of the interpreter and that of the text itself. Gadamer's philosophical hermeneutics is thus conceived in opposition to the methodological emphasis of traditional hermeneutic theories and their concern with the accuracy of interpretation. Gadamer's aim is to describe the underlying process, an existential encounter between two perspectives or horizons of expectation, which makes interpretation possible in the first place. Understanding is not just a matter of immersing oneself imaginatively in the world of the historical actor or text, but a more reflective and practical process which operates with an awareness of the temporal and conceptual distance between text and interpreter and of the ways in which the text has been and continues to be reinterpreted and to exercise an influence over us. This effective history (*Wirkungsgeschichte*), which traditional historicist hermeneutics tends to see as an obstacle, is for Gadamer an essential element that links us to the text. Our pre-judgments or prejudices are what make understanding possible.

Although Gadamer has often stressed that his philosophical hermeneutics, with its origin in Heidegger's hermeneutic ontology, is distinct from hermeneutics as a technique of interpretation, his approach clearly poses a challenge to more traditional conceptions of hermeneutics. These differences are brought out in particular in Gadamer's exchanges in the 1960s with Emilio Betti. The alternative conception of the human sciences or *Geisteswissenschaften* put forward in Gadamer's work also made it central to Jürgen Habermas' reformulation of the logic of the social Sciences (Habermas 1983). Habermas welcomed Gadamer's critique of hermeneutic objectivism, which he saw as the equivalent of positivism in the philosophy of the natural sciences, and also his stress on the totalizing character of understanding. For Habermas, however, Gadamer's stress on the fundamental nature of language, expressed in his claim that 'Being that can be understood is language,' amounted to a form of linguistic idealism. Together with Gadamer's stress on the importance of tradition and his rehabilitation of the category of prejudice, this suggested an ultimately conservative approach which was unable to deal with the systematic distortion of communicative processes by relations of power and domination. Habermas and Gadamer debated these issues in the late 1960s and early 1970s (Bleicher 1980); more recent theorists have tended to stress the compatibility of hermeneutics and critical theory (see *Critical Theory: Frankfurt School*) in a conception of critical hermeneutics (Thompson 1981, Outhwaite 1987). More recently, Gadamer also engaged briefly with the French deconstructionist philosopher Jacques Derrida, whose conception of interpretation is more skeptical.

Habermas' version of 'critical theory' can be seen, along with critical realism and Anthony Giddens' structuration theory, as one of three particularly influential attempts in the final third of the twentieth century to reconcile, as Max Weber had done at the beginning of the century, the rival claims of explanation and understanding in the social sciences. Habermas, along with his close collaborator Karl-Otto Apel, argued for a complementarity between an empirical-analytic approach oriented to the explanation, prediction, and control of objectified processes and a hermeneutic approach concerned with the extension of understanding, in an emancipatory model of critical social science, instantiated by psychoanalysis and the Marxist critique of ideology, which aims at the removal of causal blocks on understanding. Much of this remains in his more recent theories of reconstructive science and communicative action.

Critical theory tends towards a dualism of natural and social science, but in an increasingly muted form. The assumption that opposition to positivism also entailed dualism or antinaturalism was however also put in question in the late 1960s and early 1970s by the realist metatheory of science developed by Rom Harré and Roy Bhaskar (1975). Both Harré and Bhaskar, like Habermas, were substantially motivated by the desire to undermine positivistic theories and approaches in the social sciences. Harré and Secord (1972) developed a philosophy for social psychology based on the work of the later Wittgenstein and the analytic philosophy of language practiced at Oxford by J.L. Austin. Ordinary language, they argued, is better suited to the description of the mental processes of social actors than apparently more scientific artificial terminology, and they drew attention to models of research practice of this kind in the work of Goffman, Garfinkel, and others.

Harré and Bhaskar were in any case interested in giving a more adequate account of science as a whole, in a world composed of relatively enduring structures and mechanisms. Some of these could be isolated in scientific experimentation, given the contingent existence of *Homo sapiens* and *Homo scientificus*. An important aspect of the realist programme developed by Harré, Bhaskar, and others was a conception of explanation as involving not an essentially semantic reduction of causal statements to general laws but a reference to the causal powers of entities, structures, and mechanisms. Causal tendencies might or might not be outweighed by countervailing tendencies, and two causal tendencies may neutralize one another, as do the centrifugal force of the earth's rotation and its gravitational attraction, with the convenient consequence that human beings and other animals are safely anchored to the earth's surface.

This and other features of realism meant that the whole issue of naturalism could be rethought. Human beings could be seen as having causal powers and liabilities, just like other entities; it no longer mattered

so much that their relations rarely sustained any universal generalizations of an interesting kind, but only sets of tendencies regular enough to be worth exploring. The fact that many of the entities accorded causal force in social scientific explanations were necessarily unobservable was not, as it was for empiricism, a problem of principle. And the understanding of meaning could, as Bhaskar put it, be seen as in some ways equivalent to measurement in the natural sciences. Finally, it seemed natural to include among the causes of human action the agents' reasons for acting—reasons which must be understood as far as possible.

The realist critique of traditional epistemology found an echo in social theory, notably in the work of Anthony Giddens, who had become similarly impatient with the residues of positivist social science as well as the more radical contentions of social constructionism. Giddens' conception of the 'duality of structure' was designed to replace the traditional dichotomies between theories of social structure and social change, the micro–macro divide, and between interpretive and more structural approaches. Approaches like these which aim to mediate between pure hermeneutics and more naturalistic conceptions of social science coexist, at the dawn of the twenty-first century, with more explicitly hermeneutic or phenomenological conceptions (see e.g. Soeffner 1989, Müller-Doohm and Jung 1993). Hermeneutics in a broader sense continues to exist as a major research tradition in the humanities, as well as a minority one in the social sciences (Shapiro and Sica 1984). More importantly perhaps, social scientists that would not sign up to an explicitly hermeneutic program have at least accepted the importance of hermeneutic issues.

See also: Ethnomethodology: General; Hermeneutics, Including Critical Theory; Interactionism: Symbolic; Phenomenology in Sociology; Phenomenology: Philosophical Aspects; Schtz, Alfred (1899–1959); Semiotics; Weber, Max (1864–1920); Wittgenstein, Ludwig (1889–1951)

Bibliography

- Apel K-O, Bubner R, Gadamer H-G, Giegel H J, Habermas J 1971 *Hermeneutik und Ideologiekritik*. Suhrkamp, Frankfurt, Germany
- Berger P L, Luckmann T 1967 *The Social Construction of Reality*. Allen Lane, London
- Bhaskar R 1975 *A Realist Theory of Science*. Leeds Books, Leeds, UK
- Bleicher J 1980 *Contemporary Hermeneutics*. Routledge & Kegan Paul, London
- Chladenius J M 1742 *Einleitung zur richtigen Auslegung vernünftiger Reden und Schiffted. Mit einer Einleitung von Lutz Geldsetzer*. F Lanckisch, Leipzig, Germany [1969 Stern-Berlag Janssen, Düsseldorf, Germany]
- Dallmayr F R, McCarthy T A 1977 *Understanding and Social Inquiry*. University of Notre Dame Press, Notre Dame, IN

- Dilthey W 1964 *Die Geistige Welt*. Tenbner, Stuttgart, Germany
- Droysen J G 1974 *Historik: Vorlesungen über Enzyklopadie und Methodologie der Geschichte*. Oldenbourg, Munich, Germany
- Gadamer H-G 1960 *Wahrheit und Methode*. Mohr, Tübingen, Germany [1975 *Truth and Method*. Sheed and Ward, New York]
- Garfinkel H 1967 *Studies in Ethnomethodology*. Prentice-Hall, Englewood Cliffs, NJ
- Giddens A 1976 *New Rules of Sociological Method*. Hutchinson, London
- Giddens A 1984 *The Constitution of Society*. Polity, Cambridge, UK
- Goffman E 1959 *The Presentation of Self in Everyday Life*. Doubleday, Garden City, NY
- Habermas J 1978 *Knowledge and Human Interests*. Heinemann, London
- Habermas J 1987 *The Theory of Communicative Action*. Polity, Cambridge, UK, Vol. 1 (1984), Vol. 2: (1987)
- Habermas J 1988 *On the Logic of the Social Sciences*. Polity, Cambridge, UK
- Harre R, Secord P 1972 *The Explanation of Social Behaviour*. Blackwell, Oxford, UK
- Müller-Doohm S, Jung T (eds.) 1993 *Wirklichkeit im Deutungsprozess*. Suhrkamp, Frankfurt, Germany
- Mueller-Vollmer K 1986 *The Hermeneutics Reader*. Blackwell, Oxford, UK
- Outhwaite W 1987 *New Philosophies of Social Science*. Macmillan, London
- Schutz A 1932 *Der sinnhafte Aufbau der sozialen Welt*. Springer, Vienna [1972 *The Phenomenology of the Social World*. Heinemann, London]
- Shapiro G, Sica A 1984 *Hermeneutics: Questions and Prospects*. University of Massachusetts Press, Amherst, MA
- Soeffner H-G 1989 *Auslegung des Alltags—Der Alltag der Auslegung*. Suhrkamp, Frankfurt, Germany
- Thompson J B 1981 *Critical Hermeneutics*. Cambridge University Press, Cambridge, UK
- Weber M 1975 *Roscher and Knies*. Free Press, New York
- Winch P 1958 *The Idea of a Social Science and its Relation to Philosophy*. Routledge, London

W. Outhwaite

Hermeneutics, including Critical Theory

1. Johann Conrad Dannhauer (1603–66)

The term 'hermeneutics' is generally defined as 'the science or art of interpretation.' Although it denotes a practice which stretches back to the Greeks, modern hermeneutics emerges in the seventeenth century with the publication in 1654 of J. C. Dannhauer's *Hermeneutica sacra sive methodus exponendarum sacrum litterarum*. The significance and novelty of Dannhauer's approach lies in the fact that interpretation is for him no longer treated simply as a tool for biblical exegesis or textual philologists, but as a universal phenomenon. Dannhauer was led to this conclusion by virtue of the fact that in the seventeenth century there was an enormous proliferation of texts in all disciplines due to

major innovations in the world of printing. Consequently, critical attention was now directed toward the various worldviews which were circulating in textual form, and away from those systems of belief which had hitherto escaped the scrutiny of a less than literate public. Knowledge was being construed as an activity which sought to penetrate behind appearances neither to ultimate reality, nor to a stable and immutable order of things, but as a practical mode of engagement with the texts and writings of those who had made popular the view that there was such an order. Hence, texts which had hitherto assumed divine status—such as the Bible—were now in open competition with multifarious other books and treatises. The veracity of any text depended, from a hermeneutic perspective, upon how well both it and its author withstood interpretative scrutiny.

Such a procedure did not seek to engender confusion or to render truth and meaning inaccessible. For Dannhauer, on the contrary, hermeneutics amounted to a ‘universal science’ which enables us to appraise the truthfulness of any discipline by clearing away prejudice and falsehood. As such, getting at meaning or truth requires an approach which is never content simply to take things at face value. The master of suspicion will see meaning as something to be disclosed through a recovery of an author’s intention, and not as something which can be grasped intuitively.

Dannhauer’s universal science of hermeneutic inquiry was a rudimentary attempt to come to terms with the uncertainties regarding the status of truth, knowledge, and meaning occasioned by the dawn of the Renaissance. By giving interpretation a central epistemic role, he anticipated what was to become a major philosophical school of thought.

2. Romantic Hermeneutics: Schleiermacher and Dilthey

Although Dannhauer’s significance ought not to be underestimated in the history of hermeneutic thought, much of what comes under the headings of ‘hermeneutics’ and ‘critical theory’ today has its direct antecedents in the eighteenth and nineteenth centuries. It is to scholars of the Romantic period, such as Friedrich Schleiermacher and Wilhelm Dilthey, that one must look in order to gain an understanding of the many contemporary movements which are classified as being largely hermeneutic in origin and ethos.

2.1 Friedrich Schleiermacher

Schleiermacher and Dilthey are usually credited with having made hermeneutic inquiry a central *philosophical* preoccupation. Schleiermacher (1768–1834) sought to advance Dannhauer’s attempt at a universal

hermeneutics by trying to identify what was common to the technical operations and apparatuses of both the philologist and classical exegete. Consequently, he disregarded subject matter—whether classical document or scripture—in an effort to pinpoint the fundamental and universal techniques or methodology of the interpretative process.

Schleiermacher concluded that any act of interpretation involved both a ‘grammatical’ and a ‘technical’ form: while grammatical interpretation tries to understand a document by studying the broader linguistic milieu from which it derives, technical (stemming from ‘*techne*’—meaning artistic creativity or making) interpretation, on the other hand, looks to the author and the author’s personality and unique genius as a means of understanding the text in question. While both types of interpretation are significant for Schleiermacher, he nevertheless insists that they cannot be carried out concomitantly. For to interpret grammatically is to focus attention on language *stricto sensu*, and not on the individual speaker as such; conversely, technical interpretation demands that the reader look to the peculiarities of a particular author, to the author’s style and use of expression. In the latter case, the general language used by the author is overlooked in favor of the *way* in which the author uses it.

For Schleiermacher, the ultimate task of the hermeneutic project is served best by this technical approach to questions of meaning and understanding. Truly to understand a text, the readers must strive to reappropriate its subjective dimension. They must endeavor, that is, to extrapolate from the document the individual author’s ‘psychological’ intentions and motives. In so doing, the readers seek to come to terms with the text’s central purpose and aims. Technical interpretation allows us to tap into the soul of the work in question.

As stated above, Schleiermacher considered his way of dealing with the complexity of texts as more than simply a means of alleviating the difficulties confronted by biblical exegetes and classical philologists. He followed Dannhauer’s lead in trying to identify a hermeneutic methodology (*Kunstlehre*), which would govern not only specialized interpretative procedures, but all forms of understanding. In so doing, Schleiermacher insisted that we must always begin by presupposing misunderstanding, that misunderstanding is not something which may bedevil our reading at some future point, but that which bedevils us at every stage of the interpretative process. Hence, in the words of Jean Grondin, Schleiermacher ‘universalizes misunderstanding as the situation and occasion of interpretation’ (Grondin 1994, p. 70). The prospect of their not having fully understood a work, in other words, is what persuades the readers to read and reread the subject matter, to assess and reassess their findings. Furthermore, to presuppose misunderstanding sharpens the interpreters’ acuity, in that such an approach

ensures that they will not take anything for granted; that the readers will always admit to the possibility of being disproved.

This is why technical interpretation is, for Schleiermacher, to be preferred over and above grammatical interpretation. For if misunderstanding is to be surmounted, there ought to be as complete a reconstruction of the authorial intention as it is possible to undertake. It is only by trying to stand in the shoes of the author that we can best understand the meaning and *raison d'être* of the text. Indeed, he goes further by suggesting that we ought not only understand as well as the author, but that we must, if we are really to reconstruct a foreign viewpoint, cultivate an even more profound understanding of the work than that possessed by its author. Such is the ultimate objective of all hermeneutic inquiry for Schleiermacher.

To understand a text more thoroughly than its author seems a strange ideal for someone who emphasizes so emphatically the fundamental role of misunderstanding. It should be read, however, as simply that—an *ideal*. Due to the fact that one can never assume that all misunderstanding has been eradicated, it is the duty of the reader to purvey ever more valid and fruitful interpretations of a work. Such an ideal provokes the reader to overturn previously held interpretations in favor of more plausible alternatives. The desire to know the author and the author's world better than the author is what prevents, in Schleiermacher's opinion, hermeneutic complacency. It is what drives and inspires the inquisitive reader who cannot settle for the commonly held interpretation.

Such attempts to penetrate behind the mind of an author through a technical interpretation of that author's work caused many to charge Schleiermacher with psychologism. While such a criticism is valid to the extent that Schleiermacher's hermeneutics did evince a tendency to reconstruct a psychological profile of the author, or to highlight the thought processes which led certain authors to write as they did, it could be argued alternatively that such a procedure was a nascent attempt to crystallize what was to become the bedrock of all subsequent hermeneutics—the recreation of an imaginary dialog with those of our past in an effort to make sense of *who* we are today. Given that Schleiermacher did not have at his disposal the insights of Freud and other subsequent theorists of the 'self,' his tentative steps toward a universal hermeneutic methodology founded on a dialogical conception of identity should be seen as a brave attempt to give the role of interpretation a philosophical credibility which it hitherto lacked.

2.2 Wilhelm Dilthey

While Schleiermacher's contribution to hermeneutic inquiry was not insubstantial, Wilhelm Dilthey

(1833–1911) was a thinker whose work, and its subsequent appropriations and applications on the part of leading philosophers, has had a most profound impact on the formation of twentieth-century hermeneutics. For it was Dilthey who argued that if hermeneutics is to be genuinely universal, it must have as its object all of reality. The text of reality is history itself, as that which interconnects the multifarious and disparate strands of human endeavor. History, on this reading, is mankind's document—its grand narrative. As all texts have a place in this document, it behooves the inquirer to try to understand how history coheres before trying to understand how individual texts cohere. In problematizing history thus, Dilthey dramatically rendered more complex the job of the interpreter.

Dilthey's methodological approach was determined in large measure by the advent of positivism—a school of thought which insisted on the primacy of the natural sciences (*Naturwissenschaften*) as a model for understanding the world and our relation to it. For it was his conviction that if the human sciences (*Geisteswissenschaften*) were to fend off the challenge coming from the natural sciences, they would have to adopt analogous methodological procedures as their natural counterparts. Only in so doing could one come up with an adequate defense for the efficacy of historical knowledge as encapsulated in the various human sciences. Dilthey is, therefore, faced with the problem of how to rescue the laws of history from the laws of science, or of how to distinguish scientific *explanation* from historical *understanding* (*Verstehen*).

In his quest to pinpoint what it is that marks understanding off from explanation, Dilthey argues that a human being is not known solely as a physical object amongst others; it cannot, that is, be reduced to the level of scientifically observable phenomena. In contradistinction to scientific data, the human being has a mind which exceeds the purview of the natural scientist or the positivist, a mind which attests to its being. Consequently, a person may be understood by other people who are free to interpret and decode that person's intentions and expressions, desires and volitions. It is the mental faculty of human beings, in other words, which permits them to be *understood* and not merely *explained*. Thus, for Dilthey it is *psychology*, as the science of the human as historical and sociological agent, which ought to act as the bedrock of the human sciences. In this, he remains faithful to Schleiermacher's privileging of interpretative psychology as the most efficacious means of understanding the intentions of an other.

Unlike our knowledge of the external world, which requires the mediation of the senses, knowledge of the inner world of psychological life is free from temporal vicissitudes. As such, it forms the basis of a universally fixed mental structure which may be accessed immediately. But the quandary which Dilthey faced in advancing such a theory was quite considerable. For

example, in order to legitimate his claims he had to put forward incontrovertible proof that there were in fact complex mental structures which were indeed self-evident, and which could be accessed in a neutral way by any thinking subject. Given that there has never been agreement among the practitioners of psychology as to the precise nature of such structures, let alone proof of their existence, it is hardly plausible to assert that they are universally self-evident.

In an effort to counter such dilemmas, Dilthey elects to prioritize hermeneutics as that which has the best chance of permitting access to the inner life of the mind, by way of external media such as signs and texts. For it is through such documents that the mental life of one different to the self is made manifest for analysis by the human scientist. The inner life of an other, in other words, is objectified in and through the traces which he or she leaves behind. Consequently, the interpreter has the capacity to reconstruct the life of the other by recovering what the other has deposited in textual form. The text, therefore, is that which bears witness, in external form, to the inner mental life of one who was formerly alien to the self. It is that by way of which the interpreter, or the human scientist, can arrive at a credible understanding of a life. The ultimate aim of the hermeneutic enterprise is to reproduce the life that one is endeavoring to understand. If this could be achieved, the objective validity of the human sciences would no longer be in any doubt. It would also serve to avert the pending crisis, which historicism and relativism threatened to engender.

3. Phenomenological Hermeneutics: Heidegger and Gadamer

3.1 Martin Heidegger

Dilthey's legacy was far-reaching and extensive. In the domain of philosophy, it was Martin Heidegger (1889–1976), a German student of the eminent phenomenologist, Edmund Husserl, who would do most to build upon and advance Dilthey's hermeneutic insights. While Dilthey emphasized the role of hermeneutics in the human sciences, Heidegger was to go one step further by inquiring into the ontological conditions of such sciences. In other words, Heidegger considered the human sciences to be a branch of epistemology (Greek: *episteme*) or the theory of knowledge. As such, they ought not to be the focus of genuinely fundamental philosophical analysis. Rather, the task of the hermeneutic phenomenologist was to inquire into the actual *being* of such sciences, their ontological (Greek: *on*) status. Indeed, this move towards *being*, and away from *knowledge*, was

Heidegger's most significant contribution, not only to phenomenology and hermeneutics, but to the history of twentieth-century philosophy.

Central to Heidegger's thought was the belief that the dualism between subject (the self) and object (the world beyond the self—the 'external' world) was misguided and mistaken. Although such dualisms are a regular feature in the history of philosophy, they are nowhere more apparent than in the thought of the seventeenth-century French philosopher, René Descartes. Descartes's contention that the thinking self (*Cogito*) could doubt away its environment in an effort to establish its own certainty, was primarily responsible for concretizing the subject–object divide. Even Heidegger's philosophical master, Husserl, structured his thought around this dualism; his emphasis on the role of intentional consciousness in the formation of the self's constitution was seen by Heidegger to be yet another unfortunate consequence of the Cartesian legacy, even if Husserl's original aim was to surmount the subject–object divide by leading us back to an originary experience of the world. If Husserl emphasizes consciousness as that upon which our primordial experience of the world is founded, Heidegger, in an effort to reconcile both subject and object, interprets this originary experience as a *being-in-the-world*. In so doing, Heidegger replaces the question common to Descartes, Kant, Dilthey, and Husserl—'What does it mean to *know*?'—with the question, 'What does it mean to *be*?' For Heidegger, there is no way the subject can, or should, try to extricate itself from the world in which it finds itself—from its natural environment. For it is through our basic interaction with the world that we come to a genuine and fundamental understanding of the 'who' of being.

The question of being, the question which first stimulated the early Greek metaphysicians, is thus what Heidegger intends to retrieve. He does so by underscoring our *being-there* (*Dasein*), our temporal situatedness *in* the world, as distinct from our consciousness *of* the world. Due to the fact that a human being is the only being capable of asking the question: 'What does it mean to be?', any examination of being must therefore proceed by way of an inquiry into the human self as a being-there (*Dasein*), as a being who is concretely submerged in the practical affairs of the everyday world. This helps to explain the title of Heidegger's *magnum opus* of 1927, *Being and Time* (1962).

One such way Heidegger approaches the question of the being of *Dasein*, is through an analysis of the category of 'understanding.' For Heidegger, 'understanding' does not denote conscious awareness or assimilation, but rather those 'moods' (*Stimmungen*) which characterize our basic lived experience. These moods take the form of concern, anguish, fear, guilt, excitement, etc., and are determined by Heidegger as ontological acts of *preunderstanding* (*Vor-Verständnis*), or the means by which we come to grips with the

world prior to the process of making the world an object for thought. In other words, Heidegger's phenomenological analysis strives to bring the moods of *Dasein*'s preunderstanding to the level of a reflective self-awareness. Like his nineteenth-century predecessor, Søren Kierkegaard, Heidegger turns Descartes's dictum—'I think, therefore I am' (*Cogito ergo sum*), on its head. For both Kierkegaard and Heidegger, we exist before we are objectively aware that we exist; our existence is preunderstanding, in the sense that we prereflectively interpret the world as a project of possibilities for our existence, before coming reflectively to understand it as such.

Heidegger's philosophical hermeneutics, unlike the methodological hermeneutics of Dilthey, will not, therefore, reflect upon the human sciences, but will attempt an examination and explication of the very foundations of those sciences. Moreover, for Dilthey, understanding has always to do with the understanding of another psyche or mind. For Heidegger, however, the problem of other minds is replaced with *Dasein*'s preunderstanding of the world into which it is 'thrown.' Understanding the world, thus, and not an other mind, is the fundamental goal of Heideggerian hermeneutics. To understand is not to understand facts, but to understand what possibilities are open to us in the situation in which we find ourselves. As such, human existence constitutes what Heidegger famously terms 'a hermeneutic circle,' in that it implicitly interprets itself in terms of its everyday moods and projects before it raises this interpretation to the level of philosophical speculation.

For Heidegger, it is by way of language that we interpret a thing *as* something which is intimately related to our project of being-in-the-world. Language should not be thought of in this context as mere assertion, but as interpretative of hermeneutic discourse. For on Heidegger's reading, the world never shows up for *Dasein* in some raw manifestation, but is always disclosed in and through the projects which *Dasein* sets itself. For example, Heideggerian discourse does not simply assert that an object possesses a series of objective characteristics such as weight, color, and width. More essentially, it interprets the object *as* something which is useful and meaningful for its existence. Hence, the object is described, interpreted, or defined, in its relation to *my* specific projects. A rock is thus seen *as* a weapon or a sculpting stone, depending on the use *I* make of it. If objective assertion treats things as objects present-at-hand (*vorhanden*), hermeneutic discourse recognizes them as instruments ready-to-hand (*zuhanden*)—or as things which aid us in our attempts to fulfill our everyday projects. Hermeneutic discourse, that is, reveals things in terms of their possible serviceability for me, and it does so by releasing them from the abstraction of a timeless present into the temporal horizons of my concrete historical concerns—retrieving the meaningfulness of these entities from my past experience or projecting

such meaningfulness into my future possible experiences.

Heidegger's great contribution to the history of hermeneutics was his insistence that the world discloses itself through interpretative discourse. He failed, however, to explain what repercussions and ramifications this might have for the human sciences and for culture as a whole. It would be left to his student, Hans-Georg Gadamer, to fulfill this task.

3.2 Hans-Georg Gadamer

In Gadamer (1900–), we see a movement away from Heidegger's preoccupation with the question of being, back toward more epistemological concerns. His *magnum opus*, *Truth and Method* (1975), is dedicated to asking how the Heideggerian question of truth, as developed through a hermeneutic framework, can accommodate Dilthey's methodological hermeneutics of the human sciences. According to Gadamer, it was not essential for the human sciences to develop an exact method if they were to compete with the natural sciences. For while the natural sciences proceeded by way of logical induction, the human sciences had no such starting point or *method*. The human scientist, in other words, used insight and tact as tools rather than the laws of induction. Gadamer places the blame for the decline of the human sciences at the feet of those who, like Kant and his successors, considered the natural sciences as being eminently more objective than their subjective and aesthetic humanistic counterparts.

In his attempt to resuscitate the human sciences, Gadamer relies heavily on Heidegger's contention that we come to understand ourselves through our projects and possibilities. Preunderstanding, thus, is the condition of self-awareness. As such, and this is what it means to be caught in the hermeneutic circle, we come to an understanding of ourselves and our world by virtue of the prejudices of our specific time and place; that is, because *Dasein* is always already temporally situated, its preunderstanding will be conditioned and prejudiced by the world into which a person has been born. For instance, the reason why I interpret a rock *as* a weapon is because those who have preceded me in my tradition interpreted it *as such*. Interpretation, we will recall, is never neutral.

The question which Gadamer prioritizes, however, concerns the veracity of such interpretation; in an effort to avoid a crude form of relativism, Gadamer asks what we must do in order to avoid being duped by fallacious preunderstandings. As a devout hermeneut, he does not, of course, subscribe to the belief that misunderstanding can be fully eradicated. Interpretation, by its very nature, involves misreading and misconstrual. What we need to guard against, however, is the general propensity to take our prejudices at face value. While we cannot escape the prejudices of

our preunderstandings, we can nevertheless subject those prejudices and biases to critical scrutiny. Gadamer is aware that the interpreters have no vantage point outside their time and place to evaluate neutrally the efficacy of their preconceptions and prejudices. They rely, therefore, on what he terms 'temporal distance'; with the insight of hindsight, we can discern what interpretations have positively withstood the test of time. While he is cognizant of the fact that temporal distance is by no means fail-safe or fully reliable, Gadamer nevertheless acknowledges that it is perhaps the best available strategy for critically discerning between rival interpretations. The task for the individual is to try to develop a dialectical accommodation between the self that the individual has become by being on the receiving end of a tradition, and the individual's capacity to assume a critical disposition in relation to that tradition. Becoming aware that one is historically conditioned, and that one's beliefs derive not from a divine source, but from one's predecessors, is the first step towards such an accommodation.

It is through the process of responding to the claims which my tradition makes upon me that I come to understand the self that I am. Self-understanding, for Gadamer, involves dialoging with all those whose voices speak through me. I come to an understanding of *who* I am by reading and responding to those into whose heritage and legacy I have been situated. Hence, I am always already engaged in conversation with others who have gone before. The self is, in other words, dialogical through and through.

The problem with method, according to Gadamer, is that it never takes sufficient account of dialog, of the ongoing dialog within and beyond tradition. The form such dialog takes is that of a dialectic between question and answer. This ensures that no position can ever assume that it is beyond critical appraisal. The answer in such a dialectic will always beg yet another question. By standing back from one's tradition, not outside it, one can ask the relevant questions of it from one's particular vantage point, thus paving the way for our successors to ask vital questions of our own position. If there were no such dialog, tradition would simply freeze over. It is precisely by virtue of the fact that we endeavor to understand ourselves through a critical dialog of this type, that we can talk of historical progress.

Temporal distanciation is, therefore, that which allows me to observe a world which, although related in many respects to my own, is in many other senses different from the one I inhabit. The dialogue which pertains between the two worlds—my own and that of my forebear—is realized in what Gadamer calls a 'fusion of horizons.' This suggests that no tradition, horizon, or world is totally closed in upon itself, that it is indeed porous and open-ended. It demonstrates also that tradition cannot inoculate itself against the tensions which emerge in the course of a critical dialog

with its successor traditions. Such is the way we are 'affected' by history (*Wirkungsgeschichtlichkeit*).

4. *Critical Theory and Critical Hermeneutics: Habermas and Ricoeur*

4.1 *Jürgen Habermas and Critical Theory*

One way of appraising the Heidegger–Gadamer tradition of hermeneutic thought is to observe the way in which it interacts with what is known as the 'critical theory' school, the main protagonists of which are Georg Lukacs, Theodor Adorno, Walter Benjamin, Herbert Marcuse, and Jürgen Habermas. Primarily, critical theory appropriates the dialectical methods of Hegel and Marx, so as to develop a distinctive brand of socially committed analysis. The ultimate meaning of our world, they contend, cannot be divorced from its historical and political context. Benjamin and Habermas each developed the dialectical method of humanist Marxism in its dual capacity as critique of ideology and a project of emancipation. As such, they owed their greatest debt to Marx's original recognition that it was only by combining our practical and reflective activities that we might effectively surmount prevailing forces of domination and control.

It is within the context of Habermas's critique of ideology that the main differences between hermeneutics and critical theory emerge most clearly. For Habermas (1929–), what blocks the way to genuine social relations are the dominant illusions that inhibit fair and equal communication between social beings. The way to achieve such a goal is to keep in focus the utopia of, what he calls, an 'ideal speech situation.' This amounts to a democratic system of communication in which all social agents would have the chance to speak and make themselves heard in an undistorted fashion. It would be made possible when all participants in discourse possess an equal opportunity to initiate communication and become transparent by their words and actions. Hitherto, communication has been distorted by unequal distribution of dialog opportunities which sustain privileged positions.

In articulating the demands of the ideal speech situation, Habermas utilizes the psychoanalytic technique of identifying the forces which systematically distort communication. This model of *Sprachanalyse* ('the analysis of speech') amounts to a 'metahermeneutic' which endeavors to disclose the subject's quest for self-understanding through a process of desymbolization and resymbolization. Such a technique goes beyond, in Habermas's view, both the romantic and phenomenological deployment of hermeneutics which are founded upon a spontaneous recovery of intersubjective agreement. Where Gadamer sought to surmount misunderstanding by way of a question and answer retrieval of some original dialogical understanding, Habermas opts for the more arduous route

through a critique of ideology leading to the possibility of genuine language. The notion of unrestrained communication remains for Habermas, therefore, a utopian ideal.

Romantic and phenomenological hermeneutics are, according to Habermas, hindered by their refusal to think through sufficiently the sociopolitical dimension of understanding. He believes that Gadamer's ontological and poetic approach to the question of language is naive, to the extent that it is predicated upon the erroneous assumption that a consensus of dialogical communication could be achieved by recovering our traditional 'belongingness' to being. For Habermas, the construction of genuine consensus cannot dispense with a project of historical action. Hence, where the romantic hermeneutic spoke of a retrieval of lost meaning, Habermas speaks of an anticipation of possible meaning. In contrast to Gadamer's view that misunderstanding presupposes a prior or preunderstanding, Habermas insists that where there is misunderstanding there may exist a posterior understanding if and only if we first undertake a critique of those ideological distortions which reside at the origin of misunderstanding. What is given is always already deformed communication. Habermas, that is, wishes to show how the ideal of communicative competence is not an ontological precondition of understanding but a historical possibility yet to be achieved. Only in this manner can we reconcile the philosophers who interpret the world and the social agents who change it.

4.2 Paul Ricoeur and Critical Hermeneutics

If Habermas criticizes romantic and phenomenological hermeneutics for being too focused on the past, and not sufficiently attentive to its ideological deformations, French philosopher Paul Ricoeur (1913–) tries to find a mid-passage between the metahermeneutics advanced by Habermas, and the hermeneutics of tradition evoked by Gadamer. In so doing, he develops a 'critical hermeneutics' which dissents from the romantic idea that we understand the past by reproducing in the present some original production of meaning, as if the temporal distancing of meaning could be in some way erased. A critical hermeneutics of tradition insists on the necessity to discriminate between true and false interpretations of history. This, of course, raises the crucial question of legitimation. To resolve this problem, Habermas appealed to an ahistorical ideal of undistorted communication. For Ricoeur, however, such a criterion of legitimacy may be deferred to an indefinite future without any grounds in history. He recommends, accordingly, that we interpret tradition's claim to truth in a nonabsolutist sense of a *presumption* of truth. This suggests that we accept the truth claims of tradition until such a time as a better argument prevails. The presumption of truth

refers to our basic attitude of credit or trust in the propositions of meaning legacies by the past—a primary response which precedes the critical moment of distancing and reminds us that we are not the originators of truth but that we always already belong to a context of what he terms, 'presumed truth.' Ricoeur believes that such a model bridges the divide between the finitude of hermeneutic imagination (Heidegger and Gadamer) and the validity of the ideal undistorted communication (Habermas).

Ricoeur inserts Habermas's critique of ideology into the heart of the hermeneutics of tradition. This is born of his conviction that such a project demands a fundamental respect for tradition if it is to safeguard itself against arbitrary or ahistorical voluntarism—that is, a future project completely divorced from the historical heritage of the past. Where Gadamer invokes the notion of a common understanding which precedes us, the critical consciousness of Habermas reinterprets it in terms of the ideal of unrestricted and unconstrained communication. This antithesis is, however, broken down if one espouses a critical hermeneutics that realizes that critical theory cannot take up a view from nowhere, but must stick to its professed principles by presupposing some form of historical memory. For critical theory, a historical memory would not be that of romanticism (as it was for Gadamer) but rather of the Enlightenment, understood as a project of emancipation.

Tradition, Ricoeur concludes, must be understood in the dynamic perspective of our being-affected-by-the-past, which in turn is related to our historical horizon of expectancy. In this larger dialectic between tradition and expectation, we rediscover suppressed potentialities of past meaning which give substance to the ideal of undistorted communication. Indeed, it is only in terms of such an interplay between memory and anticipation that the ideal image of a reconciled humanity can be invested with an effective history. A critical hermeneutics has thus the ethical duty of ensuring that the expectancies for the future are brought nearer to the present by being sensitive to the concrete actions which need to be undertaken so as to bring to a realization what is 'desirable and reasonable.' Second, it must continually strive to liberate the still untapped potentialities of inherited meaning.

In this way, Ricoeur's particular brand of hermeneutics provides an essential link between romantic hermeneutics and critical theory. His incisive analysis of myth throughout his multifarious writings shows precisely how such a critical hermeneutics can be deployed in the service of concrete sociopolitical objectives. The modern project of unmasking myth frequently takes its cue from the investigative methods developed by Marx, Nietzsche, and Freud—the 'three masters of suspicion' as Ricoeur likes to call them. Once a hermeneutics of suspicion has unmasked the alienating role of myth as an agency of ideological conformism, there remains, for Ricoeur, the task of a

positive interpretation. Hermeneutics, that is, has the double task of *suspecting* and *listening*. Having demythologized the ideologies of false consciousness it labors to disclose the utopian symbols of liberating consciousness. It thereby seeks to rescue mythic symbols from reactionary domination, and to show that once the mystifying function has been dispelled we may recover genuinely utopian anticipations of 'possible worlds' of liberty and justice. A positive hermeneutics of this sort offers an opportunity to rescue myths from the ideological abuses of doctrinal prejudice, nationalism, class oppression, or totalitarian conformism; and it does so in the name of a universal project of freedom—a project from which no creed, nation, class, or individual is excluded. The utopian content of myth differs from the ideological in that it is inclusive. It opens up human consciousness to a common goal of liberation instead of closing it off in inherited securities. That is, having eliminated the ideological abuse of myth as a false explanation of how things are, we are now free to appreciate the properly symbolic role of myth as an exploration of how things might be.

See also: Critical Theory: Contemporary; Critical Theory: Frankfurt School; Hermeneutics, History of; Husserl, Edmund (1859–1938); Truth, Verification, Verisimilitude, and Evidence: Philosophical Aspects; Verstehen und Erklären, Philosophy of

Bibliography

- Bruns G 1992 *Hermeneutics Ancient and Modern*. Yale University Press, New Haven, CT
- Dannhauer J C 1654 *Hermeneutica sacra sive methodus exponendarum sacrarum literarum*. Augsburg, Germany
- Dilthey W 1976 *Selected Writings*. Cambridge University Press, Cambridge, UK
- Gadamer H-G 1975 *Truth and Method*. Seabury Press, New York
- Grondin J 1994 *Introduction to Philosophical Hermeneutics* [trans. Weinsheimer J]. Yale University Press, New Haven, CT
- Habermas J 1971 *Knowledge and Human Interests*. Beacon Press, Boston
- Habermas J 1974 *Theory and Practice*. Heinemann, London
- Habermas J 1979 *Communication and the Evolution of Society*. Beacon Press, Boston
- Habermas J 1984 *The Theory of Communicative Action*. Beacon Press, Boston, Vol. 1
- Habermas J 1987 *The Theory of Communicative Action*. Beacon Press, Boston, Vol. 2
- Heidegger M 1962 *Being and Time*. Blackwell, Oxford, UK
- Ricoeur P 1974 *The Conflict of Interpretations*. Northwestern University Press, Evanston, IL
- Ricoeur P 1981 *Hermeneutics and the Human Sciences*. Cambridge University Press, New York
- Ricoeur P 1984–7 *Time and Narrative*. University of Chicago Press, Chicago
- Ricoeur P 1986 *Lectures on Ideology and Utopia*. Columbia

University Press, New York

Ricoeur P 1992 *Oneself as Another*. University of Chicago Press, Chicago

M. Dooley and R. Kearney

Heterosexism and Homophobia

Homophobia is a colloquial expression that refers to negative, fearful, or hateful attitudes and behavior toward gays and lesbians. Although homophobia is a common term used to indicate antilesbian and antigay prejudice, many social scientists object to the concept. Herek (1995), for instance, objects that prejudice against homosexuality is not a phobia in the clinical sense of the word; it is not based necessarily on fear, nor is it inevitably irrational or dysfunctional for the individuals who manifest it. He defines heterosexism as 'an ideological system that denies, denigrates, and stigmatizes any nonheterosexual form of behavior, identity, relationship or community.' Adam (1998), Plummer (1992), and Stein and Plummer (1994) find it useful to distinguish analytically and conceptually between heterosexism, which is defined as societal prejudice located in the structures, institutions, and material conditions of society, and homophobia, which is a manifestation of the individual-level attitudes and behavior reflective of heterosexism. Herek (1995) prefers the term 'psychological heterosexism' to denote individual-level antigay and antilesbian prejudice. Heterosexism and homophobia are mutually constitutive in that individuals feel that their antigay attitudes and behaviors are legitimate to the extent that these prejudices are entrenched in social institutions. Individuals are more likely to act out their prejudices when they consider them to be culturally legitimate and shared by others (Berrill 1992). Expressions of antigay prejudice reinforce ideologies and practices of intolerance and hostility and reinscribe homosexuality as something that should be despised and feared and should remain hidden.

1. Attitudes and Expressions of Homophobia

One common stereotype is that homosexuality is related to gender-role nonconformity. It is assumed frequently that lesbians manifest characteristics defined as masculine and gay men exhibit behavior associated with feminine characteristics. For many people, this presumed gender nonconformity is reason enough to be suspicious and intolerant of homosexuals. Gay men are also often assumed to be hypersexual, animalistic, promiscuous, overly visible, and conspiratorial. Lesbians are assumed to be man-haters, aggressive, socially nonconforming, and sex-

ually cold (Herek 1995). Until 1973 the American Psychiatric Association considered homosexuality to be a form of mental illness. Many people continue to hold this belief and also believe homosexuality to be immoral and antisocial. Expressions of homophobia include verbal assaults and derisive joking, negative stereotypical media representations, discrimination in employment, education, housing, medical research, legal defense, and custody rights, physical attack, and tacit approval of all forms of discrimination, including gay bashing. Various studies, including surveys of gays and lesbians based on probability samples, indicate that nearly half of all respondents have experienced some form of discrimination in employment or housing based on their sexual orientation. A meta-analysis of 24 questionnaire studies showed that a median of 44 percent of the respondents have been threatened with violence and 80 percent have been verbally harassed (Berrill 1992). All of the respondents attribute these discriminatory actions to the perpetrators' perceptions that the target was homosexual. In 1987 the US Department of Justice reported that homosexuals are the most frequent victims of hate violence.

1.1 *Containing Sexuality*

Several theorists have observed that homophobia is an expression of a culture that is confused and inconsistent about sexual behavior. Attitudes and actions reflect both obsession with and repression of sexual desire. One basis of tacit agreement is that sexuality is something that must be contained. Homosexuality, along with adultery, prostitution and pederasty, mark the boundaries by existing outside culturally approved forms of sexuality—what poet Rich (1980) refers to as compulsory heterosexuality. Social historians (e.g., Boswell 1980, D'Emilio and Freedman 1988, Katz 1976) note that the emergence of a sexual binary in the nineteenth century posits heterosexuality as natural and socially desirable by contrasting it with homosexuality, which is characterized as unnatural and immoral. This binary represents the general tendency in modern western thought to conflate morality and nature. According to this cultural equation, sexual desire is a manifestation of a natural drive to reproduce. The moral, disciplined individual contains this desire within the bounds of heterosexual monogamy. These attitudes are reflected in clinical, social and religious definitions of the homosexual as cursed, perverted, sexually obsessed, and evil and, until recently, mentally and physically sick.

2. *Institutional Heterosexism*

Pervasive negative stereotypes regarding lesbians and gay men suggest that these attitudes and beliefs are

rooted in cultural ideologies and social practices. All of these attitudes and expressions reflect the prevailing belief that sexual desire for the opposite sex is expected and normal; sexual desire for someone of the same sex is abnormal and problematic. The lingering belief that homosexuality is dysfunctional and that it constitutes a state of anguish for the individual can be traced to the medicalization of homosexuality and its treatment as an illness. Many medical practitioners and clinicians continue to offer treatments intended to 'cure' lesbians and gays of their 'affliction.' These practices reflect the cultural belief that individuals will be unhappy and maladjusted if they do not experience sexual desire for persons of the opposite sex. Medical assessments that treat homosexuality as a form of disease reinscribe the notion that sexuality is primarily physiological and that heterosexuality is the healthy manifestation of human physiology.

Most contemporary religions also treat homosexuality as problematic. In some cases a distinction is made between homosexual feelings and homosexual acts (e.g., Catholic Catechism 1989). The experience of homosexual inclinations may be considered an affliction—a 'cross to bear'—but to act on one's homosexual feelings is considered a sin. Many religious leaders advocate celibacy for those 'afflicted' with homosexual tendencies. Some religious organizations sponsor intensive therapy programs intended to help the individual who is 'suffering' from homosexuality to recover. Similar to medical institutions, religious institutions take heterosexuality for granted as the normal and desired form of sexual organization and treat homosexuality as deviant, undesirable, dysfunctional, and, additionally, sinful.

Legal practices also reflect a cultural ideology that supports heterosexuality as normal and denigrates homosexuality as abnormal. The most prevalent indication of compulsory heterosexism is the widespread lack of legal protection from antigay/lesbian discrimination. Lesbians and gays who experience discrimination in employment, housing, education and immigration frequently find it difficult to obtain justice. In the US it is still legal to ban lesbians and gays from various public institutions including service in the military and employment in various educational domains. Another indication of cultural heterosexism is public resistance to attempts by lesbians and gay men to reduce discrimination and to seek justice. In the 1990s US citizens in several states voted against ballot initiatives that would ensure equal legal protection for lesbians and gays. The rhetoric employed to defeat these initiatives was that lesbians and gays were seeking 'special rights and privileges.' The logic underlying this argument was that the US constitution already guarantees every individual equal protection under the law. While this is the case in theory, advocates of the equality initiatives point out that in practice this protection is often not extended to lesbians and gays. In fact, many social officials,

including employers, housing agencies and health care providers actively refuse to grant lesbians and gays equal rights and openly cite deviant sexual orientation as the basis for this discrimination. US courts have routinely upheld these decisions

Not only does the law fail to uphold the rights of lesbians and gays, but it has been enforced differentially as a means of punishing lesbians and gays. One example of differential enforcement is the enactment of sodomy laws. A majority of states still have laws that proscribe oral or anal sex. In most cases, these laws refer only to the behavior and not to the gender of the participants. However, the only recorded arrests in the past several decades based on these archaic laws have been gays and lesbians. When challenged, several states (Kentucky, Louisiana, Michigan, and Texas) have ruled sodomy laws as unconstitutional. However, in 1986, in the case of *Bowers v. Hardwick*, the US Supreme Court upheld a Georgia sodomy law in a narrow vote of 5 to 4. The majority opinion, written by Chief Justice Warren Burger expresses the general legal failure to separate the expression of private individual sexual behaviors from perceived cultural expectations. In his brief he writes: 'To uphold that homosexual sodomy is somehow protected as a fundamental right would be to cast aside millennia of moral teaching' (quoted in Singer and Deschamps 1994).

Court decisions such as *Bowers vs. Hardwick* galvanized many US citizens who were concerned with the implications of court-mandated moral behavior and invasion of privacy. As a consequence, in the 1990s lesbians and gays have fought actively against discrimination. Several gains have been made, but these gains have been matched in intensity by antigay and antilebian legislation. The intent of initiatives such as Colorado's Amendment 2 (enacted in 1992) is to legally prohibit lesbians and gays from seeking legal protection for discrimination based on their status as homosexuals. This is the first proactive legislation in the history of US law that actively denies the pursuit of civil rights based on a single specific category—homosexuality. In an address to Congress regarding legislation intended to protect persons from discriminatory violence (hate crimes), Senator Dannemeyer makes clear this categorical prejudice. He notes: '(I)t is a Federal responsibility to ensure the equal protection of all citizens regardless of their race, religion, or ethnic origin. It is not a Federal obligation to protect citizens in their sexual orientation' (US Congress 1988, p. 12).

2.1 Permissible Prejudice

The extent to which homophobia is an institution is indicated in the decisions of law enforcement officials and judges who consider homosexuality sufficient grounds for justifying acts of prejudice, discrimination,

and violence. This tacit approval constitutes a form of permissible prejudice. Persons presume their prejudices to be legitimate or permissible when these prejudices are condoned by persons in positions of authority. Not only do the words of Justice Burger and Senator Dannemeyer reveal the extent to which heterosexuality is considered normal and legitimate, they suggest that homosexuality is a form of deviance that is wrong and shouldn't be tolerated. These institutional practices buttress individual feelings of fear and hate and send the message that discrimination and violence will be tolerated. In contrast, legal and cultural practices in many countries and some states in the US reflect the intent to extend equal rights to all citizens. Gay and lesbian marriages are legally recognized in Denmark and the Netherlands, for instance, and of the 16 countries that comprise the North Atlantic Treaty Organization, only the UK and the US ban lesbians and gays from service in the military.

2.2 Internalized Homophobia

In a climate of permissible prejudice, many lesbians and gays live in a continual state of fear. The entrenchment of heterosexist expectations and the corresponding active prejudice against homosexuality also induces feelings of 'internalized homophobia.' Internalized homophobia is a form of self-disgust and hatred experienced by many lesbians and gays. The phrase 'internalized homophobia' suggests that individuals who are trying to come to terms with their homosexual desires absorb or take in attitudes of cultural prejudice and use these attitudes as a basis of self-judgement. If the prevalent cultural story-line is one of intolerance and debasement, then individual lesbians and gays are likely to have to deal with this in formulating their own sense of themselves. In a culture that presumes heterosexuality, gays and lesbians must decide continually whether or not to voice information indicative of a deviant sexuality. 'Coming out' is an ongoing process. This process involves wrestling with and reconciling family, peer and religious expectations with personal feelings and experiences. Many lesbians and gay men fear negative consequences for revealing their sexuality. The majority experience rejection by their families, friends and religious and working communities when they reveal their homosexuality. Fear of rejection and a lack of positive role models can lead to confusion and self-hatred. Information compiled by the US Department of Health and Human Services in 1989 indicates that lesbian and gay youth are two to three times more likely to attempt suicide than their heterosexual peers. Gay and lesbian youth account for 30 percent of all completed youth suicides. In contrast, religious and community organizations that endeavor to understand and support gay and lesbian youth have been successful in reconciling family conflict and in integrating these youth into their schools and communities.

3. Explanations for Homophobia/Heterosexism

Adam (1998) delineates three theoretical approaches to understanding homophobia. Each approach is distinguished by a conceptual term denoting a specific theoretical orientation to anti-homosexual prejudice: Homophobia, Heteronormativity, and Heterosexism. Adam links these concepts to specific disciplines: Psychology, Cultural Studies (literary), and Sociology, respectively. The distinguishing aspect of each approach is the analytical unit: Personality, Discourse, and Social Structure. Explanations can be grouped accordingly. Within the field of psychology, Herek (1995) reviews the 'functional' explanations ('people hold and express particular attitudes because they derive psychological benefit from doing so'). Herek proposes three principle functions underlying attitudes about homosexuality: experiential, social identity, and ego-defense. A person's attitude may be a consequence of personal experiences (positive or negative); a prejudiced attitude may be intended as an expression of allegiance to a particular social group; or it may be a defensive response to personal psychological conflicts (such as gender expectations). He emphasizes that it is important to recognize the nexus between these functional attitudes and cultural heterosexism. The source and expression of these attitudes directly reflects a person's perceptions of cultural expectations.

The development of literary cultural studies in the 1990s has generated several interrelated theories which come together under the term, 'queer theory.' Queer theorists (e.g., Butler 1999, Sedgwick 1990, Seidman 1997) emphasize discursive practices that promote certain lines of thought and action and discourage or punish others. Heteronormativity is one of the most prevalent discourses in contemporary western culture. These theories focus on the manifestations of heterosexuality in art, media, education, legal discourse, religion, and so forth. According to these theories, the dominant cultural discourse is saturated with images and ideals that promote heterosexuality and condemn homosexuality (or anything that does not fit the expected monogamous, heterosexual, family idea). Queer theorists have particularly focused on the experience of the 'closet'—physical and discursive practices intended to keep homosexuality hidden and contained in separate spheres.

Adam subsumes the psychological and cultural studies approaches within the sociological. This approach emphasizes social structure, particularly the ways in which psychological attitudes and the values revealed in cultural discourses are manifest and maintained through institutional structures and practices. Change, he suggests, will be a result not only of education against prejudiced attitudes and transgressive acts intended as critiques of taken-for-granted cultural practices, but the extent to which this education and alternative expression permeates and unsettles existing institutional arrangements (such as the

family, law, and religion). Seidman et al. (1999) suggest that recent gay and lesbian social movements have not only served to challenge some of these existing arrangements, but that they may have the paradoxical effect of 'decentering' lesbian and gay identities. According to their logic, the conditions of the closet may compel individuals to make their homosexuality a primary focus and thus, to solidify it as a primary identity. Paradoxically, normalization of homosexuality may enable individuals to de-emphasize this aspect of their behavior.

3.1 Gender

Understanding homophobia is complicated by the strong link between gender and sexuality in this culture. Gender is the basic criterion by which we select sexual partners. The violation of expected gender behavior elicits taunting, prejudice, and physical and social abuse. Frequent targets of this prejudice include the mannish lesbian and the effeminate gay male. Both evoke stereotypes that represent violations of gender expectations and hint at sexual transgressions. Transgendered persons are frequent targets of violence (see *Transsexuality, Transvestism, and Transgender*). It is difficult to know if attackers are provoked by what they assume to be a sexual aberration or a gender violation. Given the tight connection between gender and sexuality, it is likely that any deviation is perceived as a threat to heterosexuality. In the 1990s Lambda Legal Defense documented several cases in the US of murdered transgendered persons in which the murderer was released by the courts. The rationale offered in each case was that to arouse sexual desire in another under 'false gender pretense' was to invite violent 'self-defense' when the pretense was revealed (Singer and Deschamps 1994). This rationale can be interpreted as a manifestation of homophobia; individuals who do not conform to the expected norms of gender and sexuality do not deserve social justice.

4. Social Movements

The social protests and political activism associated with lesbian and gay history can be interpreted as a response to widespread, organized manifestations of homophobia. Several observers have suggested that the struggle for gay and lesbian rights was the civil rights issue of the 1990s. This history represents a fight for rights that are constitutionally guaranteed to all Americans but frequently withheld from homosexuals. Many local and national lesbian and gay movements have been organized as a direct response to political initiatives designed to legalize discrimination against homosexuals. The Defense of Marriage Act introduced into US federal legislation in 1996 is one such case. Marriage between same-sex couples poses no legal threat or loss of benefits to heterosexual couples.

It is difficult to interpret the urgency with which states have pursued legislation to restrict same-sex marriages as anything other than a display of mass homophobia. In the early 1990s antigay/lesbian legislation appeared on the ballots of 41 states and the District of Columbia. This sort of legislation is indicative of the extent to which homophobia is culturally legitimated as an institutionalized form of prejudice. Critics who suggest that antidiscrimination laws are 'special rights' fail to comprehend the extensive reach of homophobic actions into the everyday lives of lesbians, gays, bisexuals, and transgendered persons.

See also: Feminist Movements; Feminist Theory; Radical Lesbian; Foucault, Michel (1926–84); Gay, Lesbian, and Bisexual Youth; Gay/Lesbian Movements; Gender, Feminism, and Sexuality in Archaeological Studies; Homosexuality and Psychiatry; Lesbians: Historical Perspectives; Lesbians: Social and Economic Situation; Libraries: Digital, Electronic, and Hybrid; Masculinities and Femininities; Queer Theory; Rationality and Feminist Thought; Regulation: Sexual Behavior; Reproductive Rights in Affluent Nations; Reproductive Rights in Developing Nations; Sexual Attitudes and Behavior; Sexual Behavior: Sociological Perspective; Sexual Harassment: Legal Perspectives; Sexual Harassment: Social and Psychological Issues; Sexual Orientation and the Law; Sexual Orientation: Biological Influences; Sexual Preference: Genetic Aspects; Sexuality and Gender; Sexuality and Geography; Social Movements and Gender; Transsexuality, Transvestism, and Transgender

Bibliography

- Adam B 1998 Theorizing homophobia. *Sexualities* 1: 387–404
- Berrill K T 1992 Anti-gay violence and victimization in the United States: An overview. In: Herek G, Berrill K (eds.) *Hate Crimes: Confronting Violence against Lesbian and Gay Men*. Sage, Newbury Park, CA pp. 19–45
- Boswell J 1980 *Christianity, Social Tolerance, and Homosexuality*. University of Chicago Press, Chicago
- Butler J 1999 *Gender Trouble*. Routledge, New York
- Catechism of the Catholic Church 1989 *Article Six: Offenses Against Chastity*. Paulist Press Mahwah, NJ, pp. 566–7
- D'Emilio J, Freedman E 1988 *Intimate Matters: A History of Sexuality in the United States*. Harper and Row, New York
- Herek G 1995 Psychological heterosexism in the United States. In: D'Augelli A, Patterson C (eds.) *Lesbian, Gay and Bisexual Identities over the Lifespan*. Oxford University Press, Oxford, UK, pp. 321–45
- Katz J 1976 *Gay American History*. Crowell, New York
- National Institute of Justice, US Department of Justice 1987 *The Response of the Criminal Justice System to Bias Crime: An Exploratory Review*
- Plummer K 1992 Speaking its name: Inventing a lesbian and gay studies. In: Plummer K (ed.) *Modern Homosexualities*. Routledge, London, pp. 3–28
- Rich A 1980 Compulsory heterosexuality and lesbian existence. *Signs: Journal of Women in Culture and Society* 5: 631–60

- Sedgwick E K 1990 *Epistemology of the Closet*. University of California Press, Berkeley, CA
- Seidman S 1997 *Difference Troubles: Queering Social Theory and Sexual Politics*. Cambridge University Press, Cambridge, UK
- Seidman S, Meeks C, Traschen F 1999 Beyond the closet? The changing social meaning of homosexuality in the United States. *Sexualities* 2: 9–34
- Stein A, Plummer K 1994 I can't even think straight: Queer theory and the missing sexual revolution in sociology. *Sociological Theory* 12: 178–87
- Singer B L, Deschamps D 1994 *Gay and Lesbian Stats*. New Press, New York
- US Department of Health and Human Services 1989 *Report of the Secretary's Task Force on Youth Suicide*
- US Congress 1988 Hate Crimes Statistics Act. *Report from the Committee on the Judiciary, 100th Congress, 2nd Session*. Government Printing Office, Washington, DC

J. O'Brien

Heuristics for Decision and Choice

Heuristics are approximate strategies or 'rules of thumb' for decision making and problem solving that do not guarantee a correct solution but that typically yield a reasonable solution or bring one closer to hand. As such, they stand in contrast to algorithms that produce a correct solution, given complete and correct inputs, if one exists. More specifically, heuristics are usually thought of as shortcuts that allow decisions or solutions to be reached more rapidly, even in conditions of incomplete or uncertain information—often because they do not process all the available information. Heuristics are most commonly studied in psychology within the domains of judgment and decision making, social cognition, and problem solving, and in computer-based applications in artificial intelligence (AI) and operations research (OR). This entry focuses on heuristics proposed as models of how people make decisions and choices.

1. Historical Overview

The term 'heuristic' (from a Greek root, 'to discover') was used for most of the twentieth century to refer to useful, even indispensable, strategies for finding solutions to problems that are difficult to approach by other means. Gestalt psychologists spoke of heuristic reasoning methods such as 'looking around' and 'inspecting the problem' to guide the search for useful information in the environment, while mathematicians employed heuristics including 'examining special cases', 'exploiting related problems', 'decomposing and recombining', and 'working backwards' (Groner 1983). These rather vague strategies were made more precise in computer-based models of human problem

solving and reasoning largely based on the means-ends analysis heuristic, which sought some way to reduce the distance between the current partial-solution state and the goal state (e.g., in Newell and Simon's early General Problem Solver system; see *Artificial Intelligence: Search*). Such general purpose or 'weak' methods proved insufficient to tackle many problems, so research in AI in the 1970s turned to collecting domain-specific rules of thumb from specialists in a particular field and incorporating these into expert systems. Around the same time, mathematicians working in OR faced new results from computational complexity theory indicating that efficient algorithmic solutions to many classes of challenging combinatorial problems (such as the Traveling Salesman Problem) might not be found; as a consequence, they too turned to the search for problem-specific heuristics, though through invention rather than behavioral observation (Müller-Merbach 1981).

After 1970, though, heuristics gained a different connotation in psychology: fallible cognitive shortcuts that people often use in situations where logic or probability theory should be applied instead. The 'heuristics-and-biases' research program launched by Tversky and Kahneman (1974, Kahneman et al. 1982) emphasized how the use of heuristics can lead to systematic errors and lapses of reasoning (see *Decision Biases, Cognitive Psychology of*) indicating human irrationality. The heuristics studied (see Sect. 2) were often vaguely defined and broadly applicable to judgments made under uncertainty in any domain, akin to the weak methods explored earlier in AI. This negative view of heuristics and of the people who use them as 'cognitive misers' employing little information or cognition to reach biased conclusions has spread to many other social sciences, including economics (Rabin 1998) and law (Hanson and Kysar 1999). More recently, a new appreciation is emerging within psychology that heuristics may be the only available approach to decision making in the many problems where optimal logical solutions are computationally intractable or do not exist (as OR researchers realized), and that domain-specific decision heuristics may be more powerful than domain-general logical approaches in other problems (as AI found). This has led to the study of precisely-specified heuristics matched to particular decision tasks (see Sect. 3), and the ways that learning and evolution can achieve this match in human behavior (e.g., Payne et al. 1993, Gigerenzer et al. 1999). The existence of such evolved adaptive heuristics has already been widely accepted for other animals in research on rules of thumb in behavioral ecology (Gigerenzer et al. 1999).

2. Heuristics for Probability Judgment

Since the rise of the heuristics-and-biases research program in the 1970s, the heuristics most widely studied within the social sciences are those that people

use to make judgments or estimates of probabilities or other quantities in situations of uncertainty where they do not have complete, exact information (Tversky and Kahneman 1974). The representativeness heuristic is a means to assess the probability that an object A belongs to a particular category B (e.g., that a person described as meek is a pilot) or that an event A is generated by a particular process B (e.g., that the sequence HHTTTT was generated by throwing a fair coin). This heuristic produces probability judgments according to the extent that A is representative of, or similar to, B (e.g., meekness is not representative of pilots, so a meek person is judged as having a low probability of being a pilot). The availability heuristic can be used to produce assessments of class frequency or event probability based on how easily instances of the class or event can be mentally retrieved (e.g., plane crashes may seem like a frequent cause of death because it is easy to recall examples) or constructed (via the simulation heuristic). The anchoring-and-adjustment heuristic produces estimates of quantities by starting with a particular value (the anchor) and moving away from it (e.g., people asked to quickly estimate either $8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$ or $1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$ give a higher value in the former case, which could come from estimation based on multiplying the first few values together, creating a higher or lower anchor, and then adjusting upwards, which yields a higher overall estimate of the first product).

Most researchers agree that such heuristics are widely used because they usually lead to good choices without much time or mental effort. Despite this, most of the large body of evidence amassed that is consistent with the use of these heuristics comes from showing where they break down and lead to errors (e.g., availability judgments can be manipulated by vivid examples like plane crashes, and anchoring can lead to different choices solely due to the order of information presentation—see *Decision Biases, Cognitive Psychology of*), leading many to conclude that people are poor decision makers after all. These heuristics have also been criticized as being vague redescrptions of observed choice behavior that are therefore difficult to falsify; however, more specific versions (e.g., testing whether availability works in terms of ease of recall or number of items recalled) are also being explored.

3. Heuristics for Choice between Alternatives

Specifying decision tasks in more detail—for instance, choosing between two (or more) alternatives on the basis of cues or pieces of information known about each—allows still more specific heuristics to be investigated. The typical normative approach to such choice decisions involves weighting and combining (typically adding) all of the available cues, with the weights determined by a process such as multiple regression. Heuristics can shortcut this process in two main ways.

First, rather than weighting all cues differentially, positive and negative cues for each alternative can merely be tallied (Dawes 1979). Second, not all cues need be used—instead, a stopping rule can indicate when enough cues have been considered to allow a choice (see below). Despite these simplifications, such heuristics often perform very well when compared to expert human decision makers (Dawes 1979), traditional benchmarks (Payne et al. 1993), or objective measures of correct answers (Gigerenzer et al. 1999).

There are several forms of stopping rules that can be used to limit the information considered by choice heuristics. The recognition heuristic (Gigerenzer et al. 1999) only ever considers the single cue of recognition: when one alternative is recognized and the other is not, this heuristic chooses the former (and otherwise guesses). One-reason decision mechanisms (Gigerenzer and Goldstein 1996) compare both alternatives on a single cue-dimension (or reason) at a time and stop at the first dimension that distinguishes between the alternatives (i.e., where their cue values differ, allowing a decision to be made). Accumulator or horse-race models (Busemeyer and Rapoport 1988) tally the cues in favor of each alternative until a specified difference in positive evidence is reached (e.g., until one alternative is three reasons ahead of the other). Elimination heuristics (Tversky 1972, Gigerenzer et al. 1999) begin with a set of alternatives and successively remove subsets that do not match a specified cue value until a single choice remains (e.g., to choose a restaurant, eliminate all those in town over some price threshold, then eliminate all those beyond some distance away, and so on until one restaurant is left). In every case (aside from the recognition heuristic), a search rule must also be specified to determine the order in which cue dimensions are considered; this order can (among other possibilities) be random, or based on recently used cues, or on cue validity or correlation with the choice criterion (which, when combined with a one-reason stopping rule, yields a lexicographic strategy such as the Take The Best heuristic—Gigerenzer and Goldstein 1996).

4. Heuristics for Sequential Search

When choice alternatives are encountered sequentially over time rather than simultaneously (e.g., in comparison shopping by driving from one store to the next), heuristics can be used to decide when to stop searching and select an available alternative. The traditional normative approach is to search until one finds an alternative below a precalculated reservation price balancing expected benefit of further search against its cost; this requires full knowledge of the search costs and distribution of available alternatives. Heuristics that use a simplified version of the reservation price calculation (e.g., replacing an integral with a weighted sum) come very close to normative

performance (e.g., at selecting good prices, Moon and Martin 1990). Other heuristics require less knowledge, such as ‘Keep searching until the total search cost exceeds 7.5 percent of the best price found’ (Moon and Martin 1990). Herbert Simon’s bounded rationality principle of satisficing (see *Bounded Rationality*) suggests setting an aspiration level equal to an alternative that is good enough (rather than optimal) and searching until that aspiration level is met; exactly how the aspiration level can be set varies with the search setting (e.g., whether it is a one-sided search like shopping, or a two-sided mutual search like finding a mate, Gigerenzer et al. 1999). Finally, another type of search heuristic that people use stops search after a particular pattern of alternatives is encountered, rather than after some threshold is exceeded (despite the fact that pattern should not matter from a normative perspective). For instance, the ‘one-bounce’ and ‘two-bounce’ rules say to keep searching for a low price until prices go up for the last or two last alternatives, respectively (Moon and Martin 1990).

5. Further Directions

Decision heuristics have been studied in different research traditions, in particular one that has focused on when and where verbally described heuristics can break down and yield deviations from classical norms of rationality (Kahneman et al. 1982), and another that has investigated how specific computationally modeled heuristics can exploit structured information to yield fast and accurate decisions (Gigerenzer et al. 1999). Further research should explore not only those aspects of decision environments that lead to different levels of performance, but also what factors lead people to use heuristics at all (e.g., time pressure, Payne et al. 1993), to choose between different possible heuristics (e.g., the influence of domain-specific emotions or social norms), and to learn or develop new heuristics (e.g., through social imitation).

See also: Bounded and Costly Rationality; Decision Biases, Cognitive Psychology of; Decision Making (Naturalistic), Psychology of; Decision Making: Non-rational Theories; Decision Research: Behavioral; Decision Support Systems; Decision Theory: Classical; Heuristics in Social Cognition; Informal Reasoning, Psychology of; Problem Solving (Everyday), Psychology of; Sequential Decision Making

Bibliography

- Busemeyer J R, Rapoport A 1988 Psychological models of deferred decision making. *Journal of Mathematical Psychology* **32**: 91–134
- Dawes R M 1979 The robust beauty of improper linear models in decision making. *American Psychologist* **34**: 571–82

- Gigerenzer G, Goldstein D G 1996 Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review* **103**: 650–69
- Gigerenzer G, Todd P M, ABC Research Group 1999 *Simple Heuristics That Make Us Smart*. Oxford University Press, New York
- Groner R (ed.) 1983 *Methods of Heuristics*. Erlbaum, Hillsdale, NJ
- Hanson J D, Kysar D A 1999 Taking behavioralism seriously: The problem of market manipulation. *New York University Law Review* **74**: 630–749
- Kahneman D, Slovic P, Tversky A 1982 *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York
- Müller-Merbach H 1981 Heuristics and their design: A survey. *European Journal of Operational Research* **8**: 1–23
- Moon P, Martin A 1990 Better heuristics for economic search. Experimental and simulation evidence. *Journal of Behavioral Decision Making* **3**: 175–93
- Payne J W, Bettman J R, Johnson E J 1993 *The Adaptive Decision Maker*. Cambridge University Press, New York
- Rabin M 1998 Psychology and economics. *Journal of Economic Literature* **36**: 11–46
- Tversky A 1972 Elimination by aspects: A theory of choice. *Psychological Review* **79**: 281–99
- Tversky A, Kahneman D 1974 Judgment under uncertainty: Heuristics and biases. *Science* **185**: 1124–31

P. M. Todd

Heuristics in Social Cognition

‘Heuristics’ are strategies of simplifying judgments that allow individuals to make decisions under sub-optimal circumstances. The discovery of heuristics has had a profound impact on social psychology (see *Social Psychology*), especially in the field of ‘social cognition,’ which studies the attempts by individuals to make sense of others (see *Mental Representation of Persons, Psychology of*). This article describes the general idea of heuristic information processing, looks at three specific heuristics, and examines the importance of heuristics for social psychology.

1. Judgments under Suboptimal Conditions

Whereas traditional theories of human decision-making (see *Decision Making, Psychology of*) have focused on the normative aspects of valid judgments while neglecting the context in which they occur, the ‘heuristics’ approach has directed its attention to the psychological processes that enable individuals to make judgments and decisions under situational, motivational, and cognitive conditions that are less than optimal. Because decisions typically are made in a state of uncertainty, it has proven fruitful to identify

the strategies that individuals actually use to arrive at solutions that do not represent the best possible outcome but simply meet specific criteria. The study of these simplifying rules of thumb, called ‘judgmental heuristics,’ has been spearheaded chiefly by Tversky and Kahneman (1974). Because the concept of heuristics is associated closely with these authors, the application of this term is often limited to the specific mechanisms they have identified. However, the underlying logic of heuristic processing also applies to other strategies that serve the same purpose, that of simplifying human judgments and making them feasible under suboptimal conditions.

Because generally it is not possible to identify the use of a heuristic procedure from the outcome of a judgment, researchers have had to create specific conditions under which heuristics lead to errors. As a result, heuristics have often been associated with the supposed irrationality of human reasoning (e.g., Nisbett and Ross 1980). However, just as visual illusions do not testify to the deficiency of human perception, the errors produced by judgmental heuristics are not proof of the inadequacy of human judgment in general; instead, they point to mechanisms that allow individuals to make acceptable decisions under natural constraints.

The discussion that follows will introduce the three most prominent heuristics.

2. The Availability Heuristic

In assessing the frequency or probability of an event (or the co-occurrence of several events), individuals often employ a strategy that is based on the ease with which bits of information can be retrieved or generated from memory. An employer wishing to gauge the rate of unemployment in their community may go to the trouble of obtaining the relevant information from official sources. But if they are not motivated or able to do that, they can try to think of unemployed friends or acquaintances. The more easily they are able to do so, the higher will be their estimate of the rate of unemployment. Tversky and Kahneman (1973) called this judgment strategy the ‘availability heuristic.’

Additional judgmental phenomena in the social domain are connected to the availability heuristic. One example is a risk whose assessment depends on the frequency with which a type of event occurs. Thus, riding a motor bicycle is risky to the extent that accidents occur frequently. However, the actual frequency of an event and the ease with which it comes to mind may be dissociated if certain events are more likely to be reported in the media. For example, Lichtenstein et al. (1978) found that causes of death frequently reported in the press were greatly overestimated in terms of their frequency. While heart disease causes 85 percent more deaths than accidents,

only 20 percent of those surveyed thought that heart disease was the greater risk. The conspicuousness of events also influenced their availability. For example, in the study by Lichtenstein et al. (1978), the overestimated causes of death were especially dramatic and sensational (murder, flood, automobile accident), while the rather inconspicuous causes of death (heart disease, cancer, diabetes) were underestimated. Events that are known from personal experience are also more readily available (and therefore judged more likely to occur) than events that are only known through third parties (Greening et al. 1996).

It should be noted that 'availability' has two psychological components that usually are confounded: the content that comes to mind and the ease (or effort) experienced while retrieving the information from memory (Schwarz et al. 1991). Several studies from the domain of health psychology demonstrate the direct influence of experienced ease of retrieval on the assessment of risk. For example, Rothman and Schwarz (1998) found higher ratings of risk for contacting a heart disease when participants were asked to list three (easy to generate) instead of eight (difficult to generate) factors that increased their own risk.

Attitudinal judgments (see *Attitude Formation: Function and Structure*) about everyday topics are also made on the basis of the availability heuristic under certain conditions. Wänke et al. (1996) reported more positive attitudes towards public transportation on the part of subjects if they had been asked previously to generate three instead of seven arguments in favor of public transportation. These experimental examples were constructed such that the retrieval of more information was associated with less experienced ease of recall. The results suggest that under suboptimal conditions, judgment is based on the experience of information retrieval and not on the content of what is retrieved.

Another example refers to the ease with which we are able to imagine a different course of events. Let us look at the following example of counterfactual thinking: 'If I had gotten up five minutes earlier this morning I would not have missed the train, I would not have been late for the exam and would have been able to read the one additional problem I needed to pass the exam.' What is crucial is that the ease with which an event can be undone mentally influences affective reactions to this event. While the person in this example would presumably feel a great deal of anger over being five minutes late, a person who overslept an hour instead of five minutes would be less angry. For a detailed account of recent findings on contrafactual thinking see Roese and Olson (1995).

Apart from these examples of judgments and affective reactions on the basis of the perceived ease of cognitive operations, the 'availability principle' in its general form, that is, the finding that increased accessibility of contents (see *Priming, Cognitive Psy-*

chology of) and cognitive structures influences judgments, has stimulated a host of research in social psychology. This pertains, for example, to work on the categorization of persons, causal attribution (see *Attributional Processes: Psychological*), the constancy of opinions after they have been discredited, or the testing of hypothesis.

3. The Representativeness Heuristic

The representativeness heuristic refers to people's tendency to simplify categorical judgments by relying solely or excessively on similarity. For example, a person who wants to determine the profession of another person may use the target person's similarity to the typical member of this profession. Consequently, a student might decide that a fellow student at the next table in a university cafeteria is a business administration major if they display the characteristics of a typical MBA student, such as reading the business section of the paper and talking about the stock exchange. Although this strategy has a logical basis, it tends to neglect other types of relevant information, such as base rates. For example, this university may have many more students of law than of business administration. Information about such a frequency distribution should affect the judgment in the same way as information about the target person. However, research by Tversky and Kahneman (1982) has demonstrated that such base rates largely are neglected even if the individuating information is not very diagnostic.

This neglect of base rates is strikingly similar to findings from attribution research. A number of studies discovered that when people identify the causes of an observed behavior, they attribute it to characteristics of the person and neglect situational influences. This response tendency implies that judges underestimate 'consensus information' (Kelley 1967) that reflects the power of situations. 'Consensus information' describes how other people behave under the same circumstances and serves as a basic determinant of causal attributions (see *Attributional Processes: Psychological*). Specifically, if everybody shows the same behavior in a given context there is little reason to attribute an action to the unique characteristics of the actor. The tendency to neglect such base-rate information in social judgments and to give it less weight than it should have under normative considerations can be understood as a manifestation of the representativeness heuristic because it consists of drawing inferences on the basis of similarity ('a person who behaves aggressively is aggressive') at the expense of base-rate information (see *Person Perception, Accuracy of*).

The relationship between this attributional bias and heuristic processing was demonstrated subsequently by varying the conduciveness of the judgmental

situation (for a review, see Gilbert and Malone 1995). Researchers demonstrated that the neglect of situational influences depended on judges' cognitive resources, such that this bias was more likely to occur when people were distracted.

The use of the representativeness heuristic may also result in violations of logical principles. In one study by Tversky and Kahneman (1983), judges had to assess the probability of the joint occurrence of two characteristics of a target (a liberated woman). One feature was highly representative ('She is a feminist') while the other was not ('She is a bank teller'). In line with the representativeness heuristic, judges considered it less likely that the target person was a bank teller than that she was a feminist bank teller, which is, of course, logically impossible.

Other effects of the representativeness heuristic concern the misperception of chance and the neglect of sample size (Tversky and Kahneman 1974).

4. Anchoring and Adjustment

'Anchoring and adjustment' describes the phenomenon that judgments are assimilated toward a value that was initially considered. For example, a person who has to estimate the proportion of African nations in the UN may arrive at a higher percentage if they have been exposed previously to a high rather than a low standard of comparison. While such an influence would be hardly surprising if this 'anchor' is offered as a piece of information that is relevant to the judgment in question, Tversky and Kahneman (1974) were able to show that the resulting assimilation effect occurred even if relevance was ruled out by presenting the anchor as the outcome of a probabilistic process.

The psychological process that was suggested by Tversky and Kahneman (1974) for this phenomenon is an insufficient adjustment of the final judgment. Alternatively, Strack and Mussweiler (1997) have proposed a mechanism that is related to the availability heuristic, the so-called selective accessibility model. Here the anchoring effect involves two stages. In the first stage, judges engage in biased hypothesis testing when they consider the anchor as a possible value of the target. In this process, semantic information consistent with the anchor is activated. As a consequence of this selective activation, consistent information will be more accessible in a second stage when information for the final judgment will be retrieved. In this perspective, the anchoring effect is not the result of a numeric influence or a mere effect of insufficient adjustment but caused by a mechanism of semantic priming that has been demonstrated in many studies in social cognition (e.g., Higgins et al. 1977).

These supporting findings are not only important because they reflect the dynamics of many negotiation situations in which an initial offer greatly affects the

final outcome but also because they are helpful in understanding the consequences of merely considering a possibility before forming a judgment. For example, it has been demonstrated that explaining a hypothetical event leads people to expect its actual occurrence (Ross et al. 1977). Similar outcomes were obtained for imagining a particular event (Carroll 1978). More dramatically, it was shown that merely thinking about an attitude object may lead a person to adopt more extreme attitudinal positions if he or she holds a schema about the object (e.g., Tesser et al. 1995). Perhaps the most relevant social domain is that of social comparisons (see *Social Comparison, Psychology of*). Here, work on the anchoring heuristic suggests a mechanism that explains why and under what conditions comparing ourselves to another person will make us appear similar to that person (Mussweiler and Strack 2000). The anchoring effect has also been invoked to explain the so-called hindsight bias (Fischhoff 1982).

5. Other Heuristics in Social Psychology

In the previous sections three basic judgmental strategies were introduced that were identified by Kahneman and Tversky as devices employed to form sufficiently accurate judgments under adverse conditions and with little cognitive effort, although under certain circumstances these strategies can lead to systematic distortions. The idea of heuristic information processing has been expanded subsequently to other heuristic cues, with lasting influence on theory formation in broad areas of social psychology.

A prominent example is the use of feelings as information. Specifically, it has been argued that under suboptimal circumstances, judges resort to their subjective experiences to generate a judgment on a different dimension. This has been demonstrated for both affective (e.g., Schwarz and Clore 1983) (see *Emotions, Psychological Structure of*) and nonaffective feelings, such as familiarity and mental effort (e.g., Strack and Neumann 2000). As discussed before, it has been shown that the availability heuristic operates through the use of the mental effort experienced by an individual engaged in a specific mental operation (Schwarz et al. 1991).

The idea of heuristics has become effective in a second line of research in social psychology. The study of persuasive communication (see *Communication and Social Psychology*), in particular, has profited from the distinction between heuristic and systematic processing (Chaiken 1987). In this domain, the reliance on peripheral cues (such as the expertise of the communicator or the length of the message) instead of on central features (i.e., the strength of the arguments) was shown to occur under suboptimal conditions. Similarly, it has been shown that the use of stereotypes

(see *Stereotypes, Social Psychology of*) obeyed the same regularities, such that people were more likely to use them when they were unwilling or unable to process individuating information (Bodenhausen and Lichtenstein 1987).

A final application and extension of the heuristics idea can be found in the domain of social influence (see *Social Influence, Psychology of*). As Cialdini (1993) has argued, people's proclivity to simplify their judgments can be exploited by agents of social influence. For example, an advertising agency may deliberately attempt to increase the ease with which certain information (presumably about their own product) may be brought to mind or try to associate positive feelings with the product. Cialdini has identified and described many more such strategies, which he called the most effective 'weapons of influence.'

6. Conclusion and Future Directions

Over the years, the concept of heuristics has generated many insights into the cognitive dynamics of social behavior. At the same time, however, the idea of heuristic processing has gone far beyond the three mechanisms described by Kahneman and Tversky and it has been claimed that the use of heuristics may possess biologically adaptive value (Gigerenzer and Todd 1999). In a more general perspective, a universal psychological principle has emerged that sheds light on the flexibility of human information processing under divergent epistemic goals. However, the work of finding the laws of this flexibility and integrating it into psychological theorizing still needs to be done.

See also: Decision Making, Psychology of; Heuristics for Decision and Choice

Bibliography

- Bodenhausen G V, Lichtenstein M 1987 Social stereotypes and information-processing strategies: The impact of task complexity. *Journal of Personality and Social Psychology* **52**: 871–80
- Carroll J S 1978 The effect of imagining an event on expectations for the event: An interpretation in terms of the availability heuristic. *Journal of Experimental Social Psychology* **14**: 88–96
- Chaiken S 1987 The heuristic model of persuasion. In: Zanna M P, Olson J M, Herman C P (eds.) *Social Influence: The Ontario Symposium*. Erlbaum, Hillsdale, NJ, Vol. 5, pp. 3–39
- Cialdini R B 1993 *Influence. Science and Practice*, 3rd edn. Harper Collins, New York
- Fischhoff B 1982 For those condemned to study the past: Heuristics and biases in hindsight. In: Kahneman D, Slovic P, Tversky A (eds.) *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK, pp. 335–51

- Gigerenzer G, Todd P M 1999 *Simple Heuristics that Make Us Smart*. Oxford University Press, New York
- Gilbert D T, Malone P S 1995 The correspondence bias. *Psychological Bulletin* **117**: 21–38
- Greening L, Dollinger S J, Pitz G 1996 Adolescents' perceived risk and personal experience with natural disasters: An evaluation of cognitive heuristics. *Acta Psychologica* **91**: 27–38
- Higgins E T, Rholes W S, Jones C R 1977 Category accessibility and impression formation. *Journal of Experimental Social Psychology* **13**: 141–54
- Kelley H H 1967 Attribution theory in social psychology. In: Levine D (ed.) *Nebraska Symposium on Motivation*. University of Nebraska Press, Lincoln, NE, Vol. 15, pp. 192–240
- Lichtenstein S, Slovic P, Fischhoff B, Layman M, Combs B 1978 Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory* **4**: 551–78
- Mussweiler T, Strack F 2000 Consequences of social comparison: Selective accessibility, assimilation, and contrast. In: Suls J M, Wheeler L (eds.) *Handbook of Social Comparison: Theory and Research*. Plenum, New York, pp. 253–70
- Nisbett R E, Ross L 1980 *Human Inference: Strategies and Shortcomings of Social Judgment*. Prentice-Hall, Englewood Cliffs, NJ
- Roeser N J, Olson J M 1995 *What Might Have Been: The Social Psychology of Counterfactual Thinking*. Erlbaum, Mahwah, NJ
- Ross L, Lepper M, Strack F, Steinmetz J L 1977 Social explanation and social expectation: The effects of real and hypothetical explanations upon subjective likelihood. *Journal of Personality and Social Psychology* **35**: 817–29
- Rothman A J, Schwarz N 1998 Constructing perceptions of vulnerability: Personal relevance and the use of experiential information in health judgments. *Personality and Social Psychology Bulletin* **24**: 1053–64
- Schwarz N, Bless H, Strack F, Klumpp G, Rittenauer-Schatka H, Simons A 1991 Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology* **61**: 195–202
- Schwarz N, Clore G L 1983 Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology* **45**: 513–23
- Strack F, Mussweiler T 1997 Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology* **73**: 437–46
- Strack F, Neumann R 2000 Furrowing the brow may undermine perceived fame: The role of facial feedback in judgments of celebrity. *Personality and Social Psychology Bulletin* **26**: 762–8
- Tesser A, Martin L L, Mendolia M 1995 The impact of thought on attitude extremity and attitude-behavior consistency. In: Petty R, Krosnick J A et al. (eds.) *Attitude Strength: Antecedents and Consequences*. Erlbaum, Mahwah, NJ, pp. 72–92
- Tversky A, Kahneman D 1973 Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* **54**: 207–32
- Tversky A, Kahneman D 1974 Judgment under uncertainty: Heuristics and biases. *Science* **185**: 1124–31
- Tversky A, Kahneman D 1982 Judgments of and by representativeness. In: Kahneman D, Slovic P, Tversky A (eds.) *Judgments under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York, pp. 84–98
- Tversky A, Kahneman D 1983 Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review* **90**: 293–315

Wänke M, Bless H, Biller B 1996 Subjective experience versus content of information in the construction of attitude judgments. *Personality and Social Psychology Bulletin* 22: 1105–13

F. Strack

Hidden Curriculum in the Classroom

Long after I've forgotten to use the power rule to derive an equation, or that the animals are classified by kingdom, phylum, class, order, family genus and species ... I will remember the other lessons I've learned, the lessons that taught me how to live (Jenny Lobasz, senior student at Dreyfoos School of the Arts).

The hidden curriculum generally refers to student learning that takes place in school as a result of actions by school personnel, parents, students, and other community agents. What is learned in the classroom is largely a function of the prevailing school conditions and classroom dynamics as well as such demographic factors as the social class, race, ethnicity, language, gender, and condition of disability of all school participants. Compared with the formal school curriculum, the hidden curriculum is not written, does not have explicit objectives, and varies considerably from one school setting to another. The hidden curriculum often referred to as the 'covert' or the 'informal' curriculum, is a most powerful influence in the classroom; it is estimated that up to 90 percent of what is learned is due to this influence (Massialas 1996).

1. The School and Classroom Dynamics

In many cases the formal curriculum is antithetical to the hidden curriculum. For example, the formal curriculum of the school explicitly promotes such values as academic learning, fairness, and good citizenship; the latter expressed as democratic participation in decisions affecting the individual. The hidden curriculum, on the other hand, teaches the students how to manipulate the system, i.e., how to cheat without being caught, how to 'apple polish the teacher,' and how to comply with the rules without questioning the legitimacy of the authorities for making them (Jackson 1983). The formal curriculum preaches democracy but the hidden curriculum imposes autocracy. The formal curriculum emphasizes scholarship and an equitable grading system but the hidden curriculum teaches students that scholastic attainment expressed through grades is a political process of give and take and of bargaining with the teacher. In this environment it is clear that if students do not master the lessons and the techniques of the hidden curriculum they are destined to fail.

Jackson (1983), being one in the first groups of researchers who studied the hidden curriculum in depth concluded that 'Learning how to make it in

school involves, in part, learning how to falsify our behavior' (p. 50). He found out that teacher–student relations in the classroom are hierarchical—the teacher being at the top of the hierarchy and the students at the bottom. He observed that classroom activity in elementary schools usually centers on seatwork, group discussion, teacher question and answer sessions, as well as teacher demonstrations. Students may learn about the three Rs but they also learn other things, things that are not spelled out in formal instruction. For example, they learn to work in crowds and they learn to accept praise or reproof. They also learn how power is distributed in the system, that is, how certain students gain special privileges. Students also learn how to be patient and how to conform to institutional requirements and to tolerate minor frustrations.

Many studies confirm the claim that schools indirectly teach students to be obedient, passive, and compliant. In this they are supported by other social institutions that directly relate to the life of the students, e.g., the family. One author recognizes that, on the one hand, modern society seeks to develop self-reliant and critical thinking individuals. On the other, it promotes, through the schools and other socialization agencies, obedience. 'Little notice is taken of the legions of overly obedient children in the schools; yet, for every overly disobedient child, there are probably twenty who are obeying too much. There is little motivation to encourage the wheels to develop as noisy, creative, independent thinkers who may become bold enough to disagree' (McCarthy 1998, pp. 56–7).

A review of research on the influence of the hidden curriculum in schools in the Arab world revealed that these schools, much like their Western counterparts, impart passivity and obedience among the students (Massialas and Jarrar 1991). For example, in a study conducted in Morocco it was found that the 'primary school imparts the values and behaviors of the society, especially the home—absolute dependence on patriarchal commands, suppression of creative thought, strict compliance with rules, both written and unwritten, acceptance of things as they are' (Massialas and Jarrar 1991, p. 137). Similar values to those transmitted in the Moroccan schools were reported in studies in other Arab schools, e.g., in Saudi Arabia, in Lebanon, in Qatar. On the basis of these studies it is apparent that the Arab classroom replicates to the extent possible the patriarchal family structure. The teacher is invariably the dominant figure, the patriarch who conveys awe, respect, and obedience.

2. Social Class and the Hidden Curriculum

Student background factors such membership in a social class invariably affect how students are treated in school and their relations with the teachers and other school personnel. A study conducted in New Jersey, USA, found that social class clearly influenced

the values learned in school through the hidden curriculum (Anyon 1983). The study identified four types of schools—working class, middle class, affluent professional, and executive elite. Working class schools emphasized mechanistic, step-by-step instruction conducive to memory type of work. Class work was quite structured. There was very little in trying to provide an explanation of the world. The middle class schools emphasized the ‘right answer’; instruction was largely teacher-directed; the majority of the lessons were drawn from the textbook. In the professional-affluent schools the emphasis was on fostering creativity and independent thought in the classroom. Students were constantly asked through projects to apply concepts in explaining their world. The executive elite school emphasized critical thinking, conceptual development, and decision making. Instruction stressed analysis of social issues and problem solving. The study concluded with the observation that the working-class schools prepared students to assume positions as laborers, whereas the professional and executive schools prepared students to be intellectuals and professionals. Thus the school, in a major way, reproduced the larger system of unequal social relations. One of the few advantages that students in the working-class schools could have was learning the ‘methods of resistance.’ ‘These methods are highly similar to the “slowdown,” subtle sabotage, and other modes of indirect resistance carried out by adult workers in the shop ...’ (Anyon 1983, p. 163).

Several other research studies on the influence of social class in learning confirm Anyon’s conclusions above. Although some studies recognize that there are variations in community and school expectations of what students should learn, there is evidence that working-class communities placed their priority ‘on acquiring credentials and following rules, whereas the higher-SES communities ... placed emphasis on students’ gaining particular kinds of knowledge and skills’ (Knapp and Woolverton 1995, p. 553). The research clearly indicates that in the working-class school students are taught through the hidden curriculum ‘that little is expected of them ... they are not highly valued; they cannot be trusted to guide their own learning or conduct themselves responsibly in the class, and they must accept low-level positions in a rigidly controlled hierarchy’ (Knapp and Woolverton 1995, p. 553). The higher-status school, on the other hand, encourages students to develop a greater sense of self-worth, higher expectations for school success, and a wider range of possibilities for fitting into hierarchies of control at different levels’ (Knapp and Woolverton 1995, p. 563).

3. Cultural Minorities and the Hidden Curriculum

Students who are members of ethnic, racial, or linguistic minorities are usually adversely affected by the hidden curriculum of the schools. The grading

system, the standardized tests, the reward and punishment measures, the textbooks, the school rules, all influenced by the hidden curriculum, normally have a negative effect on minorities. Available research clearly indicates that these and other educational instruments discriminate against minorities. For example, about 45 percent of Hispanics in the United States are expected to drop out from high school (Sleeter E et al. 1998). ‘In-school factors such as poor academic performance and disliking school account for 56 percent of dropouts among Black, male students’ (Massialas 1996, p. 69). School suspensions and expulsions, strong instruments of the hidden curriculum, also impact adversely racial, ethnic, and linguistic minorities. There is usually a disproportionate number of students from minority groups being punished through these measures. Language minority students are also adversely affected by the hidden curriculum in the classroom. For instance, one study on the topic indicated that in a classroom which had a mixture of mainstream and limited English proficient students, ‘the teachers interacted less frequently with the ESL students than their native-English counterparts ... the interactions with ESL students tended to be managerial in nature rather than instructional’ (McKeon 1994, pp. 25–6). Also reported in the literature is the negative result that ensues from the disability of teachers to understand the flow of discourse between them and students who have different patterns of linguistic interactions. ‘... when the patterns of language use or the patterns of social interactions that children bring to the school are ignored or denigrated by their teachers, the ensuing relationship may threaten the students’ academic development’ (Pease-Alvarez and Vasquez 1994, p. 93).

4. Gender and the Hidden Curriculum

The hidden curriculum has a very strong influence in school learning when mediated through gender. Gender-appropriate roles are conveyed through the formal curriculum as well as the hidden. Upon reviewing the works on the subject, Banks (1995, p. 622) stated, ‘Research indicates that sex-role attitudes and gender associations develop early, and that teaching materials, the mass media, and society at large often reinforce sex-role stereotyping’

5. Condition of Disability and the Hidden Curriculum

Students with learning disabilities, e.g., students with visual or hearing impairment, emotional disturbance are also affected by the hidden curriculum. Whether placed in classes with mainstream students (inclusion classes) or in special education classes there is always a built-in bias in their relations with teachers, other

students and in the school-related expectations which influence in a significant way what is to be learned, how, and for what purposes. There is an overrepresentation of cultural minorities in the special education programs that clearly point to the influence of the hidden curriculum in student selection and placement.

6. Conclusion and Outlook

Recently there have been many attempts to attend to the negative influences on learning of the hidden curriculum. Programs for teachers seek to make them aware of these often-negative influences and to take appropriate measures to alleviate them from the school settings. Instructional approaches such as 'cooperative learning,' 'peer tutoring,' and portfolio assessment are all attempts to combat the negative effects of the hidden curriculum on student learning.

See also: Education and Income Distribution; Educational Institutions and Society; Educational Systems: North America; Gender and School Learning: Mathematics and Science; Group Processes in the Classroom; Minorities; Obedience: Social Psychological Perspectives; School as a Social System; Self-concepts: Educational Aspects

Bibliography

- Anyon J 1983 Social class and the hidden curriculum of work. In: Giroux H, Purpel D (eds.) *The Hidden Curriculum and Moral Education*. McCutchan, Berkeley, CA, pp. 146–67
- Banks J A 1995 Multicultural education: Its effects on students' racial and gender role attitudes. In: Banks J A, Banks C A Mc (eds.) *Handbook of Research on Multicultural Education*. Macmillan, New York, pp. 617–27
- Jackson P 1983 The daily grind ... In: Giroux H, Purpel D (eds.) *The Hidden Curriculum and Moral Education*. McCutchan, Berkeley, CA, pp. 28–60
- Knapp M S, Woolverton S 1995 Social class and schooling. In: Banks J A, Banks C A Mc (eds.) *Handbook of Research on Multicultural Education*. Macmillan, New York, pp. 548–69
- Lebasz J 1999 Learning the lessons of living. *Sun-Sentinel August 31*; 1E, 7E
- Massialas B G 1996 The hidden curriculum and social studies. In: Massialas B G, Allen R F (eds.) *Critical Issues in Teaching Social Studies*, K-12. Wadsworth, Belmont, CA, pp. 61–79
- Massialas B G, Jarrar S A 1991 *Arab Education in Transition*. Garland, New York
- McCarthy S J 1998 Why Johnny can't disobey. In: Shapiro H S, Harden S B (eds.) *The Institution of Education*. McCutchan, Berkeley, CA, pp. 28–60
- McKeon D 1994 Language, culture, and schooling. In: Genesee F (ed.) *Educating Second Language Children*. Cambridge University Press, Cambridge, UK, pp. 15–32
- Pease-Alvarez C, Vasquez O 1994 Language socialization in ethnic minority communities. In: Genesee F (ed.) *Educating*

Second Language Children. Cambridge University Press, Cambridge, UK, pp. 82–102

B. G. Massialas

Hierarchical Models: Random and Fixed Effects

1. The Basic Statistical Model and its Probability Interpretation

The model that has dominated statistical practice recently supposes that data x have, for each value ψ of a parameter, a known probability distribution dependent on the unknown ψ . The distribution may be described by the probability density

$$p(x|\psi) \quad (1)$$

where p is a generic symbol for a density of the quantity, here x , before the bar; the symbol after the bar describing the conditions applying to the distribution, here the supposed value ψ of the parameter. Statistical inference then consists in making some statement about the parameter on the basis of the data; such as a point estimate for ψ , or a test that ψ assumes specified values. An important, special case has x a vector $(x_1, x_2, \dots, x_n)$, of n components, and ψ chosen such that, given ψ , the components are independent and identically distributed (i.i.d), when Eqn. (1) assumes the special form

$$\prod_{i=1}^{i=n} p(x_i|\psi) \quad (2)$$

The Bayesian approach would include, in addition to the density (1), a probability distribution for ψ expressed through a density $p(\psi)$. Statistical inference then consists of calculating $p(\psi|x)$ by Bayes's theorem, inverting the roles of x and ψ in (1).

A distinction is often made between the nature of the probabilities involved in Eqn.(1), and more particularly Eqn. (2), and those in $p(\psi)$. In view of the repetitions under similar circumstances of $x_1, x_2, \dots, x_n$, their probability is associated with frequency and the x_i thought of as successive realizations of a single random variable. Alternatively, $(x_1, x_2, \dots, x_n)$ is viewed as a random sample from a population. In contrast, ψ is a fixed, but unknown, quantity and, when introduced, $p(\psi)$ expresses beliefs about the values it might assume, without reference to frequency or populations. The disparity between random, for x , and fixed, for ψ , has played an

important role in the development of hierarchical models, leading to different inferences depending on whether the effect is supposed fixed or random. In the Bayesian approach there is no distinction and all probabilities express beliefs; the special case Eqn. (2) reflecting the belief in exchangeability among the observations x_i .

2. The Hierarchical Model

To illustrate the need for a hierarchical model, consider data consisting of repeated observations divided into groups. Let x_{ij} denote the j th observation in group i ($i = 1, 2, \dots, m; j = 1, 2, \dots, n$), with n observations in each of m groups, and suppose that within each group the values are conditionally i.i.d with distributions that may vary from group to group. A popular, special case has

$$x_{ij} \sim N(\theta_i, \sigma^2) \quad (3)$$

with the x_{ij} being independent given $\psi = (\theta_1, \theta_2, \dots, \theta_n, \sigma^2)$, so that they are normally distributed with common variance σ^2 but with means that depend on the group. This model is the basis for one-way analysis of variance where inference may well concentrate on whether the means are truly different and, if they are, of what these differences consist. As it stands, this model is just a special case of Eqn. (1) with x the set of x_{ij} and ψ as above; but now contrast two cases that differ in respect of the nature of the groups. In the first, the study is one investigating the potency of a drug and the groups correspond to dilution levels. In the second, the milk yields of cows are measured and the groups correspond to different sires. In both cases, it may be reasonable to think of the duplicate measurements within a group as i.i.d, forming a random sample from a possible population, in agreement with Eqn. (3). But there are differences between the nature of the groups; for in the dilution experiment, the levels are fixed, whereas, with the sires, those selected may be thought of as a sample from a population of sires. In the language of the last paragraph, the θ_i are fixed in the dilution experiment, but random with the cows. In the latter, it might therefore be reasonable to suppose that the means are capable of a frequency description and a probability distribution. A possible, simple case is to suppose that, in addition to Eqn. (3),

$$\theta_i \sim N(\mu, \tau^2) \quad (4)$$

with the θ_i being independent given μ and τ^2 . The frequency interpretation is inappropriate in the dilution experiment, where the levels have been organ-

ized, for example at levels $1, \frac{1}{2}, \frac{1}{4}$ etc. It is also possible that inferences in the two experiments could be directed to different ends. Thus the variability among sires expressed through τ^2 is meaningful, but not in the dilution experiment. There are many other possibilities besides Eqn. (4) for the second-stage distribution. For example, one group may be a control and the remainder correspond to new treatments. In this case, the distribution for the control mean may differ from that for the others. Outside the Bayesian view, only the latter may have a frequency interpretation. It is this flexibility that helps make hierarchical modeling so attractive.

The probability judgements in Eqns. (3) and (4) together form a hierarchical model with two stages to the hierarchy, the first expressing the dependence of the data on the group means in Eqn. (3), the second describing the variability of those means in Eqn. (4). The hierarchical model has been thought by some to be appropriate in studies of milk yield, but not in dilution studies, because the means are random in the former but fixed in the latter. The general form of a two-stage, hierarchical model has data x dependent on parameters (θ_1, ϕ_1) expressed through

$$p(x|\theta_1, \phi_1) \quad (5)$$

at the first stage, with θ_1 having a distribution

$$p(\theta_1|\theta_2, \phi_2) \quad (6)$$

at the second stage. The parameters of the second stage, θ_2, ϕ_2 are often called hyperparameters. Notice that in this formulation some of the parameters at the first stage are fixed, ϕ_1 , whilst others, θ_1 , are random and have their probability structure expressed in the second stage in terms of the fixed hyperparameters. In the normal example, Eqns. (3) and (4), θ_1 here corresponds to the group means there, and ϕ_1 to the common variance. The hyperparameters are $\theta_2 = \mu, \phi_2 = \tau^2$, the values σ^2 and τ^2 describing the within- and between-groups variances. There may not be any fixed parameters at the first stage, as when the within-variances are supposed different, at σ_i^2 , and given a distribution at the second stage, as with the means. It is possible to extend the number of stages in a hierarchical model, for example by adding a third stage to Eqns. (5) and (6) by supposing θ_2 depends on (θ_3, ϕ_3) expressed through

$$p(\theta_2|\theta_3, \phi_3) \quad (7)$$

Such an extension would be appropriate for a two-way analysis of variance; the first stage describing the variation within cells of the two-way layout, the second concerning the row and column means (the main

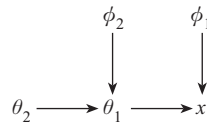
effects), the third involving the interactions between the main effects. Any Bayesian model is, in a sense, hierarchical, the two stages being $p(x|\psi)$ and $p(\psi)$ respectively, but the terminology is usually confined to the models just described. A further extension is to include regression effects. For example, Eqn. (5) may be extended to

$$p(x|z, \theta_1, \phi_1)$$

where z is an observed regressor variable. Similar regression effects may be introduced at other stages in the hierarchy.

Returning to the two-stage model in Eqns. (5) and (6), notice that x and θ_1 are random, whereas ϕ_1, θ_2 and ϕ_2 are fixed, and the stages describe the joint distribution of the random quantities, given the fixed ones, θ_2 and ϕ_2 having no effect on x except through θ_1 . This lack of effect can be expressed in the form of a directed graph, or influence diagram, which clarifies the hierarchical form. The two-stage model is represented in the figure, the lack of links from x to θ_2 and ϕ_2 expressing the feature. The extension to more stages is easily apparent.

As with the basic model in Eqn. (1), the Bayesian approach would add a probability distribution $p(\phi_1, \theta_2, \phi_2)$ for the fixed parameters and hyperparameters. A joint distribution for all the quantities, random and fixed, is not thereby specified unless the additional assumption is made that, given (θ_1, ϕ_1) , x is independent of (θ_2, ϕ_2) . This assumption is usually made, often tacitly. The independence assumption is also reflected in the directed graph by the lack of links already mentioned. There may be additional links unless the quantities $(\theta_1, \theta_2, \phi_2)$ are independent. Since the probability structure is complete, it follows that every quantity has its probability structure defined, under any conditions expressed in terms of the quantities of the problem, which probability can be found following the usual rules of the probability calculus. In particular, the posterior distribution of any of the parameters or hyperparameters, given the data, may be determined.



3. Inference in Hierarchical Models

We now return to the simple case where the observations are divided into m groups, each of size n , and possess the normal, independent distributions des-

cribed by Eqn. (3). Suppose that the common variance within groups, σ^2 , is known. The extension to unknown variance presents complications that do not affect the general thrust of what follows. If the means $\{\theta_i\}$ are fixed, then (3) completes the specification of the model and inference about them is ordinarily performed either by least-squares, or by maximum likelihood, the two methods here being equivalent, with the result that θ_i is estimated by the group mean

$$x_{i.} = \sum_{j=1}^n x_{ij} / n \quad (8)$$

with variance σ^2/n . Suppose, in contrast, the means are random and this randomness is expressed by the second-stage assumption (4), where again τ^2 , the between-group variance, is supposed known. Applying either of the inferential methods here gives another estimate for θ_i , namely

$$\left(\frac{nx_{i.} + x_{..}}{\sigma^2 + \tau^2} \right) / \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2} \right) \quad (9)$$

with variance $(n/\sigma^2 + 1/\tau^2)^{-1}$, in the form of a weighted average of the group mean $x_{i.}$, which was the estimate in the fixed case, and the overall mean $x_{..} = \sum x_{ij}/m$, the weights being inversely proportional to the variances, σ^2/n and τ^2 respectively. It is often said that the effect of randomness in the group means is to shrink the fixed estimates towards the overall mean. An important feature of the shrinkage is that the apparently best (worst) groups with highest (lowest) values of $x_{i.}$ will have their estimates decreased (increased) and, as a result, the group differences will be diminished. Notice that if τ^2 is large, the shrunk means approach the usual ones. The shrinkage feature is important in comparisons of the θ_i , for example in the construction of league tables (Goldstein and Spiegelhalter 1996), or in census returns for different communities (Ericksen et al. 1989).

In the Bayesian paradigm all the unknown quantities will have a probability distribution, in particular $(\theta_1, \theta_2, \dots, \theta_n, \sigma^2)$ in the normal, one-way case in Eqn. (3), irrespective of whether the means are fixed or random. One possibility is for the θ_i to be provided with the second-stage distribution (4) and finally (μ, σ^2, τ^2) given a distribution. If, as above, the variances are known, only the distribution of μ needs additionally to be specified. If it too is supposed normal and the variance taken to be large in comparison with σ^2 and τ^2 , Eqn. (9) is the mean of θ_i , given the data (and σ^2, τ^2). Thus the Bayesian approach favors the shrunk estimate whenever the second-stage distribution is thought to be of the form in Eqn. (4), which might be appropriate for sires but not with the dilution experiment. For a Bayesian, there is no real distinction between the fixed and random effects; what differences

there are depends on the specific form of the second-stage distribution, and perhaps on the quantities of interest. For example, interest may centre on τ^2 , the variance between means, rather than on the individual values θ_i . The situation is in marked contrast to other views where, as we have seen, the estimates of the means depend on whether or not the means are fixed or random; also the sample variance-components in the analysis of variance have different interpretations, in the fixed case being a measure of the differences between the specific means, in the random case referring to the population variance τ^2 .

Returning to the general, statistical model in Eqn. (1) with data x and parameters ψ , it is often hard, especially when the data are numerous, to specify the required density $p(x|\psi)$. In general, whenever there are a number of quantities $u_1, u_2, \dots, u_s$, it is difficult to describe a joint density $p(u_1, u_2, \dots, u_s)$; an easier course often being to think about the quantities in some order and structure a series of conditional distributions $p(u_1)$, $p(u_2|u_1)$, $p(u_3|u_1, u_2)$ etc. This is essentially a hierarchical model except that, as explained earlier, it is usual to have the distribution of one quantity dependent only on the immediately previous one and not on all the previous ones; $p(u_3|u_2)$ without reference to u_1 . In general, the order in which the quantities are introduced may be arbitrary, but in the hierarchical case this does not arise since the order is naturally dictated by the situation. For these reasons, hierarchical models have proved very popular in the description of complicated cases, enabling them to be broken down into stages, each of which is simple, and yet where they combine together to form a whole that reasonably describes complicated reality.

4. Types of Hierarchical Models

There are few hierarchical models in which it is possible to obtain explicit, analytical results for the inference desired. An exception is the normal, linear model of which Eqns. (3) and (4) provide a special case. Generally, at the first stage

$$x \sim N(A_1\theta_1, C_1) \quad (10)$$

where x ($n \times 1$) and θ_1 ($s \times 1$) are vectors, and A_1 ($n \times s$) and C_1 ($n \times n$) are matrices. A_1 is known, θ_1 is unknown. Usually C_1 is known, the simplest case, or known up to an unknown multiplier, though it is possible to describe the distribution of its elements in a second stage. Equation (10) expresses the usual linear dependence of x_1 on θ_1 and is basic to least squares, where θ_1 is estimated by $(A_1^T C_1^{-1} A_1)^{-1} A_1^T C_1^{-1} x$. The second stage imposes a similar structure on θ_1 ,

$$\theta_1 \sim N(A_2\theta_2, C_2)$$

with obvious extensions to further stages as required. Explicit results are provided by Lindley and Smith (1972) using Bayesian ideas, though many of the results can be interpreted sensibly without invoking a distribution on the final hyperparameters, as was seen in the special case.

Similar ideas had previously been used, starting in the 1960s, by Henderson in genetics and specifically in animal breeding; see Henderson (1984). These were outside the usual statistical paradigms and remained unappreciated in broader contexts. An admirable statistical survey of Henderson's and related work is provided by Robinson (1991).

Another field in which explicit results have been obtained, though usually of an asymptotic form, is that of Empirical Bayes. This is not always considered a hierarchical model but does fit into the general framework above. We illustrate with a simple case again involving several groups and data $\{x_{ij}\}$ but this time with independent, Poisson distributions of means μ_i , the same for all observations in group i . This describes the first stage of the hierarchy. At the second stage, the μ_i are supposed i.i.d with some unknown density $p(\mu)$. The model might be suitable when the groups refer to runs in an industrial process. A distinction from the model above is that hyperparameters are avoided and a nonparametric approach adopted at the second stage; no restrictions, other than of a mild nature like continuity, being placed on $p(\mu)$. The observed distribution of the group means $\bar{x}_i$ provides information about $p(\mu)$. Despite the nomenclature, the analysis is ordinarily not Bayesian. A good account is provided by Maritz and Lwin (1989). The term has also been used in connection with the equation-based approach mentioned later in Sect. 5.

Hierarchical models have also been successfully applied to stochastic processes. These may be spatial but here an exposition in the more usual temporal terms will be used. It is often reasonable to suppose that there is an underlying development process taking place over time, but that one's observation of it is disturbed by error, these two features providing the two stages of the hierarchy. A simple example is to suppose an observation x_t at time t depends on the state θ_t of the process then, with a superimposed error ε_t ,

$$x_t = \theta_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, v_t), \quad (11)$$

whereas the process is linear, autoregressive with

$$\theta_t = a_{t-1}\theta_{t-1} + \eta_t, \quad \eta_t \sim N(0, w_t), \quad (12)$$

$t = 1, 2, \dots$. Equation (11) is termed the observational equation, (12) the system equation; the two sets of error terms being supposed individually independent and independent of each other. The approach is

usually Bayesian with $\{v_i\}$, $\{w_i\}$ and $\{a_{i-1}\}$ supposed known, at least in the simpler cases, and an initial distribution supposed for θ_0 . If this last is normal, it is straightforward to update by Bayes to obtain successively the distributions of θ_i given $x^{(i)} = (x_1, x_2, \dots, x_i)$, the observations up to time i . Often the primary purpose of the study is to forecast the future state of the process, θ_{t+1} , given the current data $x^{(t)}$,

$$p(\theta_{t+1}|x^{(t)}) = \int p(\theta_{t+1}|\theta_t)p(\theta_t|x^{(t)})d\theta_t$$

involving the system equation and the posterior distribution of θ_t . A full account is provided by West and Harrison (1989).

As previously mentioned, hierarchical models are especially useful in describing complicated, but realistic, situations, where the individual stages in the hierarchy are relatively easy to assess. Nevertheless, apart from the cases outlined here, analytical results are difficult to obtain because the integrals that arise in the probability calculations are not expressible in standard forms, so that resort has to be made to numerical methods. In one sense this is no hardship, since the data are already numeric and in any application one only requires answers for that specific data set. On the other hand, the integrals are typically in high dimensions where numerical results are hard to come by. An important, recent advance has been to use Markov-chain, Monte Carlo (MCMC) methods which themselves depend, like the hierarchical models, on conditional distributions. There has therefore been a recent explosion of papers using MCMC to handle these models. A summary is given by Brooks (1998).

5. Equation-based Approaches

Hierarchical models are often described in an alternative terminology as multilevel, our stages being replaced by levels. An example is provided by students, at the first level, being in classes at the second, which form schools at the third. It is also common to alter the emphasis and not talk in terms of probability, as here, but in terms of equations which incorporate a probabilistic element, so that Eqn. (3) would be written

$$x_{ij} = \theta_i + \varepsilon_{ij} \quad (13)$$

with ε_{ij} being an error term, of zero mean. Complete adherence to Eqn. (3) would in addition require the error terms to be independent, normal and of constant variance. It is usual to employ matrix notation, as in Eqn. (10), and a popular model is

$$y = X\beta + Zu + \varepsilon \quad (14)$$

Here y , replacing x , is the observed data, dependent on observed covariates X and Z ; whilst β is a set of unknown, fixed parameters, u a set of random quantities and ε the error term. Thus ε is at the first level and u at the second level. A further complication, that X is only observed with error, is sometimes included, so that

$$X = x + \eta$$

in terms of the unknown, true value x and error η at another level. This generalizes the problem of regression with both quantities subject to error. We shall refer to this whole approach as equation based, rather than probability oriented, though it is possible to pass from one to the other, as between Eqns. (3) and (13). These ideas have been used in longitudinal studies where several observations are made on an individual over time, forming one level of variation, another level being that between individuals.

The distinction between the approaches is not confined to the model description but includes the analysis, where manipulations of equations replace the probability calculus. Thus Eqns. (14) leads to estimates $\hat{\beta}$, $\hat{u}$, $\hat{\beta}$, u given by

$$X^T X \hat{\beta} + X^T Z \hat{u} = X^T y$$

and

$$Z^T X \hat{\beta} + (Z^T Z + I) \hat{u} = Z^T y$$

under simple assumptions about u and ε . Some of the methods that are used do not employ the full probability structure but only refer to the first and second moments, a popular variant of least squares being quasi-likelihood. Other methods are based more closely on likelihood and variants thereof and often embrace ideas associated with the general, linear model. Lee and Nelder (1996) provide a good account with a lively discussion in which R. L. Smith suggests disagreement between their results and those produced by the fully Bayesian approach. This is not surprising since, as has been shown in Sect. 3, Bayesian procedures employ more shrinkage than some other methods.

An issue of the *Journal of the Royal Statistical Society A* (162, 1999, 269–448) is devoted to articles on ‘Applications of random effects/multilevel models to categorical data in social science and medicine’ and provides many examples of the ideas here briefly described. The book by Goldstein (1995) is relevant; he and his colleagues have provided a computer package (1998) to handle many models. Similar ideas have been put forward by Bryk and Raudenbush (1992). Best et al. (1996) present a Bayesian approach,

whilst Spiegelhalter et al. (1998) supply some computer facilities.

Bibliography

- Best N G, Spiegelhalter D J, Thomas A, Brayne C E G 1996 Bayesian analysis of realistically complex models. *Journal of the Royal Statistical Society A* **159**: 323–42
- Brooks S P 1998 Markov chain Monte Carlo method and its application. *The Statistician* **47**: 69–100
- Bryk A S, Raudenbush S W 1992 *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage, Newbury Park, CA
- Ericksen E P, Kadane J B, Tukey J W 1989 Adjusting the 1980 census of population and housing. *Journal of the American Statistical Association* **84**: 927–44
- Goldstein H 1995 *Multilevel Statistical Models*. Arnold, London
- Goldstein H, Rasbash J, Plewis I, Draper D, Browne W, Yang M, Woodhouse G, Healy M 1998 *A User's Guide to MLwiN*. Institute of Education, London
- Goldstein H, Spiegelhalter D J 1996 League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society A* **159**: 385–443
- Henderson C R 1984 *Applications of Linear Models in Animal Breeding*. University of Guelph, Ontario, Canada
- Lee Y, Nelder J A 1996 Hierarchical generalized linear models. *Journal of the Royal Statistical Society B* **58**: 619–78
- Lindley D V, Smith A F M 1972 Bayes estimates for the linear model. *Journal of the Royal Statistical Society B* **34**: 1–41
- Maritz J S, Lwin T 1989 *Empirical Bayes Methods*. Chapman and Hall, London
- Robinson G K 1991 That BLUP is a good thing: the estimation of random effects. *Statistical Science* **6**: 15–51
- Spiegelhalter D J, Thomas A, Best N G 1998 *WinBUGS User Manual, Version 1.1.1*. Medical Research Council Biostatistics Unit, Cambridge, UK
- West M, Harrison J 1989 *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York

D. V. Lindley

Hierarchies and Markets

Whether the first market entailed the exchange of nuts for berries on the edge of the forest or the first hierarchy involved leadership for the hunt of a wild boar, it is clear that markets and hierarchies go back to antiquity. It is also clear that organizational innovations in both markets and hierarchies have contributed significantly to economic progress.

As between the two, markets have the better reputation among economists. That goes back to Adam Smith (1922) and continues today. Thus, whereas Smith celebrated autonomy, specialization,

and the marvels of the market, he had grave concerns over the separation of ownership and control in large joint stock companies and the excesses of managerial discretion that accrued thereto. Indeed, but for grudging acceptance of economies of scale due to technology, hierarchy has been under a cloud since. Only recently has a symmetrical approach to organizations of all kinds—markets, hierarchies, bureaus—begun to emerge.

1. Foundations

The neoclassical theory of the firm that had taken shape by the 1930s described the firm in technological terms—as a production function—to which a profit maximization purpose was ascribed. The adverse consequences of monopoly, monopsony, and monopolistic competition were all readily displayed with this abstract apparatus, but issues of organization—hierarchy—were scanted by this black box construction.

That, however, began to change with the appearance of radical new ideas about organization by economists (Coase 1937, Commons 1932), lawyers (Llewellyn 1931), and organization theorists (Barnard 1938). Coase advanced the proposition that firms were not pre-existing technological entities but that firms and markets were alternative modes of organization. Whether the firm should make-or-buy was, thus, not technologically determined (although the condition of technological separability between stages was relevant) but needed to be derived. Coase further averred that the choice between market and hierarchy was decided principally by transaction cost differences. The canonical problem, then and since, for the theory of markets and hierarchies, was that of vertical integration.

In principle, this choice was easy: holding technology constant, the firm should produce to its own needs when the transaction costs in the firm were less than in the market, and it should buy in the market when this difference was reversed. Awaiting, however, a predictive theory of when transaction costs went one way rather than another, that remained a tautological construction. Thus, although after-the-fact choices could always be rationalized in transaction cost terms, that was undisciplined and such ad hoc explanations were understandably met with grave skepticism. The twin ideas of treating firms and markets as alternatives and of appealing to transaction costs to explain the allocation of activity between them nevertheless turned out to be very fruitful.

Coase went on, moreover, to pose two deep puzzles. If the market is a marvel, why do we see so much activity organized in firms? And if the firm enjoys management advantages over markets, why is not everything organized in one large firm? Any would-be

theory of firms and markets needed to come to terms with both of these issues.

Commons' approach to economic organization was different than but also complementary to that of Coase. Commons was intensely aware of the tensions that beset organization and averred that 'The ultimate unit of activity...must contain in itself the three principles of conflict, mutuality, and order. This unit is a transaction' (1932). Not only does the triple to which Commons referred—conflict, mutuality, order—prefigure the idea of governance, in that governance is a means by which to achieve order, thereby to mitigate conflict and realize mutual gains, but viewing the transaction as the unit of analysis turns out to be very instructive.

Llewellyn took exception with the legalistic way in which both lawyers and economists viewed contract. Rather than interpret contract in the prevailing legal rules tradition, Llewellyn advanced the more elastic concept of contract as framework—according to which a contract between two parties 'almost never accurately indicates real working relations, but...affords a rough indication around which such relations vary, an occasional guide in cases of doubt, and a norm of ultimate appeal when the relations cease in fact to work' (1931). Llewellyn's ideas prefigure the concepts of incomplete contracting (as against the complete contracting tradition in both law and economics) and of private ordering (as against court ordering).

Barnard brought his experience as a businessman and his instincts as a social scientist to the study of complex organization. Interestingly, both the economist Hayek and the organization theorist Barnard were in agreement that adaptation was the central problem of economic organization. But whereas the adaptations to which Hayek (1945) referred were autonomous adaptations in the market, the adaptations to which Barnard referred entailed 'the kind of cooperation among men that is conscious, deliberate, purposeful' (1938). These two work out of different mechanisms. Autonomous adaptations of a spontaneous kind are made in response to changes in relative prices signaled by the market; co-operative adaptations of an intentional kind are implemented through administrative processes within firms. Adaptations of both kinds being crucial to a high performance economy, there is a need to treat hierarchies and markets in a combined way.

This collection of deep insights—from economics, law, and organization theory—provided an interdisciplinary foundation for a science of organization. Aside, however, from work at Carnegie (by Simon (1947, 1957) and by Cyert and March (1963)), much of which was inspired by Barnard, such a program was slow to take shape. Instead, many of the prescient insights from the 1930s remained fallow for the better part of the next 35 years. The pressing but unmet needs were (a) to combine and (b) to operationalize these

profound insights. Law, economics, and organization theory needed to inform one another, thereafter to be joined in a predictive theory of economic organization.

2. Operationalization

Work of a combined kind got under way in the 1970s (Williamson 1971, 1975, 1979; Alchian and Demsetz 1972; Jensen and Meckling 1976; Klein et al. 1978) and has grown exponentially since. Six key moves in the progressive operationalization of the markets and hierarchies research program are sketched here.

2.1 Human Actors

According to Simon (1985), 'Nothing is more fundamental in setting our research agenda and informing our research methods than our view of the nature of the human beings whose behavior we are studying.' Both the cognitive and self-interestedness features of human actors are crucial to the economics of organization. Bounded rationality, according to which behavior is intendedly rational but only limitedly so, gradually has supplanted hyper-rationality as the basic cognitive condition. Opportunism, whereby economic actors are sometimes given to strategic behavior, has come into common use to describe self-interestedness. Bounds on rationality notwithstanding, economic actors are also presumed to possess foresight, which is crucial to the Commons triple. Taken together, these three attributes lead to the following: (a) all complex contracts are unavoidably incomplete (bounded rationality), (b) contract as mere promise unsupported by credible commitments poses hazards (opportunism), and (c) parties who are alert to prospective hazards will work out the ramifications and factor these into the organizational design (foresight).

2.2 Transactions

If the transaction is the basic unit of analysis, then the key dimensions with respect to which transactions differ need to be named and the ramifications worked out. Among the main attributes of transactions are: (a) the frequency with which transactions recur, (b) the uncertainty (disturbances to which transactions are subject), and (c) the condition of asset specificity. This last attribute has reference to the degree to which the assets that are used in support of a transaction can be redeployed to alternative uses and users without loss of productive value. Asset specificity takes several forms, of which physical asset specificity (specialized dies), human asset specificity (firm-specific learning), site specificity (locational), dedicated assets, brand name capital, and temporal specificity are leading types. A condition of bilateral dependency develops

when significant investments in transaction-specific assets are made. Intertemporal governance complications arise as a consequence.

2.3 Organization

Alternative modes of organization are described not in marginal but in discrete structural terms, which is to say that each mode of organization—market, hybrid, firm, bureau—is described as a syndrome of internally consistent attributes. Key dimensions for describing alternative modes of governance include (a) incentive intensity, (b) administrative controls, and (c) the contract law regime (Williamson 1991). (The market mode—with strong incentives, limited controls, and legalistic dispute settlement—favors autonomous adaptation; the firm mode—with weak incentives, extensive controls, and internal dispute resolution—favors co-operative adaptation.) The upshot is that alternative modes have distinctive strengths and weaknesses. They differ from one another in kind rather than in degree.

Whereas standard economic theory describes the firm as ‘a point or at any rate a black box,’ the economics of organization maintains that ‘firms are palpably not points. They have internal structure. This structure arises for a reason’ (Arrow 1999). The markets and hierarchies research paradigm describes the firm as a complex hierarchy in which both governance and specialization purposes are served. Cognitive and behavioral variance among the human actors in the population presents opportunities for specialization; and transactions for which co-operative adaptations are needed (namely, those for which assets are highly specific and need to be adapted to successive disturbances, given that contracts are incomplete) are ones for which added governance supports are needed. The concept of firm-as-governance structure—as a means by which to induce order, relieve conflict, and realize mutual gains for transactions that are in need of co-operative adaptations—thus comes to the fore.

2.4 Refutable Implications

Most of the refutable implications pertaining to governance are derived from the discriminating alignment hypothesis, according to which transactions, which differ in their attributes, are aligned with governance structures, which differ in their cost and competence, so as to effect a (mainly) transaction cost economizing result. Vertical integration in intermediate product markets is the canonical case, in relation to which a large number of variations on a theme have been worked out—for labor markets, capital markets, vertical market restrictions (such as franchising), regulation (and deregulation), economic reform (especially privatization), and public bureaus. More

generally, any problem that arises as or can be posed as a contracting problem can be examined to advantage in transaction cost economizing terms.

A second class of predictions arise in conjunction with specialization. Upon making provision for variance among human actors (in cognitive and self-interestedness respects), many opportunities to realize economies by matching the attributes of human agents to the needs of a task present themselves. This second class of predictions has been less thoroughly worked out but constitutes a promising research opportunity.

2.5 Empirical

Some theories of economic organization make little effort to advance refutable implications. Among those that do, few are empirically tested. Simon (1991) evidently believes that transaction cost economics is remiss in empirical respects: awaiting empirical testing, the new institutional economics and related approaches are acts of faith, or perhaps of piety.

Coase had registered similar concerns about the dearth of empirical work on contract and organization 20 years earlier, but that was before the operationalization of transaction cost economics had begun and predicted alignments were advanced. Empirical applications of transaction cost economics got under way in the 1980s and have grown exponentially since. It could have been otherwise, but the theory and evidence display remarkable congruity (Masten 1995). As Joskow (1991) puts it, ‘this empirical work is in much better shape than much of the empirical work in industrial organization generally.’

2.6 Full Formalization

The 1930s was the decade when the informal ideas for a new science of organization took shape, the 1970s was the decade when these ideas were worked up in a preformal way, the 1980s was the decade of semiformal analysis, and the 1990s witnessed a burgeoning of work of a fully formal kind. The paper by Grossman and Hart (1986) on *The Costs and Benefits of Ownership* was the first fully formal model of an incomplete contracting kind. Hart’s book on *Firms, Contracts, and Financial Structure* (1995) elaborates upon and develops these ideas. Additional work of this kind has been published since and more is in progress. As matters stand presently, however, there is a huge gap between the semiformal and fully-formal stages.

A leading reason for the break is that the fully formal analysis of incomplete contracts relies entirely on property rights assignments to assess comparative efficiency. Not only are the discrete structural differences in organization among alternative modes of governance denied, but the proposition that adaptation is the central problem of organization (of which

two types are distinguished) is ignored. These differences are consequential, whence Kreps (1999) concludes that 'If Markets and Hierarchies has been translated into game theory using notions of information economics, it is a very poor translation... In particular, mathematical-based theory still lacks the language needed to capture the essential ideas of bounded rationality, which are central to... transaction costs and contractual form.' Kreps (1999), nevertheless, remains sanguine that formal efforts will succeed eventually, possibly with the benefit of new mathematical apparatus.

3. Conclusions

Whereas once it was customary to talk about markets or hierarchies, where the former denoted decentralization and the latter central planning, that has changed. Now it is more customary to talk about markets and hierarchies. That entails much more than a replacement of the conjunction 'or' by 'and.' It represents a veritable sea change in the way economic organization is studied. The combined study of law, economics, and organization that is herein described (a) involves comparative institutional analysis in which economizing is the main case and the action resides in the attributes of human actors, transactions, and governance structures, (b) generates numerous refutable implications, (c) invites empirical analysis, (d) has many public policy ramifications, and (e) invites progressive formalization. Many economists are now persuaded that the study of economic organization needs to be studied in a combined way in which law, economics, and organization are joined. Legal scholars have incorporated many of these ideas into contract law, corporate governance, antitrust enforcement, and regulation. And organization theorists have come to appreciate the pervasive reach of economizing reasoning.

Real headway notwithstanding, new problems and challenges appear to be unending. Productive tension between transaction cost economics and each of the parts on which it stands—law, economics, and organization—can be projected safely into the future.

See also: Administration in Organizations; Llewellyn, Karl Nickerson (1893–1962); Organization: Overview; Organizational Control; Organizations: Authority and Power; Organizations, Sociology of; Transaction Costs and Property Rights

Bibliography

- Alchian A, Demsetz H 1972 Production, information costs, and economic organization. *American Economic Review* **62** (December): 777–95
- Arrow K 1999 Forward. In: Carroll G, Teece D (eds.) *Firms, Markets and Hierarchies*. Oxford University Press, New York, pp. vii–viii

- Barnard C 1938 (1982) *The Functions of the Executive*, 15th edn. Harvard University Press, Cambridge, MA
- Coase R 1937 The nature of the firm. *Economica* **4**: 386–405. (Reprinted. In: Williamson O E, Winter S (eds.) 1991 *The Nature of the Firm: Origins, Evolution, Development*. Oxford University Press, New York, pp. 18–33)
- Commons J R 1932 The problem of correlating law, economics, and ethics. *Wisconsin Law Review* **8**: 3–26
- Cyert R M, March J G 1963 *A Behavioral Theory of the Firm*. Prentice-Hall, Englewood Cliffs, NJ
- Grossman S J, Hart O D 1986 The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy* **94** (August): 691–719
- Hart O 1995 *Firms, Contracts, and Financial Structure*. Oxford University Press, New York
- Hayek F 1945 The use of knowledge in society. *American Economic Review* **35** (September): 519–30
- Jensen M, Meckling W 1976 Theory of the firm: Managerial behavior, agency costs, and capital structure. *Journal of Financial Economics* **3** (October): 305–60
- Joskow P 1991 The role of transaction cost economics in antitrust and public utility regulatory policies. *Journal of Law, Economics, and Organization* **7** (Special Issue): 53–83
- Klein B, Crawford R A, Alchian A 1978 Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics* **21** (October): 297–326
- Kreps D M 1999 Markets and hierarchies and (mathematical) economic theory. In: Carroll G, Teece D (eds.) *Firms, Markets, and Hierarchies*. Oxford University Press, New York, pp. 121–55
- Llewellyn K N 1931 What price contract? An essay in perspective. *Yale Law Journal* **40**: 704–51
- Masten S 1995 Introduction to Vol. II. In: Williamson O, Masten S (eds.) *Transaction Cost Economics*. Edward Elgar, Brookfield, VT
- Simon H 1947 *Administrative Behavior*. Macmillan, New York
- Simon H 1957 *Models of Man*. Wiley, New York
- Simon H 1985 Human nature in politics: The dialogue of psychology with political science. *American Political Science Review* **79**: 293–304
- Simon H A 1991 Organizations and markets. *Journal of Economic Perspectives* **5** (Spring): 25–44
- Smith A 1776 [1922] *The Wealth of Nations*. Dent, London
- Williamson O E 1971 The vertical integration of production: Market failure considerations. *American Economic Review* **61** (May): 112–23
- Williamson O E 1975 *Markets and Hierarchies: Analysis and Antitrust Implications*. Free Press, New York
- Williamson O E 1979 Transaction-cost economics: The governance of contractual relations. *Journal of Law and Economics* **22** (October): 233–61
- Williamson O E 1991 Comparative economic organization: The analysis of discrete structural alternatives. *Administrative Science Quarterly* **36** (June): 269–96

O. E. Williamson

High Performance Computing

Many scientific and engineering applications require a higher computing power than standard computers can provide. So in the last three decades of the twentieth century specialized computer architectures for high

performance computing (HPC) have been built. They are mainly optimized to execute floating-point operations at very high speed. The fastest HPC systems are also called supercomputers (the term supercomputing is sometimes used as a synonym for HPC). High performance is obtained by vector computing, parallel computing, use of special fast technologies or by a combination of these techniques. This article reviews the history of HPC systems, explains the most important high-performance computer architectures, and presents some typical HPC applications.

1. History of HPC

Figure 1 demonstrates that the performance of HPC systems based on the LINPACK benchmark (Dongarra 1990) was raised from about 0.1 GFlop s^{-1} (billions of floating-point operations per second) in the 1970s to more than $1000 \text{ GFlop s}^{-1}$ in the last decade of the twentieth century; so every decade the performance was increased by two orders of magnitude.

In the second half of the 1970s the introduction of *vector computers* marked the beginning of the HPC era. They offered a performance advantage of an order of magnitude compared to standard computer systems of that time, reducing the cycle time by introducing *pipelining* in the processor architecture and supporting high-level operations on complete vectors (e.g., component-wise addition or multiplication) by hardware. Typically, several functional units (FU), each dedicated to a particular arithmetic operation, are available in a vector processor (see Fig. 2(a)) and two or even more result vectors can be computed simultaneously. To achieve a high performance, the pipe-

lines of all FUs must be filled and emptied again at very high speed. Thus, input/output data and intermediate results are read from and written to fast vector registers each holding a fixed number (mostly 64 to 256) of vector elements. Separate load/store units perform an independent and overlapped data transfer from/to memory.

The first vector computer, Cray 1, was introduced in 1976. It is composed of six FUs and one load/store unit and is realized in fast emitter coupled logic (ECL) technology operating at 80 MHz. In later vector processor architectures, such as Cray 2 (1985), Cray Y-MP (1988), NEC SX-3 (1990), and NEC SX-4 (1994), performance was further increased by using up to 16 FUs and up to eight load/store units. Also the clock cycle time was decreased (e.g., by using liquid-cooled processors). During the 1980s, the number of vector processors installed in vector computers grew from one to eight and later to 16. Although they could communicate via a shared memory, it was difficult to use several processors on a single job, so mainly the throughput was increased by running several jobs in parallel.

As an alternative to vector computers, in the 1970s massively parallel processing (MPP) systems appeared as a new class of high-performance computers. Here a large number (typically ranging from 256 to 65,536) of processors, interconnected by a static network, operate in parallel to compute one task. The memory is distributed among the processors. In the first MPP systems the SIMD (single instruction multiple data) operation principle was used. The instructions are broadcast from a control unit (CU) to very simple processors (called processing elements or PEs) where they are executed simultaneously in lock-step mode on different local data items (see Fig. 2(b)). PEs are based on simple arithmetic units operating on words of limited length (often only of 1 bit). Operations on integers or floating-point numbers are performed by software routines that take several clock cycles depending on the length of the operands. Thus the power of such machines does not result from the PE itself but from the enormous number of installed PEs. So SIMD-type MPPs perform excellently on applications based on large matrices of short data elements, such as signal or image processing.

One of the first commercially available SIMD-type MPP systems was the distributed array processor (DAP) with up to 4096 PEs arranged in a 64×64 grid. The most famous machine of this class is the Connection Machine CM-1 (1985) and CM-2 (1987) with up to 64k simple one-bit processors connected by a hypercube network (Hillis 1989). For certain applications the CM-2 offered a higher performance than the CRAY Y-MP (see Fig. 1). MasPar MP-1 (with 4-bit PEs, 1990) and MP-2 (with 8-bit PEs, 1992) represent the last important machines of this architecture class, which disappeared from the market in the first half of the 1990s.

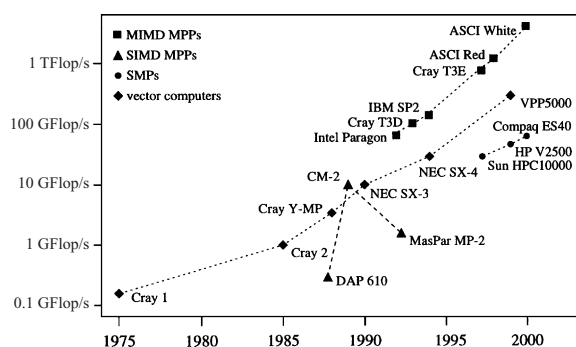


Figure 1

Performance of various high-performance computers from 1975 to 2000 for the solution of linear equation systems, measured in billions or trillions of floating-point operations per second (GFlop s^{-1} or TFlop s^{-1} , respectively)

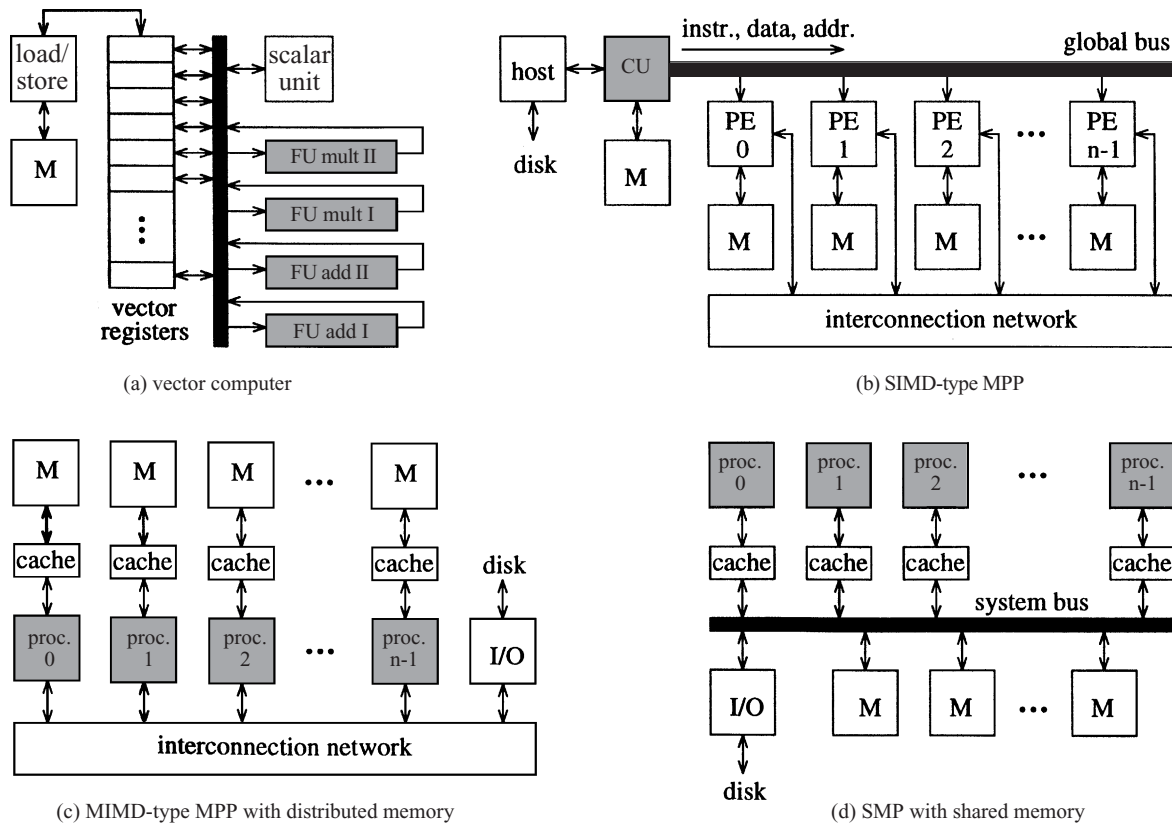


Figure 2
Architectures of high-performance computers

At the same time a new kind of HPC systems was proposed: MPPs based on advanced microprocessors operating according to the MIMD (multiple instruction multiple data) principle. In contrast to SIMD architectures the processors execute different tasks asynchronously. A few hundred or thousand cheap off-the-shelf processors designed for high clock-frequencies are interconnected by a network of high bandwidth. However, the communication is much slower than in SIMD-type MPP systems because there is a high overhead for initializing each communication and for synchronizing the processors.

In the first MIMD-type MPP systems, such as Intel's iPSC series with up to 128 processors or the nCube/2 system (1990) with up to 8,096 processors, the d -dimensional hypercube was the preferred topology. Here each processor can be reached in at most d steps; also two- and higher-dimensional grids can be embedded easily. Later systems were provided with additional communication controllers in each node and used new routing techniques so that the communication time became less dependent on the distance

between nodes. Topology was no longer important (if contention is ignored) and cheaper topologies such as a grid or a torus were favored. The Intel Paragon (1992) consists of up to 4,000 Intel i860 processors connected by a two-dimensional (2D) grid. The Cray T3D (1993) and T3E (1997) systems are based on a three-dimensional (3D) torus with up to 2,048 Alpha processors. In the second half of the 1990s, sometimes several workstations or PCs connected by an Ethernet or fiber Channel network were used instead of many integrated processors. Conceptually, this network of workstations (NOW) is not different from a distributed-memory MPP system, but the communication latency is higher.

The advantages of MIMD-type MPP systems are a good price/performance ratio and a good scalability of the architecture. On the one hand cheap and small systems with a few processors and a medium level of performance can be built easily; on the other hand the high performance of expensive vector computers can be reached or even exceeded at lower cost by using a sufficiently large number of processors. Due to the

large communication overhead, MIMD-type MPP systems are well suited for applications with a coarse-grain parallelism.

Since about 1995 several companies started to sell symmetric multiprocessor (SMP) models of their workstation families. In the first systems a rather small number (typically 4 to 16) of microprocessors access a shared memory built from several memory modules by using a central system bus (see Fig. 2(d)). Here each variable can be read and updated by each processor, which is supplied with a local cache to reduce the memory access time. The main problem in SMPs is to ensure that each variable remains consistent in all caches, especially when it is updated by one processor. This *cache coherency problem* is solved via a special protocol. Instead of a single bus often a set of several buses or even a crossbar network are used to reduce the risk of bus contention and to increase the collective bandwidth between memory and processors.

At the end of the twentieth century SMPs with a larger number of processors and a hierarchical interconnection structure became available. Here clusters of tightly-coupled processors are interconnected by a less costly and also slower network. Typical examples are Sun's HPC series (1997) and Hewlett-Packard's V-class series (1998). SMPs offer a high performance at low price. However, the scalability is limited: only a medium number of processors can be installed. So the performance levels of MPPs and vector computers cannot be reached (compare Fig. 1).

2. Programming HPC systems

On vector computers any existing sequential code can be used although it should be optimized manually to achieve high performance. For MPPs, SMPs and NOWs, however, the program must be parallelized based on either the data-parallel programming model or the message-passing model.

Data-parallel programs express all operations on vectors and matrices from a global point of view. The compiler automatically maps the data structures onto the parallel architecture. No manual synchronization or communication between the processors must be programmed, so the design of such data-parallel programs is rather simple. Examples of data-parallel languages are the several dialects of C offered by the manufactures of SIMD systems and High Performance Fortran (HPF).

In parallel programs based on the message-passing model the user must distribute tasks and data structures among the available processors. The user must also program each synchronization and communication explicitly by sending messages between the parallel tasks running on different processors. Each task can be implemented in a standard programming language such as C, C++ or Java. For the exchange of messages between tasks several standardized lib-

raries such as PVM (parallel virtual machine) or MPI (message passing interface) are available.

3. HPC Applications

In the beginning of the HPC era most applications involved the solution of large systems of linear or (partial) differential equations. Later other computation-intensive applications have emerged, such as the simulation of chemical and nuclear physical models, image processing, virtual reality, and optimization. Also artificial intelligence applications can be accelerated by parallelism—for instance, in 1997, IBM's MIMD-type MPP system Deep Blue consisting of 32 processors and 512 chess co-processors defeated world chess champion Kasparov. Deep Blue used a parallel alpha-beta tree search algorithm, tuned with expert knowledge, and could generate and evaluate 200 million chess moves per second (Hamilton and Garber 1997).

The simulation of artificial neural networks (ANN) represents another typical HPC application. Many ANN models (such as multi-layer perceptrons or radial-basis function networks) have been proposed that can be applied to many classification, approximation and optimization tasks without a formal treatment of the underlying problem. The training of ANNs by examples requires a high computational power, and many researchers have successfully simulated such ANN training on various parallel computer architectures (for an overview, see Šerbedžija (1996)). ANNs involve an inherent massive parallelism that can be exploited for a parallel implementation on different levels: either the *neurons* (basic ANN operating elements), the *synapses* (weighted connections between neurons), the *layers* (in case of a multi-layer network), or the *patterns* of the training set can be mapped onto different processors. However, especially at the first two levels, the communication/computation ratio is rather high because only simple operations are performed in each synapse or neuron between two communication steps. So, in particular, SIMD-type MPPs are well suited for a parallel implementation on all levels because of their good communication capabilities. High performance on MIMD-type MPPs, SMPs and NOWs can only be achieved by a pattern-parallel implementation; however, this technique can be applied to just a few neural learning algorithms. So for ANN simulation the paradoxical situation exists that older (and no longer available) SIMD-type systems are more suitable than modern HPC computer architectures.

The importance of high-performance computers is also steadily increasing in computational neuroscience. Complex biology-oriented neural models can be simulated on HPC systems in acceptable periods of time. Examples are: highly detailed single-cell models (such as a cerebellar Purkinje cell with 4550 compart-

ments (De Schutter and Bower 1994), medium-sized networks of neurons modeled at a less detailed level (such as hippocampal circuits with up to 128 pyramidal cells described by 70 compartments (Traub et al. 1996) or very large networks of neurons modeled at a higher abstraction level (such as the cat's primary visual system with 16384 integrate-and-fire neurons (Niebur and Brette 1994). In all cases discrete-time approximations to the solutions of a large set of nonlinear differential equations must be found.

Furthermore, HPC systems can accelerate several computation-intensive tasks in psychology such as simulating complex models of cognitive processes, simulating realistic models of visual perception, or analyzing electromagnetic activities in the brain. Especially the evaluation of psychological models and the estimation of model parameters yielding a best fit to data requires high computing power (for recent such applications, see Edwards (1997)).

4. Perspective

In the last three decades of the twentieth century, high performance computing was marked by a strong dynamic with a continuous appearing and disappearing of manufacturers, systems, and architectures. This rapid change will probably continue in the near future. Due to technological progress faster microprocessors will become available each year and cheap HPC systems with a few processors will offer a performance that could formerly be achieved only by expensive vector computers or MPP systems. Thus, high-performance computers will no longer be specialized architectures but be the high end of a continuum of fast desktop systems and servers. Scalable architectures can easily be adapted to the performance requirements of most computation-intensive HPC applications. However, for algorithms with high inter-processor communication demands (such as the simulation of ANN), it is not obvious if future HPC systems will be able to provide a sufficiently low communication latency to allow efficient parallel implementation.

A few special applications (such as weather prediction and computational fluid dynamics) require a still higher computing power than near-term HPC systems can provide. In 1996, the US Department of Energy's Accelerated Strategic Computer Initiative (ASCI) launched a program for a three-fold increase in computing performance every 18 months over a 10-year period using commercial hardware only. The first systems, ASCI Red (1997) and ASCI White (2000), consist of more than 7000 microprocessors and exceed 1 TFlops⁻¹ (see Fig. 1).

The scope of HPC applications will be further extended in the future. New applications will emerge, not only in science and engineering, but also in commerce (in areas such as information retrieval,

decision support, financial analysis, or data mining). *Data-intensive* algorithms are used here; they must operate on very large databases containing massive amounts of information (e.g., about customers or stock prices) that have been collected over many years. In particular, the rapid future increase in the amount of data stored worldwide will make HPC indispensable.

See also: Artificial and Natural Computation; Information Processing Architectures: Fundamental Issues; Information Technology; Science and Technology, Social Study of: Computers and Information Technology

Bibliography

- De Schutter E, Bower J M 1994 An active membrane model of the cerebellar Purkinje cell. *Journal of Neurophysiology* **17**: 375–419
- Dongarra J J 1990 The LINPACK benchmark: an explanation. In: Van der Steen A J (ed.) *Evaluating Supercomputers. Strategies for Exploiting, Evaluating and Benchmarking Computers with Advanced Architectures*. Chapman and Hall, London, pp. 1–21
- Edwards L (ed.) 1997 Special issue on high performance computer applications in the behavioral sciences. *Behaviour Research Methods, Instruments, and Computers* **29**: 11–124
- Fox G C, Williams R D, Messina P C 1994 *Parallel Computing Works*. Morgan Kaufmann, Los Altos, CA
- Hamilton S, Garber L 1997 Deep Blue's hardware–software synergy. *Computer* **30**: 29–35
- Hillis W D 1989 *The Connection Machine*. MIT Press, Cambridge, MA
- Leighton F T 1992 *Introduction to Parallel Algorithms and Architectures: Arrays, Trees, Hypercubes*. Morgan Kaufmann, San Mateo, CA
- Niebur E, Brette D 1994 Efficient simulation of biological neural networks on massively parallel supercomputers with hypercube architecture. *Advances in Neural Information Processing Systems* **6**: 904–10
- Šerbedžija N B 1996 Simulating artificial neural networks on parallel architectures. *Computer* **29**: 56–63
- Traub R D, Whittington M A, Stanford I M, Jefferys J G R 1996 A mechanism for generation of long-range synchronous fast oscillations in the cortex. *Nature* **383**: 621–4

A. Strey

High School Dropout

Dropping out of high school, that is, leaving school without graduating, is a central educational issue in almost all industrialized countries, although large differences between educational systems make cross-national or cross-cultural comparisons problematic. Nevertheless, there is broad international agreement that dropping out increases the risk of subsequent criminal behavior and drug abuse, lower occupational

and economic prospects, lower lifetime earnings, and becoming a member of the underclass. Furthermore, dropping out increases the probability of lower levels of mental and physical health. Studies investigating the consequences of high school dropout at a macro-level have shown that higher dropout rates result in lower tax revenues and increased expenditure for governmental assistance programs. Against the backdrop of all these negative consequences, sociological, educational, and psychological researchers have investigated the reasons for dropping out. The reasons identified can be classified as school-related, family-based, or personal. Researchers agree that the risk of dropping out begins to develop early in a student's life and that intervention programs to prevent dropout should therefore be initiated at an early age.

1. Problems in Calculating the Dropout Rate

A major challenge facing any study of high school dropout is how to define and identify the dropout population. There is some agreement, among American researchers at least, that students who were enrolled in grades 10, 11, or 12, but left school without graduating are to be defined as dropouts. However, this definition is often not viable in other countries. In international comparison, school systems not only differ with respect to the organization of primary and secondary schools, forms of ability grouping, and curricula, they also differ in terms of the age at which students are allowed to leave school. While students in the UK and most states of the USA are legally obliged to attend school until the age of 16, for example, students in other countries (e.g., Germany) are not allowed to leave school until they are 18. As a consequence of the different lengths of compulsory schooling, the probability that a student will drop out between the ages of 16 and 18 varies from country to country. Furthermore, there is some evidence that high schools tend to overestimate the percentage of dropout (e.g., Rumberger et al. 1990), and it has been argued that 'dropout data are probably the least reliable information available today regarding the reality of schools' (Frymier 1996, p. 4).

Beyond these problems of definition, it is characteristic of almost all industrialized countries that school participation rates increased continuously during the twentieth century. In the USA, for example, the proportion of 14- to 17-year-olds enrolled in high school increased from 11 percent in 1900 to 94 percent in 1978. However, dropout rates have been relatively stable since the beginning of the 1970s (between 5 and 10 percent in industrialized countries).

2. Factors Influencing High School Dropout

Three major directions of research into the antecedents of high school dropout can be distinguished, although

there is substantial theoretical overlap between the three approaches. While the first two types of research concentrate on family background and school variables from the sociological perspective, the third—more psychological—approach focuses on individual characteristics in terms of achievement and motivation.

2.1 Family Background and High School Dropout

Where family background is concerned, the determinants found to be of importance for academic achievement have also been considered with respect to dropout (e.g., Rumberger 1983, Rumberger and Thomas 2000). Parents' level of education represents a potential predictor of dropout, that is, better educated parents spend more time with their children, thus supporting their accomplishments and increasing the likelihood that they will remain in school. Furthermore, these parents influence the academic performance of their children by imparting the values, aspirations, and motivation needed to succeed and remain in school. With regard to parental income, children from poor families probably feel more pressure to contribute to the family income, and thus leave school to seek work at an earlier stage. Family structures are also important, in that children from broken homes where one or both parents are absent are less likely to receive the support and encouragement needed to stay in school. The absence of parents, however, is often combined with financial problems that might explain the dropout. In a more detailed analysis of family influences, Rumberger et al. (1990) investigated the interaction styles between parents and children and the consequences for dropout. Dropouts were more likely to live in households characterized by a permissive or an authoritarian parenting style, their parents were more likely to use extrinsic punishments as a reaction to poor grades, and reacted with more negative emotions to both good and bad school grades. Furthermore, dropout students reported that their parents were less involved in their education and gave them less help with their homework.

In the USA, in particular, relations between ethnic group membership and high school dropout have been investigated, and higher percentages of dropout have been reported for Hispanic and African-American students. In a study investigating primary reasons for dropout, Rumberger (1983) found out that African-American females most often cited marriage and pregnancy as reasons for leaving school. However, if SES and individual achievement are held constant, the relationship between ethnic background and dropout diminishes or reverses, thus suggesting that the association is mainly mediated by family background and individual levels of achievement (e.g., McNeal 1995).

2.2 School Characteristics and High School Dropout

School characteristics have been investigated in research on school effectiveness. Authors agree that not only academic achievement, but dropping out of school is a reliable indicator of school effectiveness. Rumberger and Thomas (2000) identify school resources, structural characteristics of schools, and school processes as factors, which can influence high school dropout. With regard to the financial resources of schools, research shows that the student-teacher ratio is positively related to high school dropout. Furthermore, there is some evidence that a higher percentage of teachers with advanced degrees increase the likelihood that students will remain in high school.

Structural characteristics such as school location (urban, suburban, rural), size, and type (state vs. private) also seem to influence dropout, even when important individual and family characteristics are controlled. Research by Pittman and Haughtwout (1987) revealed a positive correlation between high school size and dropout rate. Mediation analyses demonstrated that this effect was mediated primarily by the social climate (sense of cohesion, level of student participation in school activities, interaction with faculty, the magnitude of certain problems at the school), and indicated that smaller schools had a better social climate and consequently lower dropout rates. Private schools usually have lower dropout rates than state schools (e.g., Rumberger and Thomas 2000), and students attending urban schools are more likely to drop out than their counterparts in suburban schools (e.g., Frymier 1996).

Research on school processes has identified a number of school-related variables that influence dropout rates. A higher level of cooperation between teachers and students seems to reduce dropout rates and, with regard to the academic climate, Bryk and Thum (1989) found that dropout rates were lower in schools in which students took more advanced courses and reported doing more homework. Bryk and Thum also reported that high schools with serious staffing problems had higher dropout rates.

Rumberger and Thomas (2000) suggested that because schools are anxious to show that their students' test scores are improving, they actively reduce the membership of low-achieving students through a process of discharge. This process affects not only low-achieving students, but also hard-to-educate students. Since both variables are confounded with SES and ethnicity, these school processes usually lead to higher dropout rates for students with low SES, and for Hispanic and African-American students.

2.3 Individual Determinants of High School Dropout

Although many background and school variables are important antecedents of dropout, a more psycho-

logical view of the problem suggests that the concrete decision of whether or not to carry on or leave school is guided by individual characteristics such as attitudes and motivational states. Schools, parents, and teachers provide an academic environment, and students' perceptions of this environment determine their achievement motivation, their behavioral intentions, and their academic behavior and choices. In one of the few psychological studies on reasons for dropout, Vallerand et al. (1997) proposed a model assuming self-determined academic motivation to be an important factor influencing behavioral intentions and dropout behavior. Their study suggests that students who perceive their parents, teachers, and school administrations as supportive develop a higher level of perceived competence and academic autonomy, and that this leads to higher intrinsic academic motivation and a lower likelihood of leaving school without graduating.

Expectancy-value models (e.g., Wigfield and Eccles 2000) suggest that academic choices are based mainly on expectancies of success in the domain of choice, and on how students value the domain. Both expectancies and values are based on family background, school and class variables, and previous individual achievement-related experiences. Low expectancies and values with respect to the future school career increase the probability of dropping out.

3. Prevention of High School Dropout

Low levels of academic achievement (e.g., Caliste 1984) and high absenteeism (e.g., Bryk and Thum 1989) are the strongest behavioral predictors of high school dropout. Several authors (e.g., Simner and Barnes 1991) have pointed out that the earlier a student with low achievement levels and high absenteeism can be identified, the more likely is it that dropout prevention will be successful. An effective 12-week prevention program proposed by Caliste (1984) first identifies at-risk students (grade-point averages below 1.0 on a 4.0 scale). These students receive extra tuition from their teachers and participate in group counseling sessions. Topics discussed during the weekly counseling sessions include motivation, relevance of school, academic problems, career goals, and study habits. In the tutoring sessions, students are primarily assisted with current coursework and given some remedial instruction.

Participation in out-of-school activities, particularly sports clubs, can also prevent dropout. McNeal (1995) argues that higher social integration in both school and in non-academic clubs influences whether or not students remain in school. His analyses show that for low-achieving students (one standard deviation below the population mean) not involved in sports clubs, the probability of dropout is $p = 0.25$, while for members of sports clubs the probability is only $p = 0.17$.

Finally, authors agree that children from disadvantaged family backgrounds need particular social and financial support. Females, especially African-Americans, often leave school without graduating because of marriage and pregnancy. Day-care programs are viable measures to prevent drop out here.

See also: Adolescent Behavior: Demographic; Adolescent Development, Theories of; Adolescent Vulnerability and Psychological Interventions; Career Development, Psychology of; Childhood and Adolescence: Developmental Assets; Families as Educational Settings; Family Health; Gangs, Sociology of; Internet: Psychological Perspectives; Media Effects; Media Effects on Children; Motivation, Learning, and Instruction; Popular Culture; School Achievement: Cognitive and Motivational Determinants; School as a Social System; School Learning for Transfer; School Outcomes: Cognitive Function, Achievements, Social Skills, and Values; Schooling: Impact on Cognitive and Motivational Development; Urban Poverty in Neighborhoods; Violence and Media; Youth Culture, Sociology of; Youth Gangs

Bibliography

- Bryk A S, Thum Y M 1989 The effects of high school organization on dropping out: An exploratory investigation. *American Educational Research Journal* **26**: 353–83
- Caliste E R 1984 The effect of a twelve week dropout intervention program. *Adolescence* **19**: 649–57
- Frymier J 1996 Determining dropout rates in large city school districts: Problems and accomplishments. *The High School Journal* **80**(1): 1–10
- McNeal R B Jr 1995 Extracurricular activities and high school dropouts. *Sociology of Education* **68**: 62–81
- Pittman B P, Haightwout P 1987 Influence of high school size on dropout rate. *Educational Evaluation and Policy Analysis* **9**: 337–43
- Rumberger R W 1983 Dropping out of high school: The influence of race, sex, and family background. *American Educational Research Journal* **20**: 199–220
- Rumberger R W, Ghatak R, Poulos G, Ritter P L, Dornbusch S M 1990 Family influences on dropout behavior in one California high school. *Sociology of Education* **63**: 283–99
- Rumberger R W, Thomas S L 2000 The distribution of dropout and turnover rates among urban and suburban high schools. *Sociology of Education* **73**: 39–67
- Simner M L, Barnes M J 1991 Relationships between first-grade marks and the high school dropout problem. *The Journal of School Psychology* **29**: 331–5
- Vallerand R J, Fortier M S, Guay F 1997 Self-determination and persistence in a real-life setting: Toward a motivational model of high school dropout. *Journal of Personality and Social Psychology* **72**: 1161–76
- Wigfield A, Eccles J S 2000 Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology* **25**: 68–81

O. Köller

Higher Education

In most countries, teachers of the most prestigious institutions of higher education are assigned both teaching and research tasks, whereby 'academic freedom' is held in high esteem. A trend towards systematization notwithstanding, curricular approaches have remained varied across countries, fields and institutions, regarding the academic or professional emphasis, their adaptive or critical role to the world of work, the degree of specialization and the role higher education plays for personality development beyond the cognitive domain. Recent reforms react to trends of internationalization and globalization, increase the role of practical experiences, life-long education, distance education and use of new technologies.

Universities are often called surprisingly persistent institutions, though their organization is often weak and anarchic. The continental European tradition of a weak university leadership amid a major role of government and academic freedom of scholars gave way in the 1980s to reduced governmental bureaucratic control and increased formal autonomy of the central management of higher education institutions amid growing pressures for efficiency, accountability, and social relevance.

1. Definition

'Higher education' became a popular term in the second half of the twentieth century depicting the intellectually most demanding stage of pre-career education. Terms such as 'post-secondary,' 'tertiary,' or 'third-level' education underscore stages of learning: upon completion of primary and secondary education, i.e., after 10–14 years of schooling, varying according to national systems, students might enroll in the third stage of education, as a rule prior to embarking regular employment. The term 'higher,' however, suggests a specific quality, expecting students to learn questioning prevailing rules and tools and to understand theories, methods, and substance of 'academic' knowledge. Higher education also often serves a twofold function: teaching and 'research,' i.e. the creation and preservation of systematic knowledge.

2. History

Institutions and programs of advanced learning existed in various countries more than 1,000 years ago. However, the core elements of today's higher education, i.e., teaching and learning of 'analytic,' 'rational,' 'systematic,' 'critical,' 'skeptical,' and 'innovative' thinking, emerged from the European universities of the Middle Ages.

Around 1800, a new historical stage can be noted when substantial changes of the European universities occurred. The British model of the Oxford and Cambridge approach regained strength which puts emphasis on the trained mind and the personality development through close communication between the teachers and the students. The French university, reformed under the Napoleonic regime, combined strong state coordination of organization and curricula leading to state-coordinated degrees with a decentralized organization within the university, thus reinforcing a divide between the faculties and notably intellectual and cultural approach of the *facultés des lettres* and sciences on the one hand and on the other the professional emphasis of other faculties inside the universities and *écoles* outside the universities. The German model, mainly the 'unity of teaching and research' approach of the Humboldtian university, was based on the idea that academic freedom and institutional autonomy guaranteed by the state help to secure the university's most appropriate service to society, and also underscored the freedom of learning, i.e., the student's ability and responsibility of independent learning in communication with the professor. The German model became the internationally most influential in the nineteenth century.

The models spread through adaptations and misunderstandings which turned out to be creative in some instances. Notably, the US approach of being 'as British as possible' as far as undergraduate education is concerned and 'as German as possible' by combining teaching and research, and in establishing graduate schools as a completely new element, tends to be viewed as the most successful one in the twentieth century (Ben-David 1977).

The development of higher education after World War II is often regarded as the third stage of the history of modern higher education. Higher education was no longer expected to serve a small number of academics and socially exclusive professions, but rather became a 'mass' or an almost 'universal' phenomenon (see Trow 1974). Whereas less than 5 percent of the respective age group enrolled around 1950 in most industrial countries, the respective proportion was even more than 50 percent in some countries at the end of the twentieth century. The number of doctoral-granting institutions in the world had risen to about 10,000 and of other institutions to even a higher number, and the total number of students was more than 80 million in 1995.

3. Quantitative and Structural Developments

3.1 Access and Admission

According to OECD statistics (OECD 2000), about 80 percent of the corresponding age group on average of the OECD member states successfully completed

upper secondary education in the late 1990s. Upper secondary education varied between countries structurally, i.e., 'horizontal' (graduates of all institutions and programs are eligible for entry to higher education) or 'vertical' (specific types and programs preparing for higher education), as well as according to the role vocational training plays and the way it is organized.

Modes of admission to higher education can be classified according to three types (see Teichler 1988).

(a) If secondary education is organized horizontally, the large numbers of those successfully completing secondary education are eligible to apply for admission to higher education, whereby entry to higher education is the major point of selection within the education system. The entry decisions rest with the higher education institutions (in the USA) or their departments (e.g., in Japan), and might be based on school grades, combinations of school subjects, admission tests, interviews, etc.

(b) If secondary education is vertically organized, those successfully completing the academic track (being awarded the *Abitur*, *Matura*, *Baccalaureat*, etc.) tend to be entitled to enroll at institutions of higher education (e.g. Germany and The Netherlands). As a rule, entry is a nondramatic point of allocation, and selective admission applies only for few fields and institutions.

(c) Finally, some countries provide for both selective secondary education and selective admission at entry to higher education (e.g. the UK).

At all modes, most of those applying for higher education will eventually enroll. The admission system in industrialized countries is not primarily a gate-keeper of the overall quantity of enrolment, but rather allocates primarily with respect to type, institutional prestige or field of study.

3.2 Enrolment, Drop-out, and Graduation

The 'youth' in most countries acquire entry qualification for higher education at the age of 17 to 19 years. The average age of new entrant students varies in industrial countries between 18 and 23 years.

The OECD reports an average 'net entry rate' to higher education in terms of programs leading to a (at least bachelor) degree of 40 percent in industrialized countries for 1998. In addition, almost 20 percent begin short tertiary-education studies. Altogether, women have enrolled as frequently as men, on average, in industrialized societies since about 1990. UNESCO calculated a gross enrolment rate, by comparing enrolment figures of young people within five years of the normal school-leaving age: 63 percent enrolled in 'more developed regions,' 38 percent enrolled in countries in transition and 7 percent enrolled in 'less developed regions' in the mid-1990s.

About two-thirds of those enrolling in degree programs in industrialized countries are eventually

awarded a degree. The estimated dropout rates are 10 percent in Japan (lowest) and about 65 percent in Italy (highest).

In 1998, 22 percent of the respective age group on average across industrial countries eventually was awarded a degree (in comparison, 1 percent on average were awarded a doctoral degree). In addition, 11 percent completed other tertiary education programs.

3.3 Diversity of Levels, Programs, and Institutions

Universities, i.e., multidisciplinary institutions in charge of both teaching and research, entitled to award advanced academic degrees (notably the doctorate) and, where applicable, entitled to award subsequent degrees qualifying for senior academic positions (the 'doctor scientiae' or the 'Habilitation'), are considered the key institutions of higher education. Except for France, where *Grandes Écoles* lead to the highest professional careers, entry to university is viewed as most prestigious. Most other institutions, i.e., mono-disciplinary institutions of technology, theology, etc., initially viewed as less prestigious, often gained the same status over the years.

Since the 1960s, two developments reinforced a diversification of higher education. Pressure towards increasing systematic knowledge in the middle-level occupations led to an 'up-grading' of the institutions serving the training of these occupations. Concurrently, the university sector came under pressure to diversify because the growing student body became increasingly heterogeneous in terms of motivations, competencies and career prospects and because resources were not made available for an expansion of research-oriented higher education in tune with increasing students' enrolment (Teichler 1988, Meek et al. 1996).

In various countries, an additional sector of higher education was established, called for example 'polytechnics' in the UK (which eventually became universities in 1992), 'instituts universitaires de technologie' in France and 'Fachhochschulen' in Germany which initially were expected to provide applied educational programs, but not to undertake research. This sector was named 'short cycle,' 'non-university' or 'alternative' without any of these terms being generally appreciated.

Countries varied substantially, however, according to the extent of which differentiation of higher education was realized through:

- (a) types of higher education institutions;
- (b) curricular approaches of programs (e.g. academic versus professional-vocational);
- (c) levels of programs and degrees; (e.g. the OECD's 'non-university certification,' 'short first university degree,' 'long first university degree,' 'second university degree,' and doctoral or similar degrees);
- (d) length of programs (in terms of years of study or credits-points, etc.); and

(e) varied reputation and prestige among formally equal institutions and programs.

Higher education systems, for example, were characterized as 'unitary,' 'binary,' 'dual' or 'multi-type,' 'stage' or 'level,' 'comprehensive,' etc.

4. Knowledge, Teaching, and Learning

4.1 Learning and Pursuit of Knowledge

Students are expected to learn questioning the available wisdom and the established ways for problem solving and to be prepared for indeterminate work and other life tasks. A close link between teaching and research should ensure that teaching is at the 'cutting edge' of research development. In some countries, curricula are very open and teaching is conceived more or less as a by-product of research.

4.2 Trend Towards Systematization of Curricula

With the expansion of higher education, however, a trend could be observed towards systematization of content, structures, and processes. Study programs were divided into sequences and stages. Curricular regulations grew, and measures of informing and guiding students were extended. In many countries, teachers were expected to be trained in techniques of teaching, and regular evaluation of teaching practices often was introduced. The trend towards systematization affected initial course programs most strongly. But the US concept of graduate school for those preparing for a doctoral degree was also taken up or modified in most other countries in the latter half of the twentieth century (Clark 1995).

Yet major curricular approaches remained varied across and within countries, notably according to seven dimensions. First, curriculae may be directed more strongly towards preparation for research and the creation of knowledge or towards the reproduction of knowledge available. Second, curriculae may be geared closely to occupational preparation or not related to job roles at all. Third, objectives of higher education curricula differ according to the degree of specialization. Fourth, fields of study may be designed according to disciplines or take on a multidisciplinary character. In some countries (fifth), a common core of knowledge might be offered for all students. Sixth, views differ as regards the responsibility of higher education for the personality development of students beyond the cognitive domain. Finally, curricula vary according to the extent they aim to contribute to cultural enrichment of society.

The variety of curricular approaches had their pendant in the roles of the teachers and the students. Students might be viewed as young learners for which the university serves 'in loco parentis,' as customers or

as independent and self-responsible young members of the university community. Teachers might view themselves as scholars in the pursuit of knowledge, or as educators and counselors. Available research suggests that students have more similar views across countries on 'good' teaching and learning, and teachers on their role and function (Altbach 1997) than one would expect, taking into account the diversity of national systems of higher education.

4.3 *Expectation of Relevance and Utility Amid Uncertainty*

In the 1980s and 1990s, the conditions both for teaching and learning, as well as for research in higher education, changed substantially. First, pressure grew on teaching, learning and research to be more utilitarian than in the past. Public and private funds were moved increasingly towards research and establishments outside higher education, thus reducing the share of research expenditures spent in higher education in many countries to less than 20 percent. Thus, modes of research spread aiming to cut across disciplines, in spite of the persistence of distinct cultures of the disciplines (Becher 1989), and became more context aware and problem oriented (Gibbons et al. 1995).

Efforts to respond to the societal needs were often met by reinforcing a divide between basic and applied research or by a divide between academic and professional or vocational study programs, but these distinctions became blurred over time. While utilitarian pressures turn out to be imaginative for basic research in some cases, research aiming not to be applied often lay the foundation for practical use. Study programs not geared to certain professions are enriched by a confrontation with practical problem solving, both in classes and internships, while study aiming to be professionally useful is forced to become broader, owing to the uncertainties of future labor market conditions and work assignments—'key qualifications' is a widely used term underpinning this trend.

4.4 *Internationalization and Globalization*

Higher education could be viewed as very international in terms of cosmopolitan values, pride based on international reputation and global search for knowledge. However, the regulatory and funding context tends to be national, and curricula vary strikingly between countries. Kerr (1990) described a conflict between 'the internationalization of learning and the nationalization of the purposes of higher education.' International agencies, e.g. UNESCO, OECD or the World Bank often advocate concepts of convergence, but seem to be counterbalanced by a persistence of varied traditions and political options.

5. *Steering, Governance, and Organization*

5.1 *The Deliberately Weak and 'Fuzzy' Organization*

Universities are often called surprisingly persistent organizations and have survived for several centuries, though they seem to be weak or anarchic institutions in most countries. The links between the higher education institutions and the government or the society can be described as a peculiar mix of partial control and vague pressures, on the one hand, and safeguards of institutional 'autonomy' on the other hand. In the internal organization, accommodation prevailed in most countries between administrative supervision, a strong position of the professorate as a whole and the 'academic freedom' of the individual scholar.

The weakness was widely viewed as supportive for an ethos within the institution which allowed scholars to focus on the pursuit of knowledge without major concern about the social and organizational environment. The Humboldtian ideal was widely shared that a university protected from pressures of immediate utility was most likely to serve generation of knowledge and innovation eventually valuable for society.

5.2 *Traditional Modes of Governance and Administration*

Analyses of higher education governance often intend to describe both the relationships between higher education and society and the modes of governance within the institutions, thus underscoring that the internal mechanisms of power, cooperation, and communication are linked to the ways how higher education is steered from outside. Notably, US scholars have described traditional higher education systems according to four models:

(a) the collegial model—emphasizing nonhierarchical cooperative decision making, and a significant degree of self-determination by academic staff;

(b) the bureaucratic model—emphasizing legal-rational authority and formal hierarchies;

(c) the professional model—emphasizing the authority of experts and the importance of horizontally differentiated units linked to loose confederations;

(d) the political model—conceptualizing governance in terms of political conflict among interest groups with competing views and values' (Harman 1992, p. 1282).

Often, three actual modes of academic governance and administration are pointed out (see Ben-David 1977, Clark 1983). Both in the French and in the German traditions, governments, though largely respecting or protecting academic freedom, pursued certain visible ways of bureaucratic control of higher education institutions. The university leadership was

kept deliberately weak whereby the faculties in France and the chairs in Germany were the stronghold of academic freedom and institutional autonomy. In the British tradition, funds were allocated to the higher education institutions with little strings, the strongest power of the institution rested in academic collegiality, and the university leadership, though stronger than in continental Europe, exerted moderate control. The US model was characterized by a weak position of visible macrosocial forces, whereby the market seemed to play a stronger role than government and the trustees of the individual institutions of higher education, and by a strong power of the president and the deans along limited decision-making power of the academics.

Clark (1983) defined the patterns of control and coordination according to a triangle of market, state authority and academic oligarchy. He allocated US higher education as relatively close to the market, Germany to state authority, and Italy to academic oligarchy.

5.3 In Search of New Modes of Governance and Administration

Since the 1970, the climate around and within institutions of higher education has changed substantially. Obviously, institutions of higher education have lost much of their special character. They are expected to be more similar to other institutions, and the functions they perform are less exclusive. Higher education experienced:

(a) a loss of trust in the academic profession and in academic freedom. Counterbalances by other forces, scrutiny of job performance and of academic quality as well as direct incentives for social relevance and utility are viewed now as appropriate;

(b) a loss of confidence in governments' potentials of steering higher education successfully through detailed regulations and control of administrative processes; and

(c) a growing demand for efficiency in terms of serving extended duties of teaching and research with stagnating resources and for visible effectiveness.

Four major changes of governance and administration of higher education occurred. First, in the wake of the student protests in the late 1960s, various continental countries introduced models of participation whereby students, junior academic staff, and administrative staff got a substantial say, as was the case previously in various Latin American countries. This reform approach, however, lost momentum rapidly because the process of decision making seemed to become controversial and time consuming and because the actors benefiting often were disappointed as regards the limited real influence the formal decision-making process actually allotted to them.

Second, various governments in Europe moved from a 'control' approach, and in the UK from an

'institutional autonomy' approach, towards a 'supervision' approach (Neave and van Vught 1991). Mechanisms employed were, among others, political target setting, steering through indicator-based funding, contracts between the state and higher education institutions, and the establishment of evaluation systems allowing to identify the 'output' of higher education (Westerheijden et al. 1994).

Third, managerial actors (university leadership, deans) were strengthened and managerial mechanisms were established or extended. The internal processes of decision making and administration were expected to get close to industrial entrepreneurship, while respecting the specifics of a professional organization serving unpredictable innovation of knowledge.

Fourth, the setting of key actors increasingly grew complex. In addition to the state, the market and the professorate (Clark 1983), the management of the higher education institutions, and other internal actors (students, junior academic staff, and administrative staff) came into play as well as other 'stakeholders' (i.e., all other forces who could have a legitimate interest to be involved in higher education as a 'visible hand').

See also: Education and Income Distribution; Education and Learning: Lifespan Perspectives; Education (Higher) and Gender; Educational Institutions and Society; Educational Institutions, History of; Human Capital: Educational Aspects; Policy Knowledge: Universities; Universities and Science and Technology: Europe; Universities and Science and Technology: United States; Universities, in the History of the Social Sciences

Bibliography

- Altbach P G (ed.) 1997 *The International Academic Profession*. Carnegie Foundation for the Advancement of Teaching, Princeton, NJ
- Becher T 1989 *Academic Tribes and Territories*. SRHE/Open University Press, Buckingham, UK
- Ben-David J 1977 *Centers of Learning: Britain, France, Germany, United States*. McGraw-Hill, New York
- Clark B R 1983 *The Higher Education System: Academic Organization in Cross-National Perspective*. University of California, Berkeley, CA
- Clark B R 1995 *Places of Enquiry: Research and Advanced Education in Modern Universities*. University of California Press, Berkeley, CA
- Gibbons M, Limoges C, Nowotny H, Schwartzman S, Scott P, Trow M 1995 *The New Production of Knowledge*. Sage Publications, London
- Harman G 1992 Governance, administration and finance: Introduction. In: Clark B R, Neave G R (eds.) *The Encyclopedia of Higher Education*. Pergamon Press, Oxford, UK, pp. 1279-93
- Kerr C 1990 The internationalization of learning and the nationalization of the purposes of higher education: Two 'laws of motion' in conflict? *European Journal of Education* 25: 5-22
- Meek V L, Goedegebure L, Kivinen O, Rinne R (eds.) 1996 *The Mockers and Mocked: Comparative Perspectives on Differ-*

- entiation, *Convergence and Diversity in Higher Education*. IAU Press/Pergamon, Oxford, UK
- Neave G, van Vught F (eds.) 1991 *Prometheus Bound: The Changing Relationships Between Government and Higher Education in Europe*. Pergamon Press, Oxford, UK
- OECD 2000 *Education at a Glance*. OECD, Paris
- Teichler U 1988 *Changing Patterns of the Higher Education System*. J. Kingsley Publishers, London
- Trow M 1974 Problems in the transition from elite to mass higher education. In: OECD (ed.) *Policies for Higher Education*. OECD, Paris, pp. 51–101
- Westerheijden D F, Brennan J, Maassen P A M (eds.) 1994 *Changing Contexts of Quality Assessment*. Uitgeverij Lemma, Utrecht, The Netherlands

U. Teichler

Hindu Law

Hindu law claims one of the longest continuous histories of any legal system in the world. Indeed, for about 2,500 years it was based on the same primary sources. Yet, the attitude toward these sources has varied considerably over the centuries, with the result that Hindu law in 2,000 CE, is very different from what it may have been in 500 BCE.

1. The *Smṛtis*

Until the 1950s the main sources of Hindu law were contained in texts written in India's classical language, Sanskrit (Derrett 1973a, 1973b). The earliest of these texts, the *dharmaśāstras* (mostly in prose, and probably earlier) and *dharmasūtras* (exclusively in verse, and probably more recent) may be dated to approximately between 500 BCE and 500 CE. The Indian concept of *dharma*, for which there is no Western equivalent, is much broader than the Western concept of 'law.' A Hindu's *dharma* encompasses each and every rule people are supposed to live by and, even though the legal aspects of *dharma* are more and more treated in isolation in the *dharmaśāstras*, 'legal' prescriptions in these texts remain inseparable from other obligations, especially those which we would label 'religious.' A case in point is adoption, which is not valid unless accompanied by a ritual act, the *dattahoma*.

To understand the Hindu view of *dharma*, and of law in particular, it is essential to keep in mind that the *dharmaśāstras* and *dharmasūtras* are part of the revealed Veda. They make up the *smṛti* 'remembrance,' the authority of which is only slightly lower than that of the older Vedic texts, the *śruti* 'revelation.' They too, were considered to have been transmitted to mankind by *rishis* ('seers'). Like the Vedas themselves, the texts on *dharma* are perfect, eternal, and un-

changeable; they should not be interfered with by human beings.

The most important *smṛti* text is the *Manusmṛti* or *Mānavadharmasāstra* (ca. 200 BCE–200 CE), attributed to Manu, the progenitor of all Hindus (Bühler 1886). Chapters 8 and 9 of the *Manusmṛti* are entirely devoted to legal matters. Together with Chapter 7 they constitute the *rājadharmā*; the king is indeed the supreme justice in his kingdom. For the first time, an effort is made here to classify the causes of legal disputes, under 18 titles: failure to pay debts, deposit and pledge, sale without ownership, partnerships, resumption of gifts, failure to pay wages, breach of contract, rescission of sale or purchase, disputes between the owner and keeper of cattle, boundary disputes, physical assault, verbal assault, theft, violent crimes, sex crimes, rules for husband and wife, partition and inheritance, and gambling and betting (8.4–7). The subdivision into 18 'heads of litigation' is maintained in all later *dharmaśāstras*, with the difference that their eighteenth chapter is often called *prakīrnaka* 'miscellaneous,' to incorporate aspects of substantive law that were not dealt with by Manu. Only one *dharmaśāstra* is entirely devoted to law, the *Nāradaśmṛti* (Lariviere 1989). The *Manusmṛti* deals with procedure under the first head of litigation; Nārada devotes three introductory chapters to it. Most notable, in the chapters on evidence, is the highly detailed treatment of witnesses and ordeals.

In view of the later development of Hindu law, reference must be made here to one characteristic trait of *smṛtis* that has provoked much discussion in the scholarly literature. There are, in these texts, obvious contradictions, not only between different texts, but also within single texts such as Manu's translation of Lingat's 1967 work (see Lingat 1967). In one verse (9.104) Manu prescribes that, after the death of both parents, their sons divide the paternal estate equally. Without any transition, the following verse (9.105) says that the eldest brother takes the entire paternal property, while the others shall live under his authority as they did under their father's. After a few verses in praise of the eldest son, Manu continues: 'The eldest son takes an additional share of five percent, as well as the most desirable part of the estate; the next son gets half of that, the youngest one quarter' (9.112). Or, Manu acknowledges and describes in detail the duties of partners in a levirate marriage (9.60), and elsewhere rejects levirate marriages as *paśu-dharma* 'the *dharma* of cattle' (9.66). Scholars have attempted to expurgate the preserved Manu text so as to avoid the contradictions. In reality, Manu's three ways of inheriting paternal property, and his opposite views on levirate, may result from the fact that they were all *dharma* in different places, among different groups of people, etc. They all knew exactly which rule applied in their community, but all these rules were *dharma*, and deserved a place in authoritative texts on the subject (Rocher 1993).

2. The Commentaries

From about the seventh century CE onward, and well into the eighteenth century, the *smṛtis* become the object of voluminous commentaries. Both the commentaries on individual texts and the digests which treat of particular aspects of *dharma*, drew into their discussions the entire body of ancient *smṛti* texts. Among the several, often book-length, sections of the digests most often one or two are devoted to law: some deal with procedure (*vyavahāra*) and substantive law (*vivāda*) separately; others combine the two in a single treatise. There are even digests on specific aspects of substantive law, mainly on inheritance and on adoption.

The commentaries and digests follow the basic trends of the revealed, eternally valid *smṛti* texts. Even though the digests discuss the legal aspects of *dharma* in separate volumes, 'law' remains closely linked to other aspects of *dharma*: an adoption is still considered invalid if it is not accompanied by the *dattahoma*.

On the other hand, drawing on the entire body of *smṛti* literature as equally authoritative, the commentators and digest writers are able to fill in gaps in one *smṛti* by drawing on a different *smṛti*. With two exceptions, the *smṛti* texts tell us little about what happens to the inheritance of a father when he does not have any descendant beyond a great-grandson. The commentaries and digests appeal to all *smṛti* texts to establish complete, if rarely uniform, lists of heirs, from the son and other relatives, down to the king. The commentaries and digests do provide coherent, albeit different, systems of Hindu law.

Nowhere is that tendency expressed more clearly than in the contradictions referred to earlier. For the commentators, all *smṛti* texts are equally authoritative; there cannot be any contradictions, and it is their task to show this. They either present the rules as applicable under different circumstances, or they interpret one text literally and do not hesitate to force a meaning on the others, so that all texts are in perfect harmony. In the scholarly literature the often important and far-reaching difference between the interpretations put on the *smṛti* texts by the commentators are most often interpreted as attempts on the part of the latter to adapt the former to local circumstances and customs. Or, rather than being concerned with local customs, the commentators apply the same principles as their interpretations of *smṛti* texts in any other field of *dharma*: subtle grammatical explanations borrowed from Pāṇini's *Aṣṭādhyāyī* and other grammatical treatises, intricate rules of textual exegesis laid down in Jaimini's *Mīmāṃsāsūtras*, principles of logic contained in Gautama's *Nyāyasūtras*, etc. Each individual commentator brings his vast amount of traditional, panditic learning to bear on the *smṛti* texts, to mold them into one coherent system, to interpret them more correctly than his predecessors, with whom he does not hesitate to argue in the strongest possible terms.

Differently from the *smṛti* texts, which were revealed once for all through the intermediary of ancient *rishis*, the commentaries and digests were produced by human beings with whom it was possible to disagree as well as to agree.

3. Anglo-Hindu Law

In search of ways to govern Indians according to their own laws rather than by British Law, in 1772, under the impetus of the Governor of Bengal, Warren Hastings, the East India Company adopted the principle that 'in all suits regarding inheritance, marriage, caste and other religious usages or institutions, the laws of the Koran with respect to the Mohamedans and those of the Shaster with respect to the Gentoos shall invariably be adhered to.' The British administrators were familiar with Muslim law, but the law of 'the Shaster,' i.e., the *dharmaśāstras*, was totally unknown to them; in fact, at that time not a single Englishman had access to the language of the *dharmaśāstras*, Sanskrit. What followed, throughout the British period, is a series of remarkable attempts, on the one hand, to maintain the integrity of shastric law and, on the other hand, to interfere with it and abolish it whenever it came into conflict with basic principles of British justice.

In a first attempt to gain access to the sources of Hindu Law (1772), pandits were attached to the Anglo-Indian courts of law. The Court Pandit received the legal question via a translation from English or Bengali into Persian, and hence into Sanskrit. He consulted his law books, and made a decision, which was again translated, through Persian, into English. Besides being cumbersome and time-consuming, this procedure soon created uneasiness in British legal circles in India: distrust in the pandits and their decisions became a common topic in contemporary reports and correspondence. Even though the office of Court Pandits was not abolished until 1864, meanwhile some Englishmen decided that for them to get access to the sources of Hindu law, they needed to study Sanskrit. One of these early Orientalists was Sir William Jones, a judge in the Calcutta Supreme Court, who produced an English translation of the *Manusmṛti* (Jones 1794).

Even prior to Jones's translation, Hastings called to Calcutta a group of 11 pandits, to provide him with the code of Hindu law. While the pandits' composition, *Vivādārṇavasetu*, 'the bridge across the ocean of disputes,' was still going on, their work was translated into Persian, and from there into English by one of Hastings's protégés, Nathaniel Brassey Halhed. Halhed's book became known as *A Code of Gentoo Laws, or, Ordinations of the Pundits* (1776).

However, neither Halhed's *Code* nor Jones's *Institutes* proved to be helpful as law books in the Anglo-Indian courts of law. The former suffered much in the two successive translations (and is, in fact, very

different from the Sanskrit original); the latter displayed all the characteristics of an ancient *smṛti* text mentioned earlier: no answers to important legal questions, conflicting answers to the same questions, and so forth.

At that point the British turned away from the *smṛtis*, toward the commentaries which at least provided coherent legal systems and gave clear answers to questions raised in Court. On the other hand, each commentator's coherent legal system was, in varying degrees, different than that of any other commentator. Henry Thomas Colebrooke was first to introduce the concept of Schools of Hindu Law (Rocher 1972), a concept that was to dominate the history of Hindu law until the middle of the twentieth century. According to Colebrooke, there was one digest that was applicable in Bengal, and another commentary that was followed in the rest of India.

In 1810 he published a volume entitled *Two Treatises on the Hindu Law of Inheritance*, in which he translated *Jimūtavāhana's* digest *Dāyabhāga* (early twelfth century) as the text for Bengal, and the chapter on inheritance from the *Mitākṣarā*, a contemporaneous commentary on the *Yājñavalkyasmṛti*, as the text for India outside Bengal. In more recent years, as more commentaries and digests were translated, the 'Mitakshara School of Hindu law' was further subdivided into four subschools: the Benares, Mithila, Maharashtra or Bombay, and Dravida or Madras schools of Hindu law, each school and subschool administering Hindu law on the basis of its own Sanskrit text (or texts). The result was a pyramidal system in which the decision of any lower court could be appealed to the regional High Court, whose decisions were binding on the lower courts within its territory. The Judicial Committee of the Privy Council in London, the final court of appeal, made binding but different decisions depending on the region from where the appeal originated.

The basic difference between the two major Schools of Hindu law is that, in the *Mitākṣarā* school, any male individual born into a family, by his very birth acquires an unspecified right of ownership in joint family property. According to the *Dāyabhāga*, on the contrary, no one becomes the owner of any part of the family property before the time at which the right of the previous owner lapses, either by death or for any other reason that prevents him from exercising his obligations as the manager of the family property. This distinction inevitably had a profound impact on the lesser (*Mitākṣarā*) or greater (*Dāyabhāga*) power of the head of the family to manage the family property at his pleasure.

Notwithstanding these efforts on the part of the British to administer Hindu law to Hindus, and even though they continued to pay homage to the Sanskrit *dharmaśāstras*, at the same time they steered Hindu law in directions that were alien to the ancient *rishis* and the commentators.

First, based on the principle that 'we are here to administer law, not religion,' the Anglo-Indian courts drew the line between what they considered legal actions and the ritual activities that were an integral part of them in the Sanskrit texts: in Anglo-Hindu law, adoption became valid even without the performance of the *dattahoma*.

Second, since the commentaries and digests in each region became the principal sources of law, the Courts no longer considered it necessary to look for the correct meaning of the *smṛti* texts on which the commentaries were based; they relied invariably and solely on the commentators' interpretations, thereby reversing the original scale of authority between the human commentaries and the texts transmitted in the revealed *smṛtis*.

Third, even when Indians joined the Courts, the number of judges and advocates who knew Sanskrit remained small. All arguments and judgements were based on English translations. If one of the parties demonstrated that a translation was in error, the judges rejected that argument, and referred instead to precedents contained in the massive Indian Law Reports. From being based on texts, Hindu law now became a case law, and the law reports were interspersed with Latin phrases such as *stare decisis* and *communis error facit ius*.

Finally, there were traits of Hindu law that ran counter to basic British conceptions of law and justice. In many of these cases the judges overruled Hindu law on the basis of the principle of 'justice, equity, and good conscience.' When the sisters of a murderer claimed the inheritance of his victim, which in Hindu law their brother was entitled to because he was physically and mentally capable of performing the prescribed rituals (*śrāddha*) for the deceased, the case went through several appeals with opposite results, until the Privy Council rejected outright the murderer's right to inherit.

Conversely, Hindu law was expanded by the Courts' recognition of wills, which were unknown in the traditional system. In other cases, the British passed a number of Acts which overruled and abolished some existing practices of Hindu law. As the titles of the earliest Acts (*The Caste Disabilities Removal Act, 1850; The Hindu Widow's Remarriage Act, 1856; etc.*) indicate, they aimed at improving the living conditions of outcasts (a Hindu who renounced his religion or was excluded from his caste lost any right to property and inheritance) and women (when her husband died, his widow, however young she might be as a result of the common practice of child marriage, was not allowed ever to remarry).

4. Hindu Law in Independent India

In 1947 Hindu law as established by the British passed unchanged into the courts of Independent India.

Implementing article 44 of the Constitution: 'The State shall endeavour to secure for the citizens a uniform Civil Code throughout the territory of India,' was delayed. Even the more modest attempt at providing all Hindus in India with a uniform 'Hindu Code,' an idea that preceded Independence, failed. However, in 1955 the Indian Parliament enacted *The Hindu Marriage Act*, which applied not only to Hindus, but also to Buddhists, Jains, and Sikhs; in short to any Indian citizen 'who is not a Muslim, Christian, Parsi or Jew by religion' (art. 2 [1]). In 1956 *The Hindu Marriage Act* was followed by *The Hindu Succession Act*, *The Hindu Minority and Guardianship Act*, and *The Hindu Adoptions and Maintenance Act*. All four acts state, in almost identical terms, that:

Save as otherwise expressly provided in this Act,—

(a) any text, rule or interpretation of Hindu law or any custom or usage as part of that law in force immediately before the commencement of this Act shall cease to have effect with regard to any matter for which provision is made in this Act;

(b) any other law in force immediately before the commencement of the Act shall cease to have effect in so far as it is inconsistent with any of the provisions contained in this Act.

The impact of these modern Acts, overriding Hindu law as it had been practiced for a century and a half, is still a matter of dispute. The point is that, officially, they made large sections of traditional Hindu law obsolete. Any Hindu (in the broad sense) can legally marry any other Hindu, monogamous marriage becomes the rule of law, and so do annulment of marriage, divorce, and remarriage for women; succession to inheritance is made uniform, male and female heirs are treated equally; and women, who were accorded a limited estate in the past, now become full owners of their property and can dispose of it as they please.

There remains, however, an undetermined residue of Hindu law for which the Acts do not make any provision; in areas such as joint family law the Courts continue to administer Hindu law as it was prior to 1955 (Derrett 1963).

See also: India, Religions of; Law: Imposition, Reception, and Colonial; Postcolonial Law

Bibliography

- Bühler G 1886 *The laws of Manu*. In: *Sacred Books of the East*, 25. Clarendon Press, Oxford, UK
- Derrett J D M 1963 *Introduction to Modern Hindu Law*. Oxford University Press, London
- Derrett J D M 1973a *Dharmaśāstra and juridical literature*. In: Gonda J (ed.) *A History of Indian Literature*. Harrassowitz, Wiesbaden, Germany
- Derrett J D M 1973b *History of Indian Law (Dharmaśāstra)*. In: Spuler B, et al. (eds.) *Handbuch der Orientalistik*, II.3.1. Brill, Leiden, The Netherlands

- Halhed N B 1776 *A Code of Gentoo Laws, or, Ordinations of the Pundits*. East India Company, Calcutta, India
- Jolly I 1885 *Outlines of an History of the Hindu Law of Partition, Inheritance and Adoption, as Contained in the Original Sanskrit Treatise*. Thacker Spink, Calcutta, India
- Jolly J 1896 *Recht und Sitte (einschliesslich der einheimischen Litteratur)*. Trübner, Strassburg, Germany
- Jones W 1794 *Institutes of Hindu Law; or, the Ordinances of Menu*. Government Press, Calcutta, India
- Kane P V 1930–77 *History of Dharmaśāstra (Ancient and Mediaeval Religious and Civil Law)* 5 vols in 7 parts. Bhandarkar Oriental Research Institute, Poona, India
- Lariviere R W 1989 *The Nārada-smṛiti*. University of Pennsylvania Studies on South Asia 4–5, University of Pennsylvania, PA
- Lingat R 1967 *Les Sources du Droit dans le Système Traditionnel de l'Inde*. Mouton, Leiden, The Netherlands [1973 *The Classical Law of India*. trans. Manu, University of California Press, Berkeley, CA]
- Rocher L 1972 *Schools of Hindu law*. In: Ensink J, P Gaeffke (eds.) *India Maior Congratulatory Volume Presented to J. Gonda*, pp. 167–76. Brill, Leiden, The Netherlands
- Rocher L 1993 *Law books in an oral culture: The Indian Dharmaśāstras*. *Proceedings of the American Philosophical Society* 137: 254–67

L. Rocher

Hintze, Otto (1861–1940)

The Berlin historian Otto Hintze continues to fascinate those engaged in international historical research. In Germany, there was a particularly noticeable increase of interest in Hintze in the 1960s, in conjunction with an enduring resurgence of social historical schools. The rare combination of detailed investigation through decades of sources on the one hand and the elaborated approach of history's systematic contiguous disciplines on the other brought to light results which are still of great interest in modern research.

This is especially true for the comparative approach to political structural history, particularly that of the European states from the height of the Middle Ages to the nineteenth century. From an early stage, Hintze was at pains to incorporate the categories of political science and economics, constitutional law, and sociology into historiographical analysis, thus broadening the scope of the discipline. Despite modernization theory and more recent concepts of the comparative analysis of nation building, Hintze's advocacy of the comparative approach to constitutional and governmental history and the methodological initiatives taken in his works are still of great significance. Indeed, Hintze's approach remains indispensable to the investigation of the processes of state making.

Otto Hintze was born in the small Pomeranian town of Pyritz on August 27, 1861, as the son of a middle-

ranking civil servant. After studying for two years at the University of Greifswald, Hintze transferred to the Friedrich-Wilhelms-Universität of Berlin in 1880–1. There he studied, for example, under Johann Gustav Droysen, who was still in his final years of life. Hintze completed his doctoral studies in the spring of 1884 with a dissertation on medieval history (*König Wilhelm der Holländer. Ein Beitrag zur Geschichte des deutschen Interregnums*), supervised by the historian Julius von Weizsäcker. Gustav Schmoller, the economic and political scientist who was soon to have a considerable impact on Hintze, was not yet named as one of his academic teachers. Hintze followed his first rather narrow course of study in history with a second in legal and constitutional history (sciences of state). According to the records, this was a formative stage for Hintze, during which Schmoller introduced him to archival research on the Prussian administrative files. Even this, his very first research on unpublished source materials, was carried out in connection with early comparative projects. At the same time, it was Hintze's first practical preparation for his work on the *Acta Borussiae*, Gustav Schmoller's editing and research project on Prussian history in the eighteenth century in particular.

Hintze was engaged in this exhaustive project, which was based at the Prussian Academy of Sciences, in various functions over a period of decades. He was initially contracted in 1888 as a research fellow. In this capacity, he acquired an unparalleled command of the published and unpublished source material on the Brandenburg-Prussian state, its regions and provinces, classes and estates, before expanding the scope of his research to include the sixteenth and nineteenth centuries. He completed his postdoctoral thesis, a monograph on the history of the Prussian state economy (*The Prussian silk industry in the 18th century*), in 1895. In 1899 he was appointed professor at the University of Berlin, where—turning down several offers of chairs at other universities—he advanced to the status of full professor from 1900 to 1902–7, and continued to teach until prematurely given emeritus status in 1920. Besides his broad research, which had long since extended beyond the boundaries of Prussia, he owed this distinguished academic career to Gustav Schmoller's excellent contacts in political and administrative circles which is of vital importance in the view of the various animosities among contemporary historians.

The initial focus of Hintze's extensive writings was on constitutional and economic issues of Prussian history, culminating in 1901 in a lengthy monograph on the organization of the Prussian state administrative authorities around 1740. In this volume Hintze analyzed the making of the Prussian state, its development from a territorial state to a 'grand state.' He then reluctantly accepted an official commission to compile a comprehensive survey of the history of Brandenburg-Prussia. Several editions of this

work were published in 1915–16, leading to heated (interdenominational) controversies during World War I.

In the course of archival research spanning two decades, Hintze acquired an excellent grasp of the source material, establishing an empirical basis that enabled him, even in further-reaching investigations, to take concrete historical life into consideration, particularly that of more recent centuries. For Hintze, the exploration of Prussian history was never a limited end in itself, but a paradigm for the historical process *per se*. This does not mean to say that he took this historical example as a yardstick by which the development of other states was to be measured, however. By his own account, it was at an early stage of his career—during his studies of legal and constitutional history under Schmoller—that Hintze had identified a much broader area as his future field of research: general constitutional and administrative history, from the comparative perspective. Furthermore, in choosing his Prussian topics of inquiry, he had always considered their usefulness for the 'entirety of my academic plans and studies' (1892).

Long before the World War I, which marked a turning point in Hintze's life, he became engaged in other, broader fields of research. In the course of his dispute with the protagonist of the older German historical school of economics, Wilhelm Roscher, Hintze had presented his conception of a comparative approach to the analysis of world history as early as 1897. In contrast to Roscher, whom he accused of neglecting the fundamental social aspects of state structures, Hintze devised a procedure that integrated both social conditions and the effects of foreign policy and power politics into the analysis of the formation of political structures. Hintze's main concern was to identify the recurring patterns in the development of states by means of a typological and comparative procedure. 'In the social region, there is a tendency towards regulated organic development: the making of states follows this developmental trend to a certain extent, simultaneously regulating it, but is by no means dependent on it alone.' Hintze maintained that the implications of 'power interests' were not to be underestimated, that the establishment of a state was 'at least as dependent on its external position of power as on the preconditions of the particular social structure for domestic governmental activity,' and that the 'external conditions cannot fail to influence social development.' It was this 'political side' of the formation of political structures which Hintze felt was neglected in any typological theory of states which, following the Aristotelian example, merely tried to identify analogous sequences of state systems. Hintze considered the process of state making to be largely dependent on the 'interrelations of the states, on the forces prevailing amongst them, on the rise and fall of the neighbouring states,' i.e., on what Leopold von Ranke had called the 'great world relationships.' (see

Ranke, Leopold von (1795–1886)) Hintze considered power as a given category, which—like the phenomenon of state rivalries—he did not attempt to deduce in any further detail. For him, the central aim was to define and register ‘types conditioned by world history’: phenomena observed in groups of states that had maintained close political relations. In this sense, he explained afterwards, he understood state making to be a ‘general, regulative principle,’ thus implying that he did not intend strict rules to be postulated. It seems that in his early writings, Hintze still assumed a cautious primacy of foreign policy factors in the process of state making, but that—remarkable in the history of science—he turned away from this explanatory model over the course of his further research and analysis.

In a whole series of individual studies, whose scope was gradually to broaden, reaching the dimensions of world history in about 1930, Hintze applied and tested his typologically-based comparative approach. He developed what could be termed a spatial-typological procedure, which makes (European) historical development appear to be that of a bonded unit of regions. Above all, Hintze was keenly interested in the differentiation of a ‘core Europe’ (with a close genetic and spatial relationship to the Carolingian Empire) and a ring of ‘peripheral’ countries. He tested this spatial typological and structural historical differentiation at various junctures in his comparative studies of constitutional history—in his works on the history of the European estates, for example. It also becomes apparent in the comparison of continental western and central European absolutism on the one hand and the insular parliamentarianism of England on the other, a differentiation which has admittedly been called into question in recent Anglo-Saxon research. The ‘peripheral’ states of England, Sweden, Poland, and Hungary were of great interest to Hintze on account of their numerous structural parallels, especially where the ‘parliamentarian’ tradition was concerned.

Hintze set himself the aim of writing a comprehensive ‘general comparative constitutional and administrative history of the modern states, namely the Latin and Germanic peoples,’ but the manuscript was no longer published from around 1930 onwards. Central parts of this work were lost after Hintze’s death; the fragments that survived are currently being edited by an international team. New sources reveal that Hintze remained keenly interested in the ‘individuality’ of states and political processes. Of course, the historical legacy of Hintze’s work should not be allowed to conceal his innovative potential.

Otto Hintze sought to establish contact with the systematic contiguous disciplines at an early stage of his career, and not merely in reaction to the traumatic upheaval of political cultures subsequent to the 1914/1918 war. On the contrary, he had already begun to lose interest in the Prussian paradigm in the years

preceding the war. His openness to the more modern approaches taken by social history, psychology, and sociology became evident as early as 1900, in the mediating statement he made during the so-called Lamprecht dispute about the relationship of ‘the individualistic and the collectivist conception of history’ (1897). Hintze pointed to the significant contribution made to historiography by Comte and Spencer (see *Comte, Auguste (1798–1857)*; *Spencer, Herbert (1820–1903)*), both of whom he referred to in his lectures. For Hintze, individuality and collectivity constituted the two poles of historical insight, and he maintained that the methods of social psychology afforded great opportunities to go beyond Ranke. Admittedly, Hintze’s growing interest in the concepts of sociology characterizes him as an outsider among the historians of his time. Hintze’s great 1920s discussions on the theory and philosophy of history, not only with Weber (see *Weber, Max (1864–1920)*), but also with Oppenheimer, Sombart, and Troeltsch, as well as with the jurists Kelsen and Smend, merely marked the culmination of a continuous development in his work. This is demonstrated, for example, by Hintze’s discussion of Sombart’s *Modern Capitalism*, in which the Berlin historian also referred to Marx. The extent to which Hintze started from the assumption of a primary, autonomous concept of society is a matter of debate. While he regarded the state as the dominant factor, his later studies and the surviving fragments of manuscripts written prior to 1933 reveal marked changes and a broadening of approach—he now tended towards a concept of society which was not derived from the state alone. For Hintze, social and socioeconomic factors began to gain in causal-analytic significance; the same applies to the concept of class, though Hintze never approached Marxist conceptions with anything but a distant interest. It is thus also a matter of contention whether Hintze’s typological procedure, along with the principle of ‘clear abstraction’ (*anschauliche Abstraction*), *ipso facto* neglected the element of dynamics. Hintze’s procedure was related (or rather contiguous) to Weber’s conception of the ideal type (*Idealtypus*), though always referring to the central question of the individualities of historical developments. It should be noted that Georg Jellinek had made pertinent suggestions for a typological methodology as early as 1900 in his *Allgemeine Staatslehre (General Theory of State)*, and that Hintze was able to develop these ideas in his search for phenomena of ‘relative comparability’ (Kocka). In any case, Hintze saw his work as a necessary supplement to both the legacy of Ranke and modern sociology.

Hintze’s reception of Weber’s writings also predates 1914. Both the strength of his systematic interests and the positioning of methodological boundaries are evident in Hintze’s second unpublished magnum opus, which can only be reconstructed to some extent from lecture notes. After 1920, Hintze had intended to

publish this work as a 'general theory of state and society on a historical basis,' a continuation of and supplement to his general constitutional history. He had intended to paint a 'typical picture of the modern state in terms of its shared properties and developmental trends, its various historical stages and individual forms.' In this 'policy' of Hintze, which he evolved in his middle period, the primacy of the state and of the law was (still) very apparent. 'The state must be considered from the legal, the philosophical, and the historical perspective'—in this order. Social factors were suited and subordinated to this conception. In contrast to his classical studies of comparative constitutional history, in which industry and workers (or the labor movement) played at most a marginal role, Hintze integrated the social developments of the nineteenth century into his 'policy' by means of critical discussion of Marx and Lassalle. In so doing, Hintze was able to identify developments in the middle classes and nonmanual occupations that were to be examined in greater detail by later social historians.

All this makes Hintze still of interest to present-day historians. Of course, not all of his works are relevant today, as a brief glance at his political writings will show. However, his openness to the contributions of the social and political sciences—highly unusual among his contemporaries in the field—and the scope of his view, which finally extended beyond euro-centrism, mean that he remains an important, innovative figure in modern political structural history. Admittedly, the reconstruction of his *oeuvre* is essential to apply his approach in practical research also for further developing his theories, yet this is forever hindered by the fact that central works were never published and can only be substituted and inferred. After 1933 Otto Hintze, who had married the Jewish historian Hedwig Guggenhämer, became a *persona non grata* in Germany. He ordered that his personal papers should be destroyed after his death, which for the most part they were. Hintze died in his Berlin apartment on April 25, 1940, long since isolated from academic life.

See also: Comparative History; Constitutionalism; Constitutionalism, Comparative; Historiography and Historical Thought: Christian Tradition; Middle Ages, The; Nation: Sociological Aspects; Nations and Nation-states in History; Policy History: Origins; Policy History: Regimes; Policy History: State and Economy; Political History (History of Politics); State and Society; State Formation; State, History of; State, Sociology of the

Bibliography

Büsch O, Erbe M (eds.) 1983 *Otto Hintze und die moderne Geschichtsschäft* [Otto Hintze and Modern Historiography]. Colloquium Verlag, Berlin

- Hintze O 1908 *Historische und politische Aufsätze* [Historical and Political Essays]. Verlag Deutsche Bücherei, Berlin, 4 Vols.
- Hintze O 1962–1967 In: Oestreich G (eds.) *Gesammelte Abhandlungen* [selected works], 2nd and 3rd edns. Vandenhoeck & Ruprecht, Göttingen, 3 Vols.
- Hintze O 1998 Di Costanzo G, Erbe M, Neugebauer W (eds.) *Allgemeine Verfassungs- und Verwaltungsgeschichte der neueren Staaten. Fragmente* [General Constitutional and Administrative History of the Newer States. Fragments]. Palmar, Naples, Italy, Vol. 1
- Kocka J 1972 Otto Hintze. In: Wehler H-U (ed.) *Deutsche Historiker*. Vandenhoeck & Ruprecht, Göttingen, Vol. 3, pp. 41–64
- Kocka J 1981 Otto Hintze, Max Weber und das Problem der Bürokratie [Otto Hintze, Max Weber and the problem of bureaucracy]. *Historische Zeitschrift* 233: 65–105
- Neugebauer W 1993 Otto Hintze und seine Konzeption der 'Allgemeinen Verfassungsgeschichte der Neueren Staaten' [Otto Hintze and his conception of the 'General constitutional history of the newer states']. *Zeitschrift für historische Forschung* 20: 65–96
- Neugebauer W 1998 Die wissenschaftlichen Anfänge Otto Hintzes [The academic origins of Otto Hintze]. *Zeitschrift der Savigny-Stiftung für Rechtsgeschichte, Germanistische Abteilung* 115: 540–51
- Neugebauer W 1999 Zum schwierigen Verhältnis von Geschichts, Staats- und Wirtschaftswissenschaften [On the difficult relationship between history, political science and economics]. In: Kocka J (ed.) *Die Königlich Preußische Akademie der Wissenschaften im Kaiserreich*. Akademie Verlag, Berlin, pp. 235–75
- Neugebauer W 2000 Zur Quellenlage der Hintze-Forschung [On source material in research on Hintze]. *Jahrbuch für die Geschichte Mittel- und Ostdeutschlands* 45: 323–38
- Oestreich B 1989 Otto Hintze. In: Erbe M (ed.) *Berlinische Lebensbilder. Geisteswissenschaftler* [Berlin Lives. Arts Scholars]. Colloquium Verlag, Berlin, pp. 287–309
- Smith L S 1967 Otto Hintze's comparative constitutional history of the West. Doctoral dissertation, Saint Louis, MI

W. Neugebauer

Hippocampus and Related Structures

The hippocampus is part of the cortex of the forebrain in mammals and is closely linked to a restricted number of related cortical areas which are here collectively referred to as the parahippocampal region. This collection of cortical regions, which in humans is located in the most medial portion of the temporal lobe, is involved critically in learning and memory processes, in particular in long-term explicit memory. The hippocampus also has strong functional links with subcortical structures believed to play a role in the selection of appropriate behavior to favor either survival of the individual or of the species. In this

article, the wiring of the hippocampus and the parahippocampal region underlying these two proposed functions will be described, and experimental data that support the proposed functions will be summarized.

1. What is the Hippocampus and What are Related Structures?

The hippocampus is a macroscopically defined cortical structure which is present in the brains of all mammals. *Grosso modo*, its appearance looks alike in all species in that it is an enrolled continuation of the most medial rim of the cortical surface of the hemisphere. The cortices directly adjacent to the hippocampus, which are referred to collectively as the para (next-to) hippocampal region, are considered as the related structures. The particular arrangement of these different cortical structures can be appreciated easily in sections taken perpendicular to the long axis of the hippocampus (see Fig. 1).

The hippocampus and the parahippocampal region of different species are positioned differently in the brain, such that in species with a clearly developed temporal lobe, the hippocampus is positioned more ventrally and anteriorly (humans and monkeys) compared to the situation in, for example, the rat, where the hippocampus looks more like a c-shaped structure positioned in the caudal third of the hemisphere, such that its dorsal one-third has a marked spatial relation with the corpus callosum. However, these differences in position do not alter the major characteristics and the topological relations between the hippocampus and the parahippocampal structures (see Sects 2 and 3). Nor does it influence the fact that the most anterior/ventral portion of the hippocampus has a close spatial relationship with the amygdala (see *Amygdala (Amygdaloid Complex)*).

2. General Features of the Hippocampus/Parahippocampal Region and Species Differences

The hippocampus consists of three major subdivisions, which can be recognized easily in all species (see Fig. 1). The first field, the dentate gyrus, is a c-shaped, three-layered cortical structure which is characterized by a densely packed cell layer, mainly consisting of granular cells, i.e., cells with a round soma, and an apical dendritic tuft and no, or only sparse, basal dendrites. Enclosed in this c-shaped cortical blade is a polymorphic cell group generally referred to as the hilus of the dentate gyrus. The second subfield is the Ammon's Horn or Cornu Ammonis, generally abbreviated as CA-field. The CA-field is generally subdivided into CA1, CA2, and CA3. The third subdivision is the subiculum. Note that, according to some authors, a prosubiculum can be differentiated

from the subiculum proper. Also, readers may find texts in which reference is made to the subicular complex, consisting of (pro)subiculum, presubiculum, and parasubiculum. Based on connectional arguments, it is preferred to consider the presubiculum and parasubiculum as being functionally different from the subiculum (cf. Amaral and Witter 1995; and for a very detailed comparison between species the reader is referred to Stephan 1975).

The parahippocampal region, in contrast to the hippocampus, comprises cortical regions with more than three layers—generally, five or six layers are distinguished. In all species, the parahippocampal region comprises five different regions, the pre- and parasubiculum, the perirhinal and entorhinal cortices, and a fifth area which in primates is commonly referred to as the parahippocampal cortex. In nonprimates, the most likely homolog for the latter area is the so-called postrhinal cortex (cf. Burwell et al. 1995, Witter et al. 1989). The respective names of some of the areas that make up the parahippocampal region refer to the rhinal sulcus, which can be distinguished in all mammals. However, its particular features differ considerably between species. Whereas in rats it is a distinguishable fissure that runs along the rostrocaudal extent of the hemisphere, in humans it is only present in about 50 percent of the cases, as a shallow and short indentation in the most medio-anterior temporal lobe. It is of interest that the spatial relationships among the different fields of the parahippocampal region appear similar in, for example, mouse, rat, monkey, and human (see Fig. 2).

A remark is appropriate concerning the question of whether nonmammalian species do have a hippocampus and related structures. In reptiles, cortical areas have been described which resemble, both in terms of celltype as well as connectivity, the mammalian dentate gyrus and Ammon's Horn. Part of the olfactory-related cortex of reptiles may be considered homologous to the mammalian parahippocampal region (see Ten Donkelaar 1997 for further reading). Interestingly, there is some evidence that lesions of these structures in reptiles interfere with learning. With respect to birds, there is convincing evidence that a particular region in the forebrain, homologues to the reptilian cortex, is functionally important for spatial memory. This region is generally referred to as the hippocampus, but structurally it looks quite different from the mammalian hippocampus. Connection-wise, though, it is quite similar (see Clayton and Krebs 1995 for further details).

3. Wiring of the Hippocampus and Parahippocampal Region

Most detailed information about the connectivity of the system comes from anatomical tracing studies in

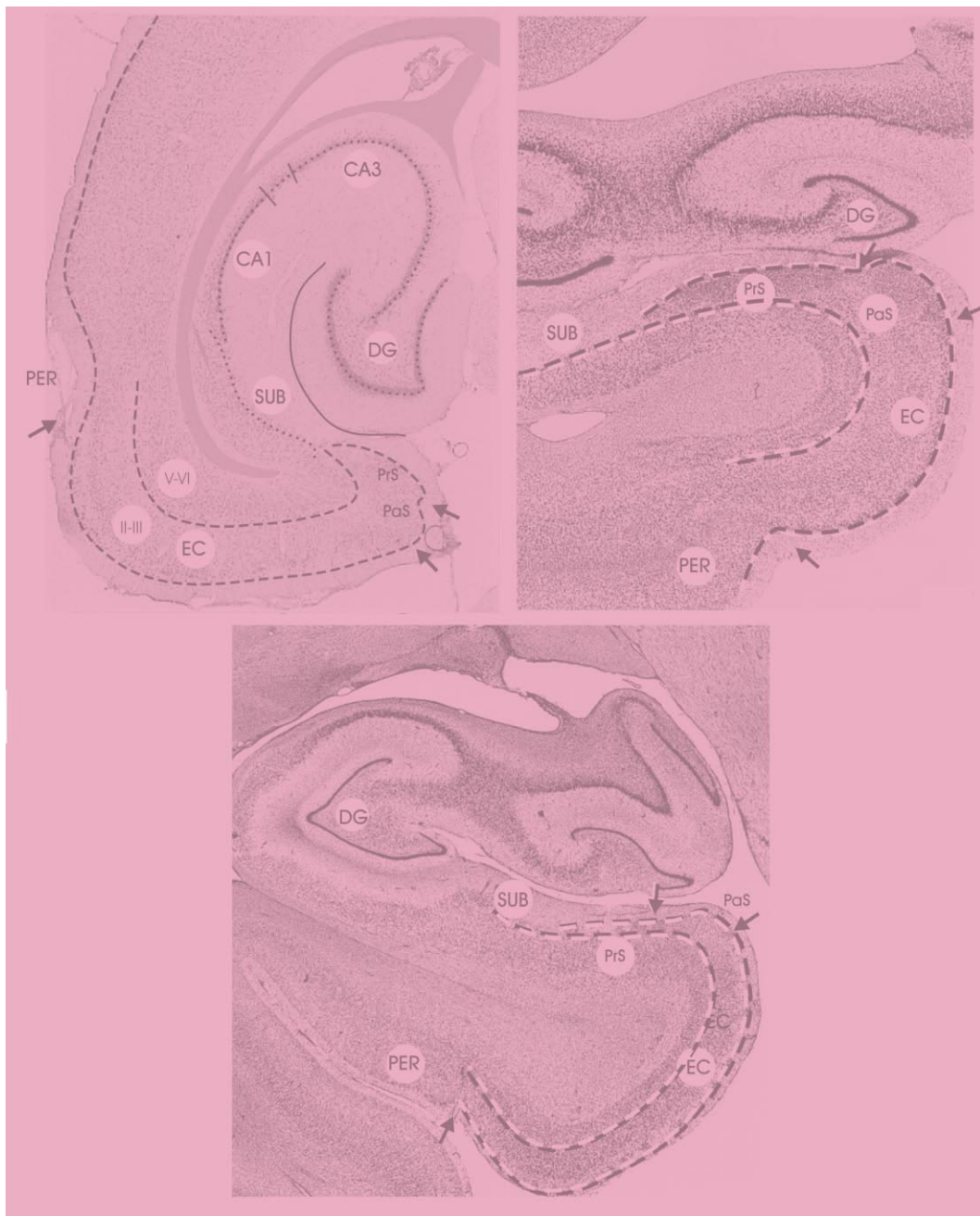


Figure 1

Three sections of the hippocampus and related parahippocampal structures cut perpendicular to the antero-posterior axis of the forebrain of the rat (upper left), the monkey (upper right), and the human (lower panel; modified after Scharfman et al. 2000, p. 11, reproduced by permission of the New York Academy of Sciences)

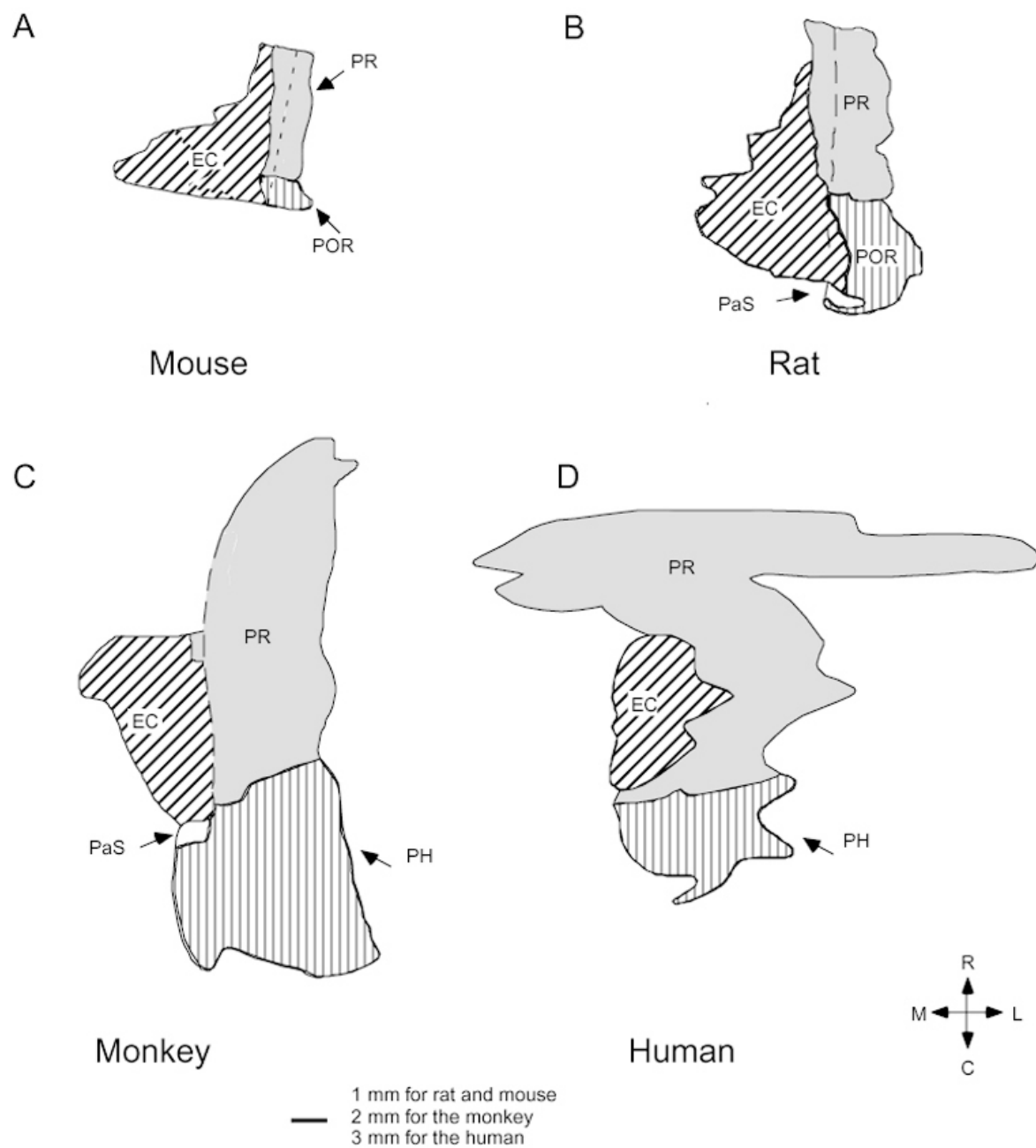


Figure 2

Two-dimensional representations of the parahippocampal region in the mouse, rat, monkey, and human, illustrating that in all four species this region comprises the same subdivisions but that their spatial arrangements and individual relative sizes differ between species. Note that the pre- and parasubiculum are not represented in these figures (from Burwell 2000, p. 29, reproduced by permission of the New York Academy of Sciences)

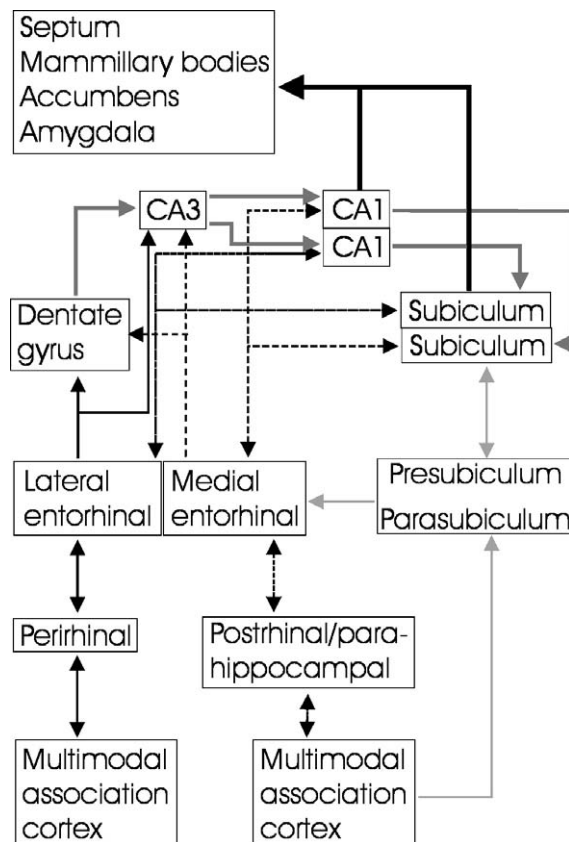


Figure 3

Summary diagram of connectivity of the hippocampus illustrating the two major communication pathways: 1, the cortico-hippocampal circuitry through the hippocampal region; 2, the fornix

the rat, although for some pathways, relevant detailed information is available for the monkey, and for the cat as well. The overall picture which emerges from all these studies is that this connectivity is rather conservative, such that a general non-species specific description may suffice. For further details, readers are referred to Witter et al. (1989, 2000a, 2000b), Amaral and Witter (1995), Burwell (2000), and Suzuki and Amaral (1994a, 1994b). The hippocampus has two major pathways through which it is connected to the rest of the brain. The first pathway makes use of the so-called cortico-parahippocampal-hippocampal circuit, and the second is formed by the fornix (see Fig. 3).

3.1 Cortico-Parahippocampal-Hippocampal Circuit

A striking feature of the cortico-parahippocampal-hippocampal circuit is that the hippocampus does not

have direct connections with the majority of the functionally different cortical regions. In contrast, this communication is mediated by way of the parahippocampal region. Of this region, the perirhinal and postrhinal/parahippocampal cortices maintain the highest number of cortico-cortical connections, including connections within all major sensory realms, such as visual, auditory, somatosensory, and gustatory ones. Also, massive connections exist with the other two major higher-order association domains, the parietal and prefrontal cortices, but connections with the motor cortex are less prevalent. The perirhinal and postrhinal/parahippocampal cortices, in turn, provide a major input to the superficial layers II and III of the entorhinal cortex, which are further reached by inputs from olfactory structures, the amygdala, and from the remaining two components of the parahippocampal region, i.e., the pre- and parasubiculum. Neurons in entorhinal layers II and III, which are the recipients of this wide variety of inputs, give rise to the major cortical input to the hippocampus, the so-called perforant pathway. This name is derived from the traditional descriptions by Ramon y Cajal (1911), who noted that fibers from the entorhinal cortex perforate the underlying white matter and the cortical lamina of the subiculum to gain access to the dentate gyrus molecular layer. In order to understand the potential functioning of the system, it is critical to point out that the perforant pathway projection harbors two different systems. Layer II cells distribute their axons to the dentate gyrus and CA3, whereas layer III cells send their axons to CA1 and the subiculum. In addition to this intriguing difference in target specificity, the terminal distributions of these two components of the perforant pathway are entirely different. Layer II cells give rise to a widespread projection to all dentate granular cells. In contrast, axons of layer III cells target only restricted groups of the available neurons in CA1 and the subiculum (see Sect. 5.1 for functional relevance; see also Fig. 3). This particular feature is of interest, since field CA1 and the subiculum constitute the major output structures of the hippocampus. Hippocampal output from these two hippocampal fields is distributed by way of projections back to the entorhinal cortex, from where ongoing output projections originate to a host of cortical and subcortical structures, including projections to perirhinal and parahippocampal/postrhinal cortices, prefrontal areas, amygdala, ventral striatum, and midline thalamus.

A further refinement has to be added to the overall organization of the connections between the parahippocampal region and the hippocampus. It has been established that the input-output system consists of at least two parallel pathways. One of the pathways carries inputs from the perirhinal cortex into and away from the hippocampus, whereas the second mediates transfer of information from and to the parahippocampal/postrhinal cortex. In view of the anatomical

organization outlined above, it has thus been suggested that in the hippocampus these two pathways converge at the level of the dentate gyrus and CA3, whereas they are kept more or less separate at the level of CA1 and the subiculum (cf. Witter et al. 2000b).

3.2 *The Fornix*

The fornix is a major fiber bundle which connects the hippocampus to the hypothalamus, in particular the mammillary bodies. On its way, the fornix also issues fibers to the septal complex, the ventral striatum, and the amygdala. As indicated above, the fornix is not a pure output pathway, since projections from the septal complex (see Sect. 5.2) to the hippocampus and in part to the entorhinal cortex travel by way of the fornix. In addition, the commissural connections between the left and right hippocampi also partially travel by way of the fornix. The hippocampal connection to the mammillary bodies is part of the traditionally described limbic or Papez circuit, which includes the mamillo-thalamic tract connecting the mammillary bodies with the anterior complex of the thalamus, which in turn project to large portions of the limbic cortex, including anterior- and posterior-cingulate cortex, and pre- and parasubiculum (see also *Cingulate Cortex* and *Limbic System*). All these structures, in turn, provide input to the hippocampus, either directly, or indirectly by way of the entorhinal cortex.

4. *The Hippocampus is not a Homogeneous Structure*

When one looks at the overall anatomical organization of the parahippocampal-hippocampal system, it appears that the hippocampus is rather homogeneous along its long axis. However, differences in numbers of neurons between the dorsal/posterior and ventral/anterior hippocampus have been reported, indicative for differences in the overall neuronal network. Moreover, in the rat and in the monkey, it has been established that the dorsal/posterior portion of the hippocampus receives its major cortical input from the more sensory-related or exteroceptive lateral and caudal portions of the entorhinal cortex, whereas the ventral/anterior hippocampus receives inputs from more medial parts of the entorhinal cortex, which most likely convey information that reflects the interoceptive status of the individual. In addition, direct reciprocal connections with the amygdala are limited to the ventral/anterior and intermediate levels of the hippocampus and are absent for the dorsal/posterior portion of the hippocampus (cf. Witter 1986). Behavioral data in support of this notion are the findings that in rats the dorsal hippocampus is critical for the initial learning and long-term memory of spatial

information, whereas the ventral hippocampus is not essential (cf. Moser and Moser 1998). Recently, brain imaging studies have revealed that in humans functional differences are also present between anterior and posterior halves of the hippocampus; however, the precise nature of these differences is still an issue for debate (Lepage et al. 1998, Schacter and Wagner 1999).

5. *Functions of the Hippocampus and Parahippocampal Region*

5.1 *Functions of the Cortico-Hippocampal System*

Ample clinical and experimental evidence supports the view that the hippocampus and parahippocampal structures are essential for memory processes, although the unique contributions of each of the different subfields are as yet incompletely understood (Scoville and Milner 1957, Squire 1992, Eichenbaum et al. 1994). The results of the behavioral studies, in particular, have provided insights into how each of the components of the hippocampal system contributes to memory processes. Instead of contributing in a unitary way to memory, it appears evident that individual structures make selective contributions. Such studies indicated that the hippocampal formation itself does not contribute critically to object recognition memory. Instead, it is the laterally positioned component of the parahippocampal region, the perirhinal cortex, that appears to be involved critically in this task (Murray et al. 2000). In contrast, the more posterior portions of the parahippocampal region, the parahippocampal cortex in primates or the postrhinal cortex in nonprimates, may be involved critically in spatial or topographical memory (Bohbot et al. 2000, Vann et al. 2000, cf. Witter et al. 2000b). Concerning the role of the entorhinal cortex, the results of lesion studies appear inconclusive; still, it is most likely that the entorhinal cortex operates cooperatively with the hippocampal formation, such that both structures process information in different, yet complementary ways. With respect to the hippocampal formation, it has been proposed that it plays a critical role in mediating memory for relevant relationships among conjunctions or associations unique to a particular episode (Eichenbaum et al. 1994).

5.2 *Functions of Hippocampus-Fornix Connectivity*

As indicated above, the fornix provides the hippocampus with another pathway, beside the cortical route, to communicate with a host of subcortical brain structures, including the mammillary bodies. Behavioral data indicate a clear functional dichotomy between the cortical versus the fornix route. However,

the role of the fornix pathway in learning and memory is as yet poorly understood. Lesions of the main target of the fornix, i.e., the mammillary complex in animals, do not result in marked memory deficits, although in Korsakoff's patients, which do show a marked anterograde as well as retrograde memory deficit, the mammillary bodies show a marked decrease in volume (cf. Aggleton and Brown 1999). In addition, there is convincing human evidence from patients with thalamic infarctions in favor of a role in memory processes of the connection from the mammillary complex to the anterior complex of the thalamus (van der Werf et al. 2000).

The fornix also mediates connectivity from the hippocampal-parahippocampal structures to the ventral striatum and the amygdala. In the ventral striatum, interactions may take place between hippocampal, amygdaloid, and prefrontal outputs, leading to the selection of appropriate behavioral strategies, for example, in learning paradigms (consult Groenewegen et al. 1999 for further reading). The functional relevance of the reciprocal connections between the amygdala and hippocampus, and for that matter also those between the amygdala and the parahippocampal region, is still poorly understood (Pitkanen et al. 2000, Stefanacci et al. 1996). It is assumed that amygdala inputs may provide the hippocampal system with a marker for relevance to a particular memory episode.

The connections to and from the septal complex of both the hippocampus and parahippocampal structures use the fornix as well as a number of other white matter pathways. Although much about these interactions has yet to be resolved, it is clear that the normal functioning of the hippocampal-parahippocampal-cortical interactions depends strongly on a functional septal complex. It is in particular the fact that the septal complex supplies the hippocampus and parahippocampal region with most of its cholinergic input, as well as the role of the septal complex in the synchronization of electrical activity which relates to its functional relevance (see Sect. 5.3).

5.3 Oscillatory Activity

The hippocampus shares with the entorhinal cortex the presence of strong rhythmical activity, in particular theta rhythm, which is coexpressed with trains of gamma waves during exploring behavior in the awake animal as well as during a particular sleep phase called Rapid Eye Movement (REM) sleep. Both structures also share the occurrence of so-called sharp waves, a high-frequency discharge which occurs predominantly during the alternating sleep phase, the so-called slow-wave sleep phase. Although the electrophysiological mechanisms underlying the occurrence of such oscillatory phenomena have been clarified to some point, the functional relevance, for example, in relation to learning, is still debated. There appears to be consensus though that the normal functioning of the

cortico-hippocampal learning system depends strongly on the normally occurring alternations between the theta and the sharp wave state (cf. Buzsaki 1996).

5.4 Clinical Relevance

The hippocampal-parahippocampal structures play an established role in learning and memory processes, which is supported by reports concerning pathological changes seen in diseases that are associated with learning and memory disorders such as Alzheimer's disease. However, other neurological/psychiatric disorders, such as epilepsy, schizophrenia, Huntington's, Parkinson's and Pick's disease are associated with pathological alterations in these regions as well. The most striking observation, however, is the specificity of the pathology in each of these diseases. In all except epilepsy and schizophrenia, the first pathological changes are seen in the anterolateral part of the entorhinal cortex, also called the transentorhinal cortex. However, in all these diseases, the layer showing the most dramatic alterations seems to be disease specific (Braak et al. 2000). In the case of epilepsy, the most striking initial pathology is observed in layer III of the medial entorhinal cortex (Schwarcz et al. 2000), whereas in the case of schizophrenia, the pathology appears to start in layer II throughout the entorhinal cortex (Arnold 2000). In all these disorders, the hippocampus becomes involved as well, leading to an overall dramatic reduction in neuronal numbers and/or connections.

See also: Alzheimer's Disease, Neural Basis of; Learning and Memory: Computational Models; Learning and Memory, Neural Basis of; Long-term Depression (Hippocampus); Long-term Potentiation (Hippocampus)

Bibliography

- Aggleton J P, Brown M W 1999 Episodic memory, amnesia and the hippocampal-anterior thalamic axis. *Behavioral and Brain Sciences* **1999**: 425–89
- Amaral D G, Witter M P 1995 Hippocampal formation. In: Paxinos G (ed.) *The Rat Nervous System*. Academic Press, San Diego, CA
- Arnold S 2000 Cellular and molecular neuropathology of the parahippocampal region in schizophrenia. *Annals of the New York Academy of Science* **911**: 275–92
- Bohbot V D, Allen J B J, Nadel L 2000 Memory deficits characterized by patterns of lesions to the hippocampus and parahippocampal cortex. *Annals of the New York Academy of Science* **911**: 355–68
- Braak H, Del Tredici K, Bohl J, Bratzke H, Braak E 2000 Pathological changes in the parahippocampal region in select non-Alzheimer dementias. *Annals of the New York Academy of Science* **911**: 221–39
- Burwell R D 2000 The parahippocampal region: corticocortical connectivity. *Annals of the New York Academy of Science* **911**: 25–42

- Burwell R D, Witter M P, Amaral D G 1995 Perirhinal and postrhinal cortices of the rat: a review of the neuroanatomical literature and comparison with findings from the monkey brain. *Hippocampus* **5**: 390–408
- Buzsaki G 1996 The hippocampal–neocortical dialogue. *Cerebral Cortex* **6**: 81–92
- Clayton N S, Krebs J R 1995 Food-storing memory and the hippocampus. *Current Opinion in Neurobiology* **5**: 149–54
- Eichenbaum H, Otto T, Cohen N J 1994 Two functional components of the hippocampal memory system. *Behavioral and Brain Sciences* **17**: 449–72
- Groenewegen H J, Mulder A B, Beijer A V J, Wright C I, Lopes da Silva F H, Pennartz C M A 1999 Hippocampal and amygdaloid interactions in the nucleus accumbens. *Psychobiology* **27**: 149–64
- Lepage M, Habib R, Tulving E 1998 Hippocampal PET activations and memory encoding and retrieval: The Hiper model. *Hippocampus* **8**: 313–22
- Moser M B, Moser E I 1998 Functional differentiation in the hippocampus. *Hippocampus* **8**: 608–19
- Murray E A, Bussey T J, Hampton R R, Saksida L M 2000 The parahippocampal region and object recognition. *Annals of the New York Academy of Science* **911**: 166–74
- Pitkanen A, Pikkariainen M, Nurminen N, Ylinen A 2000 Reciprocal connections between the amygdala and the hippocampal formation, perirhinal cortex and postrhinal cortex in the rat. *Annals of the New York Academy of Science* **911**: 369–91
- Ramón y Cajal S 1911 *Histologie du Système Nerveux de l'Homme et des Vertébrés*. Maloine, Paris
- Schacter D L, Wagner A D 1999 Medial temporal lobe activations in fMRI and PET studies of episodic encoding and retrieval. *Hippocampus* **9**: 7–24
- Scharfman H E, Witter M P, Schwarcz R 2000 Preface. *Ann. New York Acad. Sci* **911**: ix–xiii
- Schwarcz R, Eid T, Du F 2000 Neurons in layer III of the entorhinal cortex. A role in epileptogenesis and epilepsy? *Annals of the New York Academy of Science* **911**: 328–42
- Scoville W B, Milner B 1957 Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry* **20**: 11–21
- Squire L R 1992 Memory and hippocampus: A synthesis of findings with rats, monkeys and humans. *Psychology Review* **99**: 195–231
- Stefanacci L, Suzuki W A, Amaral D G 1996 Organization of connections between the amygdaloid complex and the perirhinal and parahippocampal cortices in macaque monkeys. *Journal of Comparative Neurology* **375**: 552–82
- Stephan H 1975 Allocortex. In: Borgman W (ed.) *Handbuch der Mikroskopischen Anatomie des Menschen* IV-9. Springer-Verlag, Berlin, pp. 1–998
- Suzuki W A, Amaral D G 1994a The perirhinal and parahippocampal cortices of the Macaque monkey: Cortical afferents. *Journal of Comparative Neurology* **350**: 497–533
- Suzuki W A, Amaral D G 1994b Topographic organisation of the reciprocal connections between the monkey entorhinal and the perirhinal and parahippocampal cortices. *Journal of Neuroscience* **14**: 1856–77
- Ten Donkelaar H J 1997 Reptiles. In: Nieuwenhuys R, Ten Donkelaar H J, Nicholson C. (eds.) *Central Nervous System of Vertebrates*. Springer-Verlag, Berlin, Vol 2, pp. 1315–524
- van der Werf Y D, Witter M P, Uylings H B M, Jolles J 2000 Neuropsychology of infarctions in the thalamus: A meta-analysis. *Neuropsychologia* **38**: 613–27
- Vann S D, Brown M W, Erichsen J T, Aggleton J P 2000 Fos imaging reveals differential patterns of hippocampal and parahippocampal subfield activation in rats in response to different spatial memory tests. *Journal of Neuroscience* **20**: 2711–18
- Witter M P 1986 A survey of the anatomy of the hippocampal formation, with emphasis on the septotemporal organization of its intrinsic and extrinsic connections. In: Schwarcz R, Ben-Ari Y (eds.) *Advances in Experimental Medicine and Biology, Vol. 203: Excitatory Amino Acids and Epilepsy*. Plenum Press, New York, pp. 67–82
- Witter M P, Groenewegen H J, Lopes da Silva F H, Lohman A H M 1989 Functional organisation of the extrinsic and intrinsic circuitry of the parahippocampal region. *Progress in Neurobiology* **33**: 161–254
- Witter M P, Naber P A, van Haeften T, Machielsen W C M, Rombouts S A R B, Barkhof F, Scheltens Ph, Lopes da Silva F H 2000a Cortico-hippocampal communication by way of parallel parahippocampal–subicular pathways. *Hippocampus* **10**: 398–410
- Witter M P, Wouterlood F G, Naber P A, van Haeften T 2000b Anatomical organization of the parahippocampal–hippocampal network. *Annals of the New York Academy of Science* **911**: 1–24

M. P. Witter

Historical Archaeology

Historical archaeology is the concurrent study of material culture and documentary evidence. This article outlines the scope of the practice of historical archaeology, some of its principal features as a discipline, and some of the major research questions for which historical archaeologists seek answers.

In a strict sense, historical archaeology should begin with the discovery of writing, for it is from this time onwards that societies are able to express themselves both in the ways in which they fashion the world around them—their artifacts—and also through recorded self-expression: the written record. But in practice, these early literate periods have been left to classical archaeology, and historical archaeologists have defined their domain as the last 500 years. This half millennium saw the creation of extensive ‘world systems’ and, in particular, the colonization of the Americas, Africa, and parts of Asia by Europeans. This focus has given historical archaeology its particular character, and some have gone so far as to describe it as the archaeology of European colonial expansion. But more recently—as we shall see—these conventions have been challenged, opening up the possibility of a dynamic and expanding area of enquiry.

The practice of historical archaeology is rooted in the study of the European settlement of North America and, in particular, of the early seventeenth century colonial settlement in New England and the Chesapeake. Here, a rich heritage of material culture

has long been closely associated with a sense of national identity, best represented by the archaeology of Colonial Williamsburg, the eighteenth century capital of Virginia which has been preserved and reconstructed as an interpretation of daily life in America's early years (Colonial Williamsburg 2000). This focus has resulted in a rich narrative of early settlement, best represented by Ivor Noel Hume's excavation and interpretation of Martin's Hundred (Noel Hume 1983). Here, we see the carefully excavated traces of daily domestic life matched with an imperfect documentary record to reveal a far fuller narrative history than words alone could tell. Martin's Hundred is matched by numerous other studies that provide an ever-emerging picture of life in the early American colonies.

More recently, this emphasis on the settler and colonist has been balanced by the emergence of an archaeology of slavery. Archaeologists of slavery—whether working on former plantations in the south or on slave living sites in northern towns—have made the point that archaeology's claim to be the history of ordinary people must remain hollow if all dimensions of colonial life are not fully explored (Singleton 1995). In this area, historical archaeology has made notable gains in revealing domestic life on plantations and some of the aspects of identity, including the persistence of elements of African material culture in house design and personal artifacts, distinctive culinary practices, and patterns in composite material culture assemblages.

Historical archaeology in North America now extends well beyond this 'classic' focus and there is a variety of specialized subfields. These include the archaeology of western frontier settlements, battlefield archaeology, the archaeology of industrial sites, and urban archaeology. The practice of historical archaeology has also been strongly influenced by the professional field of cultural resource management, (CRM) as large State and federal projects have required numerous surveys and excavation projects ahead of development and urban renewal projects. CRM work has ranged from small studies of individual archaeological features to comprehensive surveys and extensive excavations in the face of freeway construction. Such work tends to be characterized by a heavy emphasis on fieldwork and recording, and less attention given to analysis, interpretation and historical synthesis, leading to a division between field-oriented archaeologists working in cultural resource management and archaeologists working in museums, universities, and other public institutions.

Despite a fondness for the self-image of a rugged fieldworker, uncluttered by intellectual pretensions, historical archaeology has reflected the major theoretical trends of the social sciences over the past half century (Hall 2000, Orser and Fagan 1995). Early work—such as Noel Hume's study of Martin's Hundred—was avowedly narrative. In this approach

(which has continued through to the present) the results of archaeological excavations serve to amplify and illustrate an incomplete documentary history, contributing to telling a story that is made richer by the tangible traces of the past. This empirical emphasis was challenged from several directions. Firstly, the 'New Archaeology' of the 1960s and 1970s sought law-like generalizations of human behavior through the evidence of material culture—an approach represented in Stanley South's search for standard patterns across archaeological sites (South 1977).

This approach was challenged by structuralism, best represented by James Deetz' widely read *In Small Things Forgotten* (1996). Here, material culture in its full breadth—from pipe-stems to houses—was organized as evidence of the human mind at work in creating systems of binary oppositions, together constituting 'world views.' Structuralism reassessed the principles of narrative history in seeking to show how European settlers 'became American' when they adopted an embedded mental template that determined how their world was arranged—the essence of cultural identity. But, in emphasizing the commonality of world views, structuralism lost a clear sense of cause and effect, and the processes of history unfolding. This prompted new interpretative approaches that turned to historical materialism for theoretical models. Well represented by Mark Leone's work in Maryland's Annapolis (1988), this approach has seen in the archaeological record the evidence for the emergence of merchant capitalism and social and economic control through the manipulation of material culture to create a false consciousness that disguises the true location of power. For example, Leone showed how garden design served to naturalize the authority of the new class of merchant capitalists in Annapolis, and how such naturalization extended to other categories of material culture, including musical instruments and ceramics.

More recently, the overwhelming association between historical archaeology and North America has been challenged by strong developments in other parts of the world. Each regional focus has tended to develop its own identity while remaining linked to the core concepts of this discipline. This is opening up the possibilities of a richer body of theory and of new areas of comparative study. One such area of development has been in Australia, where the archaeology of colonial settlement has gained increasing attention (Connah 1988). Here, studies have ranged from the historical archaeology of colonial buildings and early settler footholds on the continent to women's prisons in Tasmania. This has been matched by the development of historical archaeology in South Africa, where Dutch settlement in the mid-seventeenth century was followed by a century of British colonial settlement and a diverse material record (Hall 1993). Newer work in areas such as the Caribbean and in South America is adding a comparative richness to the field.

Despite these developments, though, historical archaeology still remains overwhelmingly the archaeology of European colonial expansion. Despite notable exceptions (for example, Johnson 1996), the archaeology of Europe itself in recent centuries still remains a separate field—'post-Medieval archaeology.' The archaeology of colonial contact—and the impact of European colonial expansion on the other side of colonial frontiers—remains largely unexplored. Claims for other forms of historical archaeology, for example the use of oral traditions as verbal sources (Andah 1990), remain largely unacknowledged. And the historical archaeologies of other large-scale, literate systems such as those of China or of Islamic expansion and settlement remain outside the scope of the field.

For some, historical archaeology is no more than a transition zone between preliterate periods, in which material culture is the only source of evidence, and documented history. But for others, historical archaeology has a distinctive character that makes it a discipline in its own right. There is some justification in this second point of view. Increasingly, historical archaeology has a distinctive set of methodologies and a suite of particular research questions of its own. Historical archaeology has its own professional journals, and its practitioners attend their own conferences. If this were the world of natural organisms, we could say that speciation is well advanced.

Historical archaeology shares with the rest of archaeology the basic methodologies of excavation. But although historical archaeologists often work with the precision of the prehistorian, teasing evidence from the slightest detail, they also often work to a larger scale. This is because the archaeological sites are large, with complex histories of deposition and disturbance which make the mechanical excavator a more appropriate research tool than the trowel. In addition, a large proportion of historical archaeological evidence is not beneath the ground. Standing buildings, field systems, roads and town layouts can reveal as much about the past as a conventional excavation. This brings historical archaeology close to fields such as ecology, cartography, architectural studies, and historical geography in its methodologies. Indeed, while retaining its ties with the time-honored skills of excavation, historical archaeology is quintessentially interdisciplinary in its search for, and analysis of, evidence.

See also: Art History; Cartography; Cultural Resource Management (CRM); Conservation of Cultural Heritage; Environmentalism; Preservation and Conservation; Historical Geography; Writing, Evolution of

Bibliography

Andah B W 1990 The oral versus the written word in the cognitive revolution: Language, culture and literacy. *West African Journal of Archaeology* 20: 18–45

- Colonial Williamsburg Foundation 2000 Williamsburg restored. <http://www.history.org>. 10 January 2000
- Connah G 1988 *'Of the Hut I Built': The Archaeology of Australia's History*. Cambridge University Press, London
- Deetz J 1996 *In Small Things Forgotten. An Archaeology of Early American Life*, (expanded and revised). Anchor Books, New York
- Hall M 1993 The archaeology of colonial settlement in southern Africa. *Annual Review of Anthropology* 22: 177–200
- Hall M 2000 *An Archaeology of the Modern World*. Routledge, London
- Johnson M 1996 *An Archaeology of Capitalism*. Blackwell, Oxford, UK
- Leone M 1988 The Georgian order as the order of merchant capitalism in Annapolis, Maryland. In: Leone M, Potter P (eds.) *The Recovery of Meaning. Historical Archaeology in the Eastern United States*. Smithsonian Institution, Washington DC, 235–61
- Noel Hume I 1982 *Martin's Hundred: The Discovery of a Lost Colonial Virginia Settlement*. Knopf, New York
- Orser C E, Fagan B M 1995 *Historical Archaeology*. Harper-Collins, New York
- Singleton T A 1995 The archaeology of slavery in North America. *Annual Review of Anthropology* 24: 119–140
- South S A 1977 *Method and Theory in Historical Archaeology*. Academic Press, New York

M. Hall

Historical Change and Human Development

Human development refers to patterns of change and stability that characterize a person from birth to death. Many theories maintain that development is highly plastic, which means that it is not predetermined in its processes or outcomes. In fact, human development is highly responsive to social and historical change. This article examines how behavioral scientists interrelate history and biography: how the particularities of historical circumstance lead to unique patterns of development.

Historical data are also used to test general theories of the family, life course, and life span. This approach does not deny that historical forces operate in unique ways to shape development, but it does consider the generalizability of theories across historical settings. Accordingly, this article then examines how historical data are used to test general theories of behavior.

1. The Historical Time Frames of Human Development

Human development reflects mechanisms that operate in three historical time frames, each of which is studied from different disciplinary standpoints. First, human behavior reflects evolutionary mechanisms that operate across historical epochs covering many

millions of years. These mechanisms are studied by biologists, anthropologists, and evolutionary psychologists who share a set of Darwinian assumptions. Second, individual development reflects the unique influences of historical periods that are comparatively limited in their duration. For example, social historians and students of psychohistory examine how the organization and ideology of Medieval society led to specific biographical patterns. Finally, life-span psychologists and life course sociologists examine connections between human lives and the 'historical present,' which encompasses the historically specific circumstances of a person's life (e.g., the Great Depression or the 1970s).

1.1 Historical Epochs: Phylogeny, Ethology, and Evolution

One of the most influential (albeit erroneous) contributions to the study of human development has been the principle of recapitulation, which emerged in the late eighteenth and early nineteenth centuries. Although many nineteenth century biologists subscribed to recapitulation, its greatest proponent was Ernst Haeckel, whose biogenetic law maintained that ontogeny (the life history of the person) recapitulates phylogeny (the evolutionary history of the lineage leading to the person). That is, the morphological development of each person quickly repeats—indeed, is caused by—the adult forms of the evolutionary history that preceded it. For example, Haeckel argued that because 'the human egg is a simple cell, we may at once infer ... a unicellular ancestor of the human race resembling an Amoeba.' Similarly a later embryonic form in humans points to a 'primitive worm-like ancestral form which is related to the sea squirts or Ascidians ...' (quoted in Gottlieb 1992, pp. 29–30). Because this recapitulation takes place in the embryonic and juvenile stages of development, the further evolution of a species depends on terminal addition, the appending of new characteristics to 'the end of an unaltered ancestral ontogeny' (Gould 1977, p. 74).

The biogenetic law fails principally because there is no known mechanism that links the evolution of a lineage with the development of an individual. And yet this law is important for two reasons. First, even though it was commonly rejected in biological circles in the early twentieth century, various forms of psychological recapitulation continued to influence many developmentalists, including Baldwin, Freud, Hall, Piaget, and Preyer (Gould 1977). For example, the biogenetic law (as it addressed both morphological and behavioral development among its proponents) and Lamarckian biology influenced Freud in pervasive ways. Thus, Freud held that the psychoanalytic stages of development (e.g., the oral and anal stages) were recapitulations of the phylogenetic sexual stages of our remote ancestors (Sulloway 1979).

This influence also is illustrated clearly in *Totem and Taboo* (Freud [1913] 1962) and such later works as *Moses and Monotheism* ([1939] 1967). In these works, Freud elaborates his theory of psychosexual recapitulation—this time focused on the Oedipus complex—involving an early trauma, a defensive reaction, a latency period, and then the outbreak of a neurosis. This theory, he believed, characterized the history of the human species, the history of the Judeo-Christian tradition, and the life history of the individual.

In the history of the human species, following the Ice Age a horde of brothers killed and consumed their father because of their sexual desire for his wives. Later, the brothers atoned for this act by deifying the slain father, now represented by a totem, and by observing prohibitions against parricide and incest. This sequence of acts was probably repeated over the millennia and was traumatic; consequently, by way of Lamarckian inheritance (which posits that acquired characteristics can be transmitted to successive generations), the feelings that surrounded these events were passed to later generations and now reside deep in the memory of all people. That is, according to Freud, the Oedipus complex is an inherited psychological structure rooted in phylogeny.

The life history of the individual repeats this sequence. Childhood sexual traumas (either real or conjured from phylogenetic fantasies that often involve the Oedipus complex) are repressed and forgotten during a period of latency. Additional traumas or disappointments in adulthood can then lead to high levels of distress and often to a neurosis. Thus, mental disturbances in adulthood need not trace back to a trauma that is consciously experienced by people during their own childhood; all humans are born with inherited traumas that reside deep in the unconscious and that can compromise pathways of adult development. Unfortunately, Freud's theoretical edifice is flawed because of its strong and erroneous biological assumptions. Referring to Lamarckian biology, Freud himself remarked in *Moses*, 'If things are different, then we are unable to advance one step further on our way, either in psychoanalysis or in mass psychology.' (Freud [1939] 1967 p. 128, Sulloway 1979, see also *Freud, Sigmund (1856–1939)*).

Second, although the biogenetic law was a flawed attempt to link the evolutionary time frame with individual development, the link is still being pursued at the beginning of the twenty-first century, albeit in very different ways, including, for example, through ethology and life history theory. Ethology is the study of animal behavior, its causes, development, functions, and, importantly, its evolution. Developmentalists have drawn on ethological research to provide principles whose applicability to humans can then be assessed (Hinde 1982). This approach is vividly illustrated by Bowlby's attachment theory, which relies heavily on ethological research (e.g., animal im-

printing). In *Attachment and Loss* (1969–80), Bowlby explains that the attachment behavior of young children, which includes variations in proximity seeking to the primary caregiver, is an instinctual behavior observed in many birds and mammals because it facilitates survival (but see Belsky 1999). According to Bowlby, the attachment relationship in humans between child and primary caregiver leads to an ‘internal working model’ of felt security, which in turn has pervasive implications for one’s later relationships and psychological development (see *Attachment Theory: Psychological*).

Life-history theory refers to evolutionary explanations for the specific timing of biological events during the life course. Life-history theory assumes that natural selection has influenced when and how energy is allocated to growth, maintenance of the self, and reproduction so as to maximize reproductive output. Hence, life-history theory attempts to explain, for example the age of maturation, fertility rates and the spacing of children, menopause, senescence, and the length of the life span. For example, ‘disposable soma theory’ provides an evolutionary explanation for why people senesce (i.e., experience a decrease in bodily functions with increasing age). According to this theory, death from unavoidable causes (e.g., accidents, predation) in the course of evolution was inevitable and thus energy was diverted from attempts to live for exceedingly long periods to reproductive efforts earlier in the life span. Indeed, empirical evidence supports the proposition that the rate of senescence is directly related to the probability of death due to difficult-to-avoid causes. One of the major challenges to life-history theory, however, is its application beyond nonhuman species and primitive peoples. Many of its expectations refer to environments that were typical of the evolutionary past and so require further elaboration or restriction when applied to modern societies, especially societies that have experienced the demographic transition (Hill and Kaplan 1999; see also *Education and Learning: Lifespan Perspectives; Psychological Development: Ethological and Evolutionary Approaches*).

1.2 Historical Periods: Social History and Psychohistory

Compared to evolutionary epochs, historical periods are more limited in both time and place. Historical periods typically refer to intervals in history that represent relatively cohesive and distinct patterns of material living conditions, ideologies, norms, social organizations, and institutions.

Social history is often concerned with how these factors jointly shape biography, especially factors such as childhood socialization, school, work, and family. For example, many historians have studied how changing social circumstances have altered the mean-

ing of ‘childhood,’ ‘adolescence,’ and indeed the functions of the family before, during, and after the Early Modern Period, which is often defined in terms of the Industrial Revolution. Ariès’s path-setting *Centuries of Childhood* (1962) provided the impetus for many contemporary inquiries. According to Ariès, in medieval society (and much later among the lower classes), childhood and adolescence did not exist; rather, ‘as soon as the child could live without the constant solicitude of his mother, his nanny or his cradle-rocker, he belonged to adult society’ (p. 128).

Beginning in the sixteenth century, parents began to educate their children before they could enter the adult world. As educational systems developed, youth were taught in age-segregated settings according to curricula that were increasingly less concerned with vocational training than the medieval guilds had been. With this prolongation of education and segregation from the adult world, childhood and adolescence emerged as distinct age-graded identities. While *Centuries of Childhood* has been criticized in matters large and small, its basic analytic orientation—one that emphasizes the interconnectedness of work, school, family, and the phases of the life course—remains a point of departure for historians interested in human development (e.g., Kett 1977).

Psychohistory emphasizes the psychological development of the individual. Its primary focus is on the complex relationships between historically unique processes and the psychological functioning of individuals or, in some cases, mass psychology. Numerous biographies that utilize psychological data or theories have followed Freud’s early attempts. Most draw on some form of psychoanalytic theory, although there are notable exceptions (e.g., B. F. Skinner’s application of behavioral theory to his autobiography).

Erikson’s *Young Man Luther* ([1958] 1993) is often acknowledged as one of the great contributions to psychohistory. He focuses on the adolescent identity crisis because, according to his general theory of the life span, the resolution of this crisis ideally leads to the rejuvenation of the person and, potentially, to the rejuvenation of society. Erikson claimed that Luther’s crisis consisted of an intense Oedipus complex that prompted Luther to question the moral authority invested in father figures, be they his own father or the Pope. Both his father and the Catholic Church attempted to control Luther with a ‘negative external conscience’ since both had contributed significantly to his superego, which emphasized his inherently flawed nature: In the case of his father, Luther did not find support in his choice of occupations, and, in the case of the Church, he felt sinful and unworthy of his priesthood. According to Erikson, Luther’s uniquely strong sense of conscientiousness led to an affirmation of his inner voice as a valid instrument of both self-determination and religious belief, thereby justifying salvation based on faith rather than on obedience to an ecclesiastical (paternal) authority. ‘In laying the

foundation for a religiosity for the adult man,' Erikson argued, Luther 'displaced the attributes of his own hard-won adulthood; his renaissance of faith portrays a vigorous recovery of his own ego-initiative.' (Erikson [1958] 1993 p. 206).

Erikson's biographical approach is unusual for its reciprocal view of history and human development: historical forces clearly shaped Luther and yet, in resolving his identity crisis, Luther in turn shaped history. There are, however, several basic difficulties common to analyses such as Erikson's, including, for example, a lack of pertinent biographical information about the subject's childhood, difficulty establishing causal links between early and later events in a person's life, and identifying why one person—and not another among many people in similar circumstances—becomes a historically great figure (Runyan 1982; see also Erikson, *Erik Homburger (1902–94)*).

1.3 The 'Historical Present': The Life Span and Life Course

Social historians and psychohistorians emphasize common elements of context that people share in a historical period. Yet within historical periods, unique trends and events occur that differentiate age groups and even individuals within age groups. Life-span psychology and life course sociology represent two approaches to the historical present and development.

Life-span developmental psychology refers to a series of interrelated concepts and propositions that emphasize the complexity of development in historically changing contexts. According to this perspective, age-graded development is markedly influenced by existing sociocultural conditions and by social change. Indeed, development can be viewed as the interactions among age-graded, historical, and nonnormative (or idiosyncratic) influences (see Baltes et al. 1998). The life-span paradigm has been especially influential in charting connections between age trends in psychological phenomenon and historical time.

For example, Nesselroade and Baltes (1974) examined the relative contributions of age and history to the development of adolescent personality and ability between 1970 and 1972 in the USA. They reasoned that because youth are highly influenced by social circumstance and because personality reflects situational factors, trends in variables such as adolescent ego strength, independence, and ability may be explained by (a) unique historical experiences shared by young people born in the same year (a birth cohort effect) and (b) historical experiences that exert a similar influence across many age groups (a period effect). To examine these possibilities, they studied a sample of adolescents of diverse ages (ranging from 13 to 18) and, within age groups, adolescents from different birth cohorts (e.g., 15 year olds born in 1955, 1956, and 1957).

Empirical analyses reveal significant declines in self-rated ego strength and achievement and a significant increase in independence between 1970 and 1972, irrespective of age, which suggests that the development of adolescent personality and ability are less dependent on age than on historical setting. According to Nesselroade and Baltes, the observed effects may reflect several historical features that characterized the early 1970s and affected many adolescents (i.e., a period effect), including the tendency of youth to value moral, social, and political questioning (especially of the Vietnam war) rather than to value cognitive achievement, and the widespread decline in confidence in public and educational leadership. This research and related studies suggest that psychological development does not proceed as 'an orderly, sequential unfolding of universal, stage-like behavior patterns' (Nesselroade and Baltes 1974, p. 57), but rather is shaped in significant ways by historically based experiences (see also Schaie 1996, for notable cohort and period effects on adult cognitive development).

The life course refers to the sequence of age-graded roles that typically characterize development. These roles (e.g., student, worker, mother) bring with them expectations, opportunities, and limitations, all of which shape the contours of people's lives. This sequence of roles, their timing, and content are sensitive to immediate historical circumstances. Thus, the study of the life course typically begins with social changes and traces the implications of these changes to the developing person (see *Life Course in History*).

Drawing on data from a wide variety of sources, for example, Modell (1989) shows that the 'normative' sequence of roles often associated with the transition to adulthood—the completion of school, entry into work, marriage, and becoming a parent—became less common through the twentieth century. Modell juxtaposed his quantitative findings with the qualitative study of US cultural practices (dating, for example) to suggest that as youth were freed from the constraints of family and as they acquired financial resources of their own by wage labor, they increasingly became 'shapers of youth,' making choices about when and how they would enter adulthood. The aggregate of these choices creates a 'demonstration effect' for subsequent cohorts, until the next demonstration effect comes along to replace it.

The study of the life course owes much to Elder's watershed *Children of the Great Depression* ([1974] 1999), which integrated a focus on historical and social change with the study of the family and the social and psychological development of the person. A key analytic tool in Elder's research is the birth cohort (i.e., the year a person is born), which situates the person's development in the context of historical events. For example, Elder compared the lives of older children (ages 11–13 in the depths of the Depression) and younger children (ages 3–5 in the depths of the Depression) during and after this event. He further

distinguished between families who did and did not experience serious financial burdens because of economic downturn. Although the Depression led to similar changes in the deprived families of both cohorts—involving the division of labor, altered family roles, and social strains—its developmental effects varied by the age of the children. For example, children in the younger cohort from deprived families were more vulnerable to family strains and indeed were less likely to be hopeful, self-directed, or confident about their future than were their counterparts in families not adversely affected by the Depression. In contrast, the older boys from deprived families were likely to take active steps to help their families and themselves; these experiences in turn enhanced their social independence, which had lasting effects on their lives (see Elder et al. 1993).

Elder's research established a general analytic model that interrelates economic downturn and individual development. The basic model posits that economic stressors lead to marital tensions, which in turn have negative implications for parenting and children's well-being. Although the model was developed in the context of the Great Depression, it has proven useful in understanding other economic downturns (e.g., the Farm Crisis of the 1980s). Elder's research also established a series of analytic principles for the study of change and human development across historical circumstances. For example, the life-stage principle asserts that the influence of a historical event depends on the age at which the person experiences it. This principle is illustrated by the differences observed among the older and younger children of the Great Depression, but it has also been used, for example, to study the effects of military service, German reunification, and China's Cultural Revolution on people's lives.

2. Historical Data and Theories of Human Development

While the theories and empirical research discussed in the previous section emphasize how particular historical forces lead to specific patterns of development, some scientists use historical data to test general ideas about human behavior. This approach has yielded valuable insights, for example, about the factors that affect physical height (Fogel et al. 1983), human creativity (Simonton 1999), and the timing of puberty (Mitterauer 1993).

One of the landmark contributions in the use of historical data to test general psychological principles is Sulloway's *Born to Rebel* (1996). Sulloway developed an elaborate theory that interrelates birth order, family dynamics, and aspects of personality. For example, his theory predicts that later-borns are more likely to subscribe to radical ideas than first-borns. He tested

this and other expectations by drawing on a comprehensive database of 6,566 participants in famous scientific, religious, and political movements in history. A noteworthy feature of this research is the use of experts to rate the participants in these events in terms of their personality and social attitudes.

Drawing on these unique data, Sulloway found support for his theory. For example, the Copernican revolution and Darwin's theory of evolution by natural selection are both rated as radical scientific theories and in fact the leaders of both were likely to be later-borns, including Copernicus and Darwin themselves. (Alfred Russell Wallace, sometimes considered the co-originator of evolution by natural selection, was also a later-born.) In contrast, scientific breakthroughs of a technical nature are likely to be achieved by first-borns as found, for example, in the cases of Watson and Crick, the cofounders of the structure of DNA (as well as in the case of their chief rival, Linus Pauling, also a first-born). Sulloway's research is exemplary for its attempt to document person-by-situation interactions involving the historical context, such as birth order and the social implications of a scientific revolution.

3. Integrating History and Development

The preceding discussion suggests two directives for future research. First, accounts of human development will ideally incorporate explanatory factors from all three historical time frames. The case of attachment illustrates the potential of this approach. In terms of epochs, ethological research has contributed to an understanding of the instinctual basis of attachment behaviors. In terms of historical periods, the parent-child relationship is heavily conditioned by social institutions such as education, work, and family. And, in the historical present, parental loss or changes in the parent-child relationship are often precipitated by immediate circumstances of historical import. Indeed, attachment theory was inspired in part by the countless children left without parents after World War II.

Similarly, Sulloway's work on sibling strategies and family dynamics unites evolutionary theories about sibling competition with a consideration of historical practices (such as primogeniture), and finally with statistical patterns associated with individual development.

Second, interdisciplinary collaborations will benefit from the integration of diverse methods of inquiry. Many developmentalists rely on statistical analyses of data that were recently collected; these accounts typically emphasize how contextual and personal factors predict behavior. In contrast, many historians rely on the qualitative study of archival materials; these accounts typically emphasize the complex combinations of structural contingencies and cultural

meanings that shape biography. If developmentalists often ignore the interactionist, cultural stance of historians, historians are often skeptical as to whether their interests can be measured and subjected to statistical analyses. Yet a growing number of studies illustrate the value of integrating diverse methods, including, for example, Modell's (1989) study of the transition to adulthood and Elder's study of the Great Depression (Elder et al. 1993; for another example see Hareven 1982).

In short, as research on attachment and family dynamics illustrates, all three historical time frames come together to shape one's life, and thus render the study of human development an interdisciplinary task that encompasses diverse historical influences. And as research on the transition to adulthood and the Great Depression illustrates, the quantitative and qualitative methods of the behavioral sciences can be used in complementary ways to uncover these many historical influences as they shape development and the biography.

See also: Human Development, Bioecological Theory of; Life Course in History; Life Course: Sociological Aspects; Lifespan Development, Theory of

Bibliography

- Ariès P 1962 *Centuries of Childhood: A Social History of Family Life*. Vintage Books, New York
- Baltes P B, Lindenberger U, Staudinger U M 1998 Life-span theory in developmental psychology. In: Damon W (ed.) *Handbook of Child Psychology*. Wiley, New York, Vol. 1, pp. 1029–43
- Belsky J 1999 Modern evolutionary theory and patterns of attachment. In: Cassidy J, Shaver P R (eds.) *Handbook of Attachment*. Guilford, New York pp. 141–61
- Bowlby J 1969–80 *Attachment and Loss*, 3 vols. Basic Books, New York
- Elder G H Jr [1974] 1999 *Children of the Great Depression: Social Change in Life Experience*. University of Chicago Press, Chicago
- Elder G H Jr, Modell J, Parke R (eds.) 1993 *Children in Time and Place: Developmental and Historical Insights*. Cambridge University Press, New York
- Erikson E [1958] 1993 *Young Man Luther: A Study in Psychoanalysis and History*. W W Norton, New York
- Fogel R W, Engerman S L, Floud R, Friedman G, Margo R A, Sokoloff K, Stekel R H, Trussel T J, Villaflor G, Wachter K W 1983 Secular changes in American and British stature and nutrition. *Journal of Interdisciplinary History* 14: 445–81
- Freud S [1913] 1962 *Totem and Taboo*. Norton, New York
- Freud S [1939] 1967 *Moses and Monotheism*. Vintage, New York
- Gottlieb G 1992 *Individual Development and Evolution: The Genesis of Novel Behavior*. Oxford University Press, New York
- Gould S J 1977 *Ontogeny and Phylogeny*. Belknap Press of Harvard University Press, Cambridge, MA
- Hareven T 1982 *Family Time and Industrial Time: The Relationship Between the Family and Work in a New England*

- Industrial Town*. Cambridge University Press, Cambridge, UK
- Hill K, Kaplan H 1999 Life history traits in humans: Theory and empirical studies. *Annual Review of Anthropology* 28: 397–430
- Hinde R A 1982 *Ethology: Its Nature and Relations with Other Sciences*. Oxford University Press, New York
- Kett J F 1977 *Rites of Passage: Adolescence in American, 1790 to the Present*. Basic Books, New York
- Mitterauer M 1993 *A History of Youth*. Blackwell, Oxford, UK
- Modell J 1989 *Into One's Own: From Youth to Adulthood in the United States, 1920–1975*. University of California Press, Berkeley, CA
- Nesselroade J R, Baltes P B 1974 Adolescent personality development and historical change: 1970–1972. *Monographs of the Society for Research in Child Development* 154 (vol. 39) no. 1
- Runyan W M 1982 *Life Histories and Psychobiography: Explorations in Theory and Method*. Oxford University Press, New York
- Schaie K W 1996 *Intellectual Development in Adulthood*. Cambridge University Press, New York
- Simonton D K 1999 *Origins of Genius: Darwinian Perspectives on Creativity*. Oxford University Press, New York
- Sulloway F J 1979 *Freud, Biologist of the Mind: Beyond the Psychoanalytic Legend*. Basic Books, New York
- Sulloway F J 1996 *Born to Rebel: Birth Order, Family Dynamics, and Creative Lives*. Pantheon Books, New York

M. J. Shanahan

Historical Demography

Some of the central questions historical demographers are concerned with may be traced back to the writings of Thomas Robert Malthus. His influential thesis was that human beings are able to reproduce themselves 'in a geometrical ratio' while the capacity of the earth to produce subsistence increases 'only in an arithmetical ratio.' In the long run, therefore, population increase of the past must have been kept controlled either by the 'positive check' of higher mortality or by the 'preventive check' of abstinence or late marriage. However, it was not until the post-World War II period that historical demography was born as a separate discipline in historical science: in the 1950s concepts and techniques of modern demography were applied by French scholars to historical material. Its progress and achievements since then have changed our perceptions on the Malthusian and other questions.

1. A Social Science History

Historical demography is a social science history. It is distinguished from other branches of history by its explicit use of concepts and tools derived from a

discipline in social sciences, i.e., demography in this case. Its goal is to establish historical facts by applying such conceptual frameworks to historical sources, and its methodological novelty lies in quantification. The research results of historical demography as such, have often questioned the conventional pictures of past populations, thereby throwing doubts about assumptions many theories rest upon. The position of historical demography in social and behavioral sciences thus resembles that of New Economic History, but it should be realized that there are a few notable differences in emphasis between these two social science history fields. While the application of theoretically constructed econometric models is given prime importance in New Economic History, emphasis in historical demography has always been on measurement. Attempts to formulate models to explain historical facts have never been put aside, but more effort has been directed to the development of finer measures and the identification of patterns, temporal and cross-cultural, by using those quantitative measures.

2. Sources, Methods, and Impacts

The innovation in methodology often goes with the discovery of new data sources. In historical demography too, its birth was closely tied with the French scholars' application of family reconstitution method, a form of nominal linkage, to registers of baptism, burial, and marriage kept in Christian parish churches. Parish register demography, as one may call this type of research, is an attempt to reconstruct a demographic profile of the past based on a set of family histories reconstituted according to the rules set out by the demographer Louis Henry who turned to historical material to identify what he called 'natural fertility,' that is, fertility before the deliberate effort to reduce family size came to be practised.

Information that historical demography started supplying since then is a kind that they had always thought would only be available from modern vital records or sample surveys. Thanks to laborious works by historical demographers, demographers are now able to describe past populations in much the same vocabulary as for the present. However, there are a few drawbacks. In this kind of research, analysis ought to be restricted to persons whose burial dates are known, or who are known to have survived until their reproductive period was over. Since more people tended to leave the community with advancing age, it is very difficult to compute death probabilities across all the age groups, although a large enough sample, such as the one constructed by E. A. Wrigley and his associates at the Cambridge Group for the History of Population and Social Structure, does enable researchers to estimate age-specific patterns of mortality even from family reconstitution (Wrigley et al. 1997).

Furthermore, demographic research based on historical micro data has enabled the researchers to work with 'cohort' measures. Most demographic concepts, such as life table and marital fertility, are designed to portray the life events of a group of people who were born or married in the same period (termed 'cohort'), but usually it is not possible for present-day demographers to single-out people born in a particular period and to trace them to the moment they die. Instead of using cohort measures, they have to rely on 'period' rates to account for demographic events in a specified period, in which the behavior of one cohort is bound to be mixed with that of other cohorts. In this sense, historical micro data have given demographers an opportunity to work with methodologically ideal populations.

For historians, on the other hand, historical demography has offered them a different set of merits. First, population becomes tractable as 'population change' and can now be decomposed into demographic components such as births, deaths, and migrations. This in turn enables them to relate those demographic phenomena to changes in economic circumstances. For example, one of the results of the earliest family reconstitution studies in France showed us how changes in grain prices were correlated with those in births and deaths. Similarly, one may examine interrelationships between time series of demographic rates and real wages in detail.

Another impact historical demography has had is being felt in social history. Since parish register demography reconstructs families and their life events that occurred in familial contexts, it has offered social historians invariable insights into the demographic domain of family life. Marriage, for example, is no longer a subject tied only with legal and ecclesiastical concerns. Who married whom, at what age, with what age gap between spouses—all these questions can now be answered. In addition, information about pre-nuptial pregnancy, gender differentials in infant and child mortality, and other related topics, can shed light on the intimacy of family life in the past.

Also related, but somewhat different approaches, have been adopted by another group of social historians. Statistical studies of the family and household, launched by Peter Laslett at the Cambridge Group in the 1960s, have been conducted in close association with research in parish register demography (Laslett 1983, Laslett and Wall 1972). The research along this line strongly suggests that types of family system, whether or not a new household was formed on marriage, whether the presence of live-in servants in the household was frequent or not, the mean age at marriage, and the frequency of celibacy, were all closely interrelated.

These links with social history have led to the burgeoning of family and household studies in history. Topics studied thus far include servanthood, gender inequality within the household, care for the elderly,

widows and orphans, and the succession of headship and property. There are also issues that have enabled historians to get further into the inner domain of the individual. Abortion, infanticide, child neglect, and other means of family limitation embedded in traditional reproductive culture, are such issues.

However, source material being used in historical demography and its related research fields is not confined to the parish registers. In fact, there is a serious drawback in the use of parish registers as a source material. Since only people who registered baptism, marriage, and burial in a single community were recorded, it is impossible to reconstruct families who eventually left the community. Also impossible is to know how many people existed in that community at a particular point in time. All this implies that aggregative studies are extremely difficult to conduct, implying that the size of a settlement population, its birth and death rates, and in- and out-migration ratios can hardly be computed. Also difficult to conduct, for the reasons already mentioned, are mortality studies.

It is not surprising, therefore, that in many west European countries other micro data are being used. Especially useful are listings of households and population, most of which were one-off 'censuses' taken for various purposes. In other cultural regions, there are places where census-type records were regularly taken. In traditional Japan, for example, registers of village population were compiled by officials every year while in some parts of imperial China similar records were taken triennially. Such repetitive census-type data are useful not only for household analysis, but also for demographic studies since probabilities of entry or disappearance can be calculated at a regular interval, which, as monographs such as Smith (1977) and Lee and Campbell (1997) illustrate, allow researchers to draw demographic interpretations. One of the few shortcomings of such population registers is that it is very difficult to estimate the infant mortality rate since those born after the previous compilation, but died before the next, were not recorded at all.

3. Two Dimensions: Micro and Macro

Two broad areas of issues have been occupying a central place in the historiography of historical demography. One is the Malthusian models of positive and preventive checks, and the other the thesis of demographic transition. The former concerns pre-modern settings. The latter is about the transition processes from the pre-modern to the modern age, as Demeny (1972) put it: 'In traditional societies, fertility and mortality are high. In modern societies fertility and mortality are low. In between there is the demographic transition. In both fields of research, however, it is useful to distinguish micro-behavioral from macro-aggregative dimensions.'

3.1 Micro Demography

As noted in Sect. 2, parish register demography was born as a fertility-oriented research field. Malthus's notion of positive check posits that overpopulation tended to be checked by a rise in mortality, which postulates that reproductive behavior of couples was governed by the invariable 'passion between the sexes.' The classic theory of demographic transition assumes that pre-transition populations were under 'strong pressure on their members to reproduce,' and hence that marital fertility was uncontrolled and its level close to the physiological maximum. Both Louis Henry, who coined the term 'natural fertility,' and Ansley Coale, who developed a technique to measure a deviation from the natural, uncontrolled fertility pattern in the course of his own project on the European fertility decline, assumed that if the couple has no target size of family and is not concerned with the existing number of children (termed 'parity'), marital fertility is not controlled and hence its level must be invariably high. In other words, the decision made by a couple to end childbearing (termed 'stopping') is assumed to mark the onset of fertility limitation. Much effort, therefore, has been made to measure fertility levels and to identify patterns by examining age, parity, birth interval, and other factors associated with individual couples' reproductive behavior (see Coale and Watkins 1986).

However, according to a large body of empirical studies so far conducted the original assumptions about reproductive behavior prior to the demographic transition are no longer tenable (see Cleland and Wilson 1987, Friedlander et al. 1999). First, pre-transition of marital fertility was far more variable than previously imagined. Few populations exhibited levels comparable to that exemplified by the Hutterites, a small group of anabaptist settlers in north America, whose total marital fertility rate (TMFR, defined as the number of children a married woman would have within her reproductive period) stood as high as 11. In most cases, in fact, it was kept at moderate levels with their TMFR ranging from 6 to 9.

Second, it is confirmed that stopping behavior was never common among pre-modern populations. It should be noted, however, that some elite sub-populations, such as Catholic aristocrats in France and Italy and bourgeois families in Calvinist Geneva, exhibited a distinct stopping behavior, so that they can be regarded as forerunners of modern family limitation. Third, moreover, the couple's behavior to 'space' births at particular intervals, as contrasted with stopping, seems to have been more important in accounting for the moderate levels of marital fertility in traditional times. Unlike stopping, birth spacing could have been achieved either through deliberate means such as contraception, or through unconscious actions such as breastfeeding, and many research

results suggest that the role played by differences in feeding practices of the infant was unmistakable. Fourth, it is now widely recognized that in many instances marital fertility was rising, rather than falling, in the period immediately prior to the onset of modern fertility decline. This was the case in the UK and other European populations as well as in twentieth-century developing countries (Knodel 1988, Wrigley 1998). One of the factors suggested as having affected this rising tendency is a reduction in the level of breastfeeding, which was often associated with early phases of urbanization. Another is a decline in the incidence of stillbirths, which has recently been suggested for eighteenth-century England. The latter is a promising area of micro-demographic study, because even from family reconstitution, it is possible to investigate into the interplay between fertility and mortality measures at very early phases of life.

The conclusion that pre-transition fertility was not uncontrolled, be it consciously or unconsciously, demolishes the conventional divide between the modern and the pre-modern regime. Furthermore, findings from micro-demographic data analysis have made it difficult to see the demographic transition in a dichotomous framework. To what extent the transition was a behavioral revolution, still remains debatable.

3.2 Macro Demography

The concept 'reproduction' can also be applied to a population as a whole. It means the way in which the population replaces itself through natural processes of birth and death. The overall level of such replacement may be indicated by the total number of children that an average woman will have during her childbearing period. This is called the total fertility rate (TFR). Since it does not take mortality into account, the population's natural process should be measured in terms of both fertility and mortality such as life expectancy at birth (e_0). Thus one may describe historical populations by referring to the two, TFR and e_0 . According to Livi-Bacci (1997), for example, Paleolithic populations and the Russians in 1879 happened to roughly exhibit the same level of TFR, a little lower than 6.5 children per woman. But mortality was different. The life expectancy is estimated to have been at below 20 for the Paleolithic, whereas that for Russia was a little below 30. As a result, the difference in the rate of natural increase was considerable. It was only marginally above zero in the Paleolithic case whereas, more than 10,000 years later, the Russian population was growing at somewhat over 1 percent per annum.

This, however, should not be taken to imply that the driving force in population history was always mortality decline, with fertility remaining high throughout the long, long period before the modern age. Indeed,

over time both fertility and mortality may have exhibited substantial volatility. Neolithic settlers are supposed to have been characterized by fertility and mortality higher than Paleolithic hunters and gatherers, reflecting a combination of higher population density and lower mobility in the farming population. A much recent case in point is England from 1550 to 1800. Wrigley and Schofield (1981) have documented that from the late sixteenth century, when England's rate of natural increase stood at below 1 percent per annum, the country's TFR and e_0 simultaneously declined towards the late seventeenth century. From that period to the second half of the eighteenth century the two measures came back to a position of higher fertility and higher life expectancy again. This time, however, the speed of rise in fertility exceeded that in life expectancy, so that the country's population grew faster than 1 percent.

Since information from family reconstitution studies does not add up to macro, aggregate measures, complex and sophisticated estimation techniques, such as back projection, are needed to reconstruct the history of a past population at large. Judging from a few well-grounded back projection works, such as Wrigley and Schofield (1981) for England from the sixteenth to the nineteenth century and Weir (1994) and Bonneuil (1997) for eighteenth- and nineteenth-century France, first, there is now agreement that the role of nuptiality should be stressed in accounting for how overall fertility changed in pre-transition western Europe. The age at marriage was instrumental in adjusting the actual level of fertility in relation to changing economic circumstances. This what the classic Malthusian model of the preventive check predicted, thus suggesting that such a homeostatic mechanism was at work in the long- to medium-run time frame of the European past.

Second, this does not imply that mortality crises were not important. Short-term fluctuations in mortality were frequent in various regions in Europe, most of which were subsistence crises occasioned by crop failure. The frequency of famines, therefore, could account for the long-run level of death rates, but it should also be noted that changing levels of 'exposure' to viruses and other micro-organs were equally—sometimes more—responsible in determining that level of mortality in the past. The supposed link between mortality and the general standard of material life was thus less direct than often imagined (see papers in Bengtsson and Saito 2000).

It is true that during the period of fertility decline, neither nuptiality, nor mortality, could offset the declining trend of marital fertility and, hence, the corresponding rise in population. This, however, should not be taken to imply that economic growth, which accompanied this demographic transition in most cases, changed the whole relationship between the demographic variables. Neither industrialization, nor urbanization, was the crucial determinant of the

onset of fertility decline. Even the modern decline of mortality was not in close association with an increase in the general standard of living.

Finally, mention must be made to the role of migration. The basic demographic equation tells us that a population will change according to the addition of births and immigrants, and the subtraction of deaths and emigrants. In other words, without information about in- and out-migration it is impossible to account for any macro change. Migration, therefore, is as important a topic as nuptiality, fertility, and mortality in historical demography.

Unfortunately, however, as historical sources rarely recorded actual population movements, not much has been done on this important topic. The historiography reveals that the volumes of in- and out-migration, distance, pull and push mechanisms are issues relatively better covered thus far. Much less known is the demographic impact of such migration flows. Were death probabilities in areas migrants headed for, different from those in their native communities? Was their marriage and reproductive behavior altered by their geographical mobility? Or was even their parents' demographic decision-making affected by the prospect that their children would eventually leave for a place of, presumably, better opportunities? It is only in cases where no effect of these was substantial, that we do not need to know anything other than the rate of net migration (immigrants minus emigrants), which tended to be rather small. More specifically, any interpretation of results from family reconstitution studies rests upon the assumption that 'leavers' would behave just as 'stayers' did. All the demographic measures derived from this method only relate to the population of stayers, i.e., those born locally who continued to live in the their native community. The mean age at marriage, for example, could have differed between these two populations. Although it turns out that the difference was minimal in the case of English family reconstitution, the possible effects of migration on marriage as well as the other demographic variables can only be determined empirically (Wrigley et al. 1997, Chap. 5).

One of the few findings we have in this regard is about rural-to-urban migration. In a traditional setting the urban population was distinguished by a greater exposure to diseases and, hence, higher mortality than in the countryside. Obviously, even in such a mortality-dominant urban demography, the marriage market may also have been affected by the size of in-migration as well as that of labor demand. Moreover, the rural sector should not be considered demographically homogenous. There are European case studies suggesting that 'proto-industrial' villages exhibited an earlier age at marriage and higher fertility rates than farming villages. To what extent this was universal is debatable. What is certain, however, is that the destination and volume of each migration stream between the rural and urban sectors could

affect the overall rate of population change in a pre-industrial setting (de Vries 1984).

The picture becomes less clear for the period of modern fertility transition, which in many cases coincided with extremely large volumes of migration streams, both long- and short-distance movements. It is probably safe to suppose that rural-to-urban migration in this period also meant deterioration in the disease environment, while trans-oceanic migration to land-abundant settlements led to lower age at marriage and higher fertility. What is less clear is its influence on the population sending those migrants in the course of change. In the literature of demographic transition, migration, especially townward migration, tends to be regarded as having facilitated the decline of fertility. On the other hand, it is often noted that rural populations did not readily respond to changing economic and social circumstances by lowering their fertility. One suggested reason why it was still beneficial for peasant families to keep their fertility relatively high is the expanding prospects for the younger generation. When it was felt that they could emigrate and form their own households elsewhere, the older generation's immediate need for the limitation of fertility may have been reduced. In nineteenth-century Scotland, for example, people left when circumstances allowed, and the same set of factors accounting for the emigration forced stayers to postpone marriage. Once married, however, they were less pressurized to limit the family size (see Andersson and Morse 1993, Friedlander et al. 1999).

4. *The Demographic Landscape in Cross-Cultural Perspective*

At a more generalized level, historical demographers have developed a notion of homeostatic demographic regime. Many societies in the past had an ability to maintain equilibrium between their resources and the demographic characteristics. It is the ability to restore the original state of equilibrium, even if an external shock forces the system away from that position. When the ways in which this homeostasis is achieved are similar for a group of populations, it is probably appropriate to call that particular mode of inter-relationships between the demographic variables a 'demographic regime.' Given empirical evidence so far accumulated, therefore, we may see the traditional west European populations as characterized by a particular mode of demographic regime. In this regime, both fertility and mortality levels were relatively low, and the key to its adjusting mechanism was nuptiality acting to ensure the system to restore equilibrium.

Turning to the nuptiality patterns, John Hajnal in his influential articles has shown that in traditional Europe, west of the line from St Petersburg to Trieste,

the mean age at marriage was relatively high and celibacy not negligible, and that all this was closely tied with the nuclear family system and its associated principle of neo-localism (Hajnal 1965, 1983). In contrast, non-European countries exhibited comparatively low ages at marriage and very low proportions never married, suggesting that different demographic regimes were operating in other parts of the world.

Take China and Japan. According to in-depth regional studies in traditional periods, the level of life expectancy at birth (e_0) in northern China was 32, while it stood at 39 in central Japan (see Lee and Wang 1999 for China, Bideau et al. 1997, Chap. 8 and, Liu et al. 2001, Chap. 7 for Japan). Compared with the western European average, China's mortality was on the high side and Japan's was on the low side. On the other hand, TMFR in both countries fell between 6 and 6.5, unmistakably on the low side of the European standard.

The Chinese and Japanese results raise some interesting questions. First, it is clear that the much lower level of marital fertility did not reflect stopping behavior, but spacing, thus confirming much of what the work on Europe has suggested. What is not certain at this stage is whether the observed long birth interval was a mere reflection of low fecundity, or a product of deliberate actions. While the factors accounting for the former vary from breastfeeding to malnutrition, the latter raises further questions: to what extent abortion, infanticide, and child neglect were practised, which of the three was most responsible, and whether it was motivated by despair or guided by the notion of 'planning,' however, rudimentary it would have been. Second, the nuptiality pattern was undoubtedly different from that of western Europe. There was no neo-local principle working. Nor was the age at marriage functioning so as to adjust overall fertility to economic conditions in Asia. Rather, nuptiality appears to have been more closely related to mortality. Indeed, the female mean age at marriage was unmistakably higher in Japan than in China, reflecting the aforementioned difference in e_0 , although both countries seem to have shared the characteristics of universal marriage.

Third, however, if attention is turned to the status of females and family systems, we realize that there were some fundamental differences between China and Japan. For example, while no particular anomalies are observed from Japanese sex-specific life tables, the Chinese studies point out that there were persistent gender inequalities in mortality: the probability of girls dying between ages 1 and 5 was about 20 percent higher than that of boys, and such gender imbalance continued up to age 45. Clearly the status of girls and women was very different in traditional Chinese society. More specifically, it is the family system that differed. Predominant in Japan was the stem family system, under which the family household was vertically structured, and customs such as adoption, marriage, and remarriage were all for the family line to

perpetuate. In China, on the other hand, what was operating was the rules of joint family. Under that system the household was laterally structured and the composition tended to be perennially complex.

To conclude, historical demography as it now stands, is both interdisciplinary and comparative and has raised new questions of its own, as it has evolved over time as well as cross-culturally.

See also: Demography: Twentieth-century History; Families and Households, Behavioral Demography of; Family and Kinship, History of; Family Systems in Europe; History and the Social Sciences; Life Course in History; Quantification in History; Social History

Bibliography

- Andersson M, Morse D J 1993 High fertility, high emigration, low nuptiality: Adjustment processes in Scotland's demographic experience, 1861–1914. Parts I and II. *Population Studies* 47: 5–25, 319–43
- Bengtsson T, Saito O (eds.) 2000 *Population and Economy: From Hunger to Modern Economic Growth*. Oxford University Press, Oxford, UK
- Bideau A, Desjardins B, Perez-Brignoli H (eds.) 1997 *Infant and Child Mortality in the Past*. Oxford University Press, Oxford, UK
- Bonneuil N 1997 *Transformation of the French Demographic Landscape 1806–1906*. Oxford University Press, Oxford, UK
- Cleland J, Wilson C 1987 Demand theories of the fertility transition: An iconoclastic view. *Population Studies* 41: 5–30
- Coale A J, Watkins S C (eds.) 1986 *The Decline of Fertility in Europe*. Princeton University Press, Princeton, NJ
- Demeny P 1972 Early fertility decline in Austria-Hungary: A lesson in demographic transition. In: Glass D V, Revelle R (eds.) *Population and Social Change*. Arnold, London
- de Vries J 1984 *European Urbanization 1500–1800*. Methuen, London
- Friedlander D, Okun B S, Segal S 1999 The demographic transition then and now: Processes, perspectives, and analyses. *Journal of Family History* 24: 493–533
- Hajnal J 1965 European marriage patterns in perspective. In: Glass D V, Eversley D E C (eds.) *Population in History*. Arnold, London
- Hajnal J 1983 Two kinds of pre-industrial household formation system. In: Wall R, Robin J, Laslett P (eds.) *Family Forms in Historic Europe*. Cambridge University Press, Cambridge, UK
- Knodel J E 1988 *Demographic Behavior in the Past: A Study of Fourteen German Village Populations in the Eighteenth and Nineteenth Centuries*. Cambridge University Press, Cambridge, UK
- Laslett P 1983 *The World We Have Lost—Further Explored*. Methuen, London
- Laslett P, Wall R (eds.) 1972 *Household and Family in Past Time*. Cambridge University Press, Cambridge, UK
- Lee J Z, Campbell C D 1997 *Fate and Fortune in Rural China: Social Organization and Population Behavior in Liaoning 1774–1873*. Cambridge University Press, Cambridge, UK
- Lee J Z, Wang Feng 1999 *One Quarter of Humanity: Malthusian Mythology and Chinese Realities*. Harvard University Press, Cambridge, MA

- Liu Ts'ui-jung, Lee J, Reher D, Saito O, Wang Feng (eds.) 2001 *Asian Population History*. Oxford University Press, Oxford, UK
- Livi-Bacci M 1997 *A Concise History of World Population*, 2nd edn. Blackwell, Cambridge, UK
- Smith T C 1977 *Nakahara: Family Farming and Population in a Japanese Village, 1717–1830*. Stanford University Press, Stanford, CA
- Weir D R 1994 New estimates of nuptiality and marital fertility in France, 1740–1911. *Population Studies* 48: 307–31
- Wrigley E A 1998 Explaining the rise in marital fertility in England in the 'long' eighteenth century. *Economic History Review* 2nd ser. 51: 435–64
- Wrigley E A, Schofield R S 1981 *The Population History of England, 1541–1871: A Reconstruction*. Arnold, London
- Wrigley E A, Davies R S, Oeppen J E, Schofield R S 1997 *English Population History from Family Reconstitution 1580–1837*. Cambridge University Press, Cambridge, UK

O. Saito

Historical Explanation, Theories of: Philosophical Aspects

There are three fairly distinct kinds of theory about the nature of historical explanations. First, there are reflections upon accounts people have given of basic processes of historical change which determine the general direction of history. These accounts are what J.-F. Lyotard calls 'the grand narratives' or 'meta-narratives' of history. In Western culture, these have generally been theories of historical progress and theories of cyclical change. Such theories have been proposed by theologians, political theorists, and philosophers, but seldom by professional historians, who are understandably wary of generalizations of such immense scope. In recent years both the truth and the point of such explanatory theories have been queried, and as a result such theories of world history are now seldom produced. Some historians still offer quite general explanatory interpretations of national and international history, but these too are often held to be partial, in both senses. For convenience historical explanations of this first kind will be referred to as 'general interpretations of historical change.' The closest thing to a theory about such explanations is the discussion by J.-F. Lyotard, so his writing will be used as a platform for some remarks about them.

Second, there are four carefully articulated theories of how historians should explain events. These are written by philosophers rather than historians. Some think historical explanations should be scientific, employing general laws. Some declare that history is not a scientific but a hermeneutic discipline looking for the meaning of events rather than their causes. A few favor comprehensive, many-layered structural history, noting how basic geographic and social structures

influence historical change. Finally, there are those who, in a postmodern vein, see history as no more than a form of literature, with historical events acquiring meaning from their place in a discourse. The philosophers who promote these theories of historical explanation would say they are describing historical practice, but in fact they are much less interested in analyzing historical practice than in promoting a form of explanation which they think very important in understanding history. For that reason these will be called 'prescriptive theories of historical explanation.'

The third kind of theory of historical explanation is a theory which claims to be descriptive and analytic, focusing on historical practice to discover the variety of historical explanations historians actually produce, and their criteria of acceptability. The task of this theory is to provide clear analysis of typical historical explanations, pointing out the logical standards which it seems they aim to satisfy.

The article will briefly describe each of these theories of historical explanation, and make one or two comments upon them.

1. General Interpretations of Historical Change

In the culture of Western Europe, there have been several grand theories of historical change. Christianity, more particularly the Bible, brought with it a theory of progress in history, of God establishing his kingdom on earth, first through his prophets, then through the Israelites, then through Jesus, and finally through the Church, until Christ returns in triumph. St. Augustine presented an immensely influential version of this story in *The City of God*. With the progress in science achieved by the application of reason in the seventeenth and eighteenth centuries, a secular theory of historical progress was created by Kant, Condorcet, and others, who saw reason and the methods of science driving human progress towards ever-increasing, this-worldly happiness. Hegel and Marx described the process of human progress as dialectical, contradictions being overcome by new stages of development, either in the advancement of reason, freedom, and self-consciousness in the case of Hegel, or in autonomy and prosperity, in the case of Marx. The optimism implicit in all these theories of progress was largely shattered by the experiences of the twentieth century, of hideous world wars, of vast economic depression, of rampant infectious diseases, of tyranny, greed, and madness defeating enlightenment in so many countries. As Lyotard puts it: Auschwitz invalidates the Hegelian assumption that the real is rational; Budapest 1956 invalidates the Marxist claim that what is communist is proletarian (Lyotard 1987, p. 162).

A few writers have argued that progress has indeed continued throughout the twentieth century. With the defeat of fascism and the disintegration of Soviet

communism, F. Fukuyama sees the emergence of liberal-democratic capitalism as the final stage of historical development, 'The End of History.' He says that it best satisfies the two driving forces in history, the desire for material prosperity and the desire for personal recognition. Tyrannies had to go, as tyrants showed no respect for the common people. In democracies everyone has a say, or at least a vote. It seems obvious, however, that if the vote is not enough to ensure the recognition and respect people desire, then history has not ended after all.

Lyotard has three main objections to the grand theories of historical progress mentioned above. The first has already been stated: all are contradicted by the horrors of the twentieth century. Second, Lyotard objects that these theories have constrained historical writing in an unacceptable way. Under the influence of such theories, historians have written particular histories as instances of the theory of progress they most favored. For example, they depicted political revolutions as steps towards political freedom, and scientific revolutions as evidence of the growth of reason. In this way Lyotard finds metanarratives oppressive, and prefers that people should write their narrative stories as they like (Lyotard 1984).

Finally, Lyotard says that 'all great historical upheavals' such as the French Revolution 'are formless and figureless in historical human nature' (Lyotard 1987, p. 174). The enthusiasm such an upheaval generates is sublime, he says, and incapable of true representation, although we like to suppose that all who participated were motivated by a desire for greater freedom. All that historians can truly describe, according to Lyotard, are individual human actions and responses, 'little narratives' of particular events.

A few comments on Lyotard's ideas seem called for. First, he is right to deny inevitable progress toward a more rational, free, self-conscious, prosperous, and autonomous society. Theories of progress all underestimate the desire and ability of powerful people to exploit others for their own gain. More realistic, it seems, are some cyclical theories of historical change. Aristotle and Polybius produced such theories which were revived during the Renaissance in the writing of Louis Le Roy and Jean Bodin in France, and of Machiavelli and later Vico in Italy. In Aristotle they found a doctrine of the natural succession of political structures from monarchy to aristocracy to democracy and back to monarchy again, each stage established to resolve injustices which had arisen in the previous one. From Polybius they came to see cycles in the growth, flourishing, and decay of states, some attributing decay to states overextending themselves, others seeing it as a result of moral failure among its leaders. However, as Bodin said, the causes of a state's collapse can be varied, including military conquest by its enemies (1965, p. 59). There is no general theory of historical change which is true without exception. Cyclical theories were popularized in the twentieth century by

Manuel Spengler's (1926–8) *The Decline of the West* and Toynbee's (1951–61) *A Study of History*.

Second, while metahistories may have been used to legitimate certain accounts of particular historical events, to the exclusion of others, some generalizations about the motivation for political revolutions and other events seem in order. The leaders of the French Revolution and the pamphlet writers who supported them did legitimate their attack on the monarchy by an appeal to reason, liberty, and human rights. Of course there were other, more self-interested motives at work as well, but that is not to deny the legitimation given to the revolutionary movement by these ideas. Rather than dismiss generalizations about the motives for revolutions and other large-scale events as all mistaken, historians should take care to qualify them so that they are fairly accurate. Generalizations are useful ways of summing up individual stories and describing social movements.

Finally, one of the chief objections to metahistorical explanations of historical change is something that Lyotard does not mention. They frequently assume that a whole series of events were all justified by the same ideals, when in fact this is false. For example, many political events and revolutions have been justified in the name of liberty and human rights, but the ideas of liberty and human rights were not always the same in every case. The rights and liberties sought in the English Revolution, the French Revolution, and the Russian Revolution, for example, differed significantly. Therefore, it is misleading to suggest that they were all legitimated by the same ideals.

This typically historicist objection to metahistory is important, but not entirely decisive. While different particular rights and liberties were sought at different times and places, all can still be called rights and liberties. Indeed, all drew upon traditions of thought about human reason and dignity found in classical and biblical writing, which were applied to the particular situations. It is wrong to suggest that these events all had precisely the same legitimation, but it is also wrong to deny that the values they appealed to were quite unrelated.

2. Prescriptive Theories of Historical Explanation

In the last section we considered some philosophical responses to theories of historical change, that is to theories which identified certain more or less constant sources of historical developments. These theories provided substantive explanations of history. We turn now to theories of historical explanation of a quite different kind. These are theories about the logical structure of explanations of historical events. They are not theories about what produces historical change, but theories about how historians should explain what has happened in the past. These theories are prescriptive, describing how their authors think historical

events should be explained. There are in fact many forms of explanation commonly used by historians, but these theories focus upon just those which their authors judge, for different reasons, to be the most important.

2.1 Theories of Scientific Explanation

In the nineteenth century both John Stuart Mill and the historian Henry Thomas Buckle argued that historians should try to explain historical events as scientists do, by discovering the laws of society which governed their occurrence. With these laws the occurrence of historical events could be deduced, given certain antecedent conditions, just as the occurrence of natural events can be deduced from the laws of nature.

When K. R. Popper came to propose a theory of historical explanation, he quoted Mill's theory with approval, but thought it could be made more precise. 'I suggest,' he wrote, 'that to give a causal explanation of a certain *specific event* means deducing a statement describing this event from two kinds of premises: from some *universal laws*, and from some singular or specific statements which we may call the *specific initial conditions*' (Popper 1961, p. 122). Popper went on to point out that the laws used by historians are familiar, everyday ones of no particular interest. He opposed those historicists who believe there are general laws of historical development; and he opposed those who claim that historical events are unique, and so have nothing general about them whatever. Those who present general laws of history, such as the Marxist theory that all history is of class struggle, are really offering interpretations, Popper said, describing those changes in history which interest them.

The philosopher who most clearly and forcefully presented a deductive model of historical explanation was C. G. Hempel, in his article 'The function of general laws in history.' Drawing upon the tradition of Mill, Popper, and others, he produced the following analysis: The explanation of an event of some specific kind *E*

consists of

- (1) a set of statements asserting the occurrence of certain events $C_1 \dots C_n$ at certain times and places,
 - (2) a set of universal hypotheses, such that
 - (a) the statements of both groups are reasonably well confirmed by empirical evidence,
 - (b) from the two groups of statements the sentence asserting the occurrence of event *E* can be logically deduced.
- (Hempel 1965, p. 232)

The letters *C* and *E* are intended to suggest cause and effect. The aim of historical explanations, he wrote, is that of 'showing that the event in question was not "a matter of chance," but was to be expected in view of certain antecedent or simultaneous conditions' (Hempel 1965, p. 235). He admitted that the

relevant laws are often difficult to formulate precisely, and that often the consequence of the initial conditions occurs only with a degree of probability (Hempel 1965, p. 237). Consequently he called the explanations offered in history incomplete 'explanation sketches' (Hempel 1965, p. 238).

Hempel's formulation of the deductive theory of explanation was subject to much discussion and criticism. The main objections were as follows:

(a) The model does not state conditions *sufficient* to produce a good explanation,

(i) because sometimes events which enable one to predict an outcome are not the events which caused it. For instance, from a sudden fall in a barometer one can predict a storm, but that event does not explain the occurrence of the storm, because the fall in the barometer was not necessary for the storm's occurrence.

(ii) Furthermore, for an explanation to be satisfactory, it must describe the kinds of cause which are of interest to the reader. If a man with a gastric ulcer is suffering from stomach pain after eating parsnips, a doctor would attribute the pain to his gastric ulcer, whereas his wife would attribute it to his eating parsnips. For the doctor, eating parsnips is normal, but having an ulcer is not, whereas the wife is familiar with the man's ulcer and picks out his eating parsnips as the event which made the difference. This objection assumes that the function of causal explanations is to explain a contrast between a given event and another expected, normal event, rather than to show why an event was predictable.

(b) The model does not state conditions *necessary* for a good explanation,

(i) because often all that is required to explain an event is to show that the event was more probable than a contrasting state, which might not be very probable at all. Consider the question: 'Why was the monarchy replaced by a republic in France in 1792?' This question could be asking: why was there a change in constitution and not just a change of government, why was the republic set up in 1792 and not earlier, and why was the establishment of a republic in France more likely than any other kind of constitution, such as a limited monarchy of the sort which resulted from the English Revolution?

(ii) Furthermore, many explanations in history are not intended to show that an event was very probable, or that it was more probable than a contrasting state. For instance, historians often want to know the reasons why people behaved as they did, reasons which may or may not have made their actions very probable.

There is some merit in Hempel's theory that historical explanations are meant to show events to have been very probable. Admittedly, some explanations have different aims. And certainly causal explanations must refer to events which were necessary for the occurrence of the event being explained. Never-

theless, some causal explanations are not contrastive, but are meant to explain what brought about a certain event by pointing out events and circumstances which made its occurrence very probable. However, this can very seldom be done by citing a single law relating cause and effect. Normally historical events are the product of several tendencies at work in a situation, and are explained as an outcome of those. Historians try to make the outcome appear as highly probable as possible, but often that is not very probable, just more probable than anything else in the circumstances.

2.2 Theories of Hermeneutic Explanation: Explaining the Meaning of Actions

While G. Vico and J. G. von Herder had written of the importance of understanding human cultures as complex, almost organic units, the theory of hermeneutic understanding, in opposition to scientific understanding, was first fully applied to history by W. Dilthey. The theory had its origin in the lectures on hermeneutics by F. Schleiermacher, whose biography Dilthey published early in his career (in 1871). Following Schleiermacher, Dilthey said that the way to understand human actions is by using empathy to imagine the experiences and thoughts which moved people to act as they did. We interpret what people do and say, he said, by drawing upon our own experiences of life to imagine what was going on in their minds to make them produce the expression which is being interpreted. After he had developed this psychological theory of understanding, Dilthey became interested in the meaning human actions have in furthering the aims of the social or cultural structures to which they contribute. He wrote:

To start with, the parts of a life have a meaning according to their relation to that life, its values and purposes, and according to the place they occupy in it. Furthermore, historical events become significant through being links in a system of interactions in which they co-operate with other parts to bring about values and purposes in the whole. (Dilthey 1961, p. 148)

The structures he mentions as being 'the simplest and most homogeneous' are 'education, economic life, law, politics, religions, social life, art, philosophy, science' (Dilthey 1961, p. 146). These different structures often express similar values and beliefs, and so are interconnected. These common ideas can create a common tendency in a particular culture for an age. People are constantly improving the structure and effectiveness of these organizations, and their actions can be interpreted as having such historical significance.

The main objection to Dilthey's theory of historical understanding was clearly stated, and addressed, by Gadamer in *Truth and Method* (1986). Dilthey had said that we understand other people's actions on the

basis of our general knowledge of life in the world. Gadamer pointed out that often the historian's world is remote from the world of the people being studied. Dilthey, he said, 'took no account whatsoever of the historical nature of experience' (Gadamer 1986, p. 212). Heidegger explained that understanding things is not something we do after we have apprehended them, but rather it is what we do *in* apprehending them, in reading a text or in recognizing a machine. And our interpretation of things is a function of our own historical nature (Gadamer 1986, pp. 230–2): 'that we study history only insofar as we are ourselves "historical" means that the historicalness of human There-being [*Dasein*] in its expectancy and its forgetting is the condition of our being able to represent the past' (Gadamer 1986, p. 232). Our understanding of what things mean is first of all provided by the traditions in which we have been raised (Gadamer 1986, p. 250).

Once the historicity of historical understanding is fully understood, the question of the objectivity of history becomes quite urgent. Is there any sense in which one's understanding of the past could be called 'objective'? Gadamer has a couple of things to say about this. First, he insists upon coherence as a condition of correct understanding: 'The movement of understanding is constantly from the whole to the part and back to the whole. Our task is to extend in concentric circles the unity of the understood meaning. The harmony of all the details with the whole is the criterion of correct understanding' (Gadamer 1986, p. 259). He remarks that when it is difficult to find a coherent interpretation, then we suspect that we are dealing with a world whose patterns of meaning we do not really understand. This leads to his second important point. Gadamer employs the concept of a 'horizon,' which he describes as 'the range of vision that includes everything that can be seen from a particular vantage point' (Gadamer 1986, p. 269). Historians occupy a particular vantage point in the history of the tradition which formed them. To *understand* a past event, they must do so from the point of view of those involved. They can find that historical horizon because they have learned of it from within their own tradition. Then when historians come to *report* their understanding of the past, they do so in terms of their present worldview. Gadamer explains: 'In the process of understanding there takes place a real fusing of horizons, which means that as the historical horizon is projected, it is simultaneously removed' (Gadamer 1986, p. 273). It is but a short step from Gadamer's account of historical understanding to postmodern theories, which describe history as a cultural construct, usually in narrative form, having no particular relation to the real events of the past whatever.

At the same time as German philosophers were reacting to attempts to make historical explanations scientific by developing theories of meaning, British

philosophers were pointing out that historical explanations are typically accounts of the reasons for which people acted as they did. R. G. Collingwood wrote as follows:

For science, the event is discovered by perceiving it, and the further search for its cause is conducted by assigning it to its class and determining the relation between that class and others. For history, the object to be discovered is not the mere event, but the thought expressed in it. To discover that thought is already to understand it. After the historian has ascertained the facts, there is no further process of inquiring into their causes. When he knows what happened, he already knows why it happened. (Collingwood 1961, p. 214)

Collingwood said that knowledge of past minds is arrived at by drawing inferences from observable facts, more particularly by forming and testing hypotheses about what people had in mind when acting as they did.

Collingwood's attack on scientific models of explanation and espousal of rational explanations was given fresh expression by W. H. Dray in his first book *Laws and Explanation in History* (1957). He criticized the 'covering-law models' of Popper and Hempel, and developed an account of rational explanation which was less idealistic and more logical than Collingwood's. He said that in deciding reasons for actions, historians relied on general knowledge of the form: 'When in a situation of type $C_1 \dots C_n$ the thing to do is x ' (Dray 1957, p. 132). Such statements, he said, were 'principles of action' which are universally known and respected. They explain actions by showing that the action was the rational thing to have done, in the circumstances.

Today we would give more respect to Collingwood's account of explanation than to this early statement of Dray's, because people's reasons for acting are not always universally appropriate. Indeed, even from the agent's point of view, they are sometimes not very rational at all. All that matters when explaining the reasons for an action is that they were the reasons for which the agent acted.

Hempel responded to Dray by arguing that reasons only explain actions if we know that people holding such reasons were likely to act on them (Hempel 1963). According to Dray, historical explanations should show that actions were reasonable responses to a situation. Hempel said they must always show that the action was to be expected in the circumstances. I think reasons can cause actions by inclining people to perform them, but they do not always make the actions appear either very rational or predictable.

2.3 Theories of Structural Explanation in History

There are two distinct kinds of structural explanation in history, both closely associated with the work of the French Annales school in the twentieth century.

2.3.1 Braudel's structural theory. The first kind of structural explanation is most famously exemplified by F. Braudel's large history, *The Mediterranean and the Mediterranean World in the Age of Philip II* (1972b). In it, Braudel explains the geographic and social structures in which observable actions occurred, and how those structures limited or promoted certain possibilities. He wrote:

when I think of the individual, I am always inclined to see him imprisoned within a destiny in which he himself has little hand, fixed in a landscape in which the infinite perspectives of the long term stretch into the distance both behind him and before. In historical analysis as I see it, rightly or wrongly, the long run always wins in the end ... it indubitably limits both the freedom of the individual and even the role of chance. (Braudel 1972b, p. 1244)

The constraints which Braudel had in mind were largely geographical and technological, but he included frameworks of belief as well (see Braudel 1972a, p. 18).

Braudel showed how geography and technology limited agricultural production and trade, and so the growth of wealth in the sixteenth century. The size of populations, and their vulnerability to famine and disease, also played a part. In addition to these slowly changing structures of geography and technology, Braudel also described cycles of economic and social change which occurred more frequently. These constituted processes of change within economic, social, and institutional structures which historians could model and describe in a theoretical way. He provided a careful analysis of the cycles of mercantile capitalism between the fourteenth and eighteenth centuries, which formed the background for the history of Philip II in the sixteenth century. As Braudel showed, these structures limited the possibilities of human action, as well as making the effects of certain events very probable. However, Braudel gave no systematic account of the relation between these two kinds of structures and the actions of individuals, which constituted the third level of historical description he identified.

2.3.2 Structures of meaning. The second kind of structural explanation favored by historians of the Annales school, such as Lucien Febvre, Marc Bloch, and more recently by Emmanuel Le Roy Ladurie and many others, is one which explains the meaning of an utterance, a gesture or a symbol by referring to a structure of meanings, a language, of which it is a part. Dilthey had anticipated this theory of meaning in some of his later writings, although he is best known for relating meaning to intention. This French structuralist theory of explanation emerged from work by Saussure in linguistics and Lévi-Strauss in anthropology, which explained the mean-

ing of one part of a language by referring to other parts to which it was systematically related. Historians found such patterns of meaning in social events. As Jonathan Culler has put it:

The cultural meaning of any particular act or object is determined by a whole system of constitutive rules [such as rules of a marriage ceremony or of football]—rules which do not so much regulate behaviour as create the possibility of particular forms of behaviour ... make it possible to marry, to score a goal, to write a poem, to be impolite. It is in this sense that a culture is composed of a set of symbolic systems. (Culler 1973, p. 22)

People are seldom aware of the rules governing the symbols they use, any more than they are aware of the rules of language, but the rules can be discovered by historians and anthropologists studying their behavior. The rules can be implicit or explicit. Historians prefer to study the explicit rules governing behavior, usually found in the discourse by which it is described, because then the meaning given to the behavior is truly historical, and not something merely inferred by the historian.

2.4 Theories of Narrative Explanation

Once historical explanations were thought to display the meaning of historical events, rather than just describe their causes, the way was open to view historical narratives as forms of explanation. This is precisely what Hayden White did in his seminal book *Metahistory* (1973). In the theoretical introduction to that book White describes 'explanation by emplotment,' 'explanation by formal argument,' and 'explanation by ideological implication.' A historical event which is merely listed in a chronicle, he said, has no particular meaning or significance. But once it is located in a story, it is given both meaning and significance, by being related to what preceded and followed it. Following Northrop Frye, White identified four forms of emplotment: Romance, Tragedy, Comedy, and Satire, each with its characteristic structure. Thus, for example, 'Romance is fundamentally a drama of self-identification symbolised by the hero's transcendence of the world of experience, his victory over it, and his final liberation from it.' (White 1973, p. 8). As well as telling stories of different kinds, he said, historians also describe the relations between events in the past in several different ways. Following Stephen C. Pepper, he identified four kinds of 'argument' about the nature of historical events. The Formist depicts them as unique; the Organicist sees particular events as parts of a whole; the Mechanist presents events as instances of general laws at work in history; and the Contextualist describes events as products of tendencies at work in their environment. Finally, narratives can represent certain ideological views of historical change, and give events meaning by locating them within one of those. Following Karl

Mannheim, White picks out four such ideologies: the Conservatives,' who see history as the evolution of existing structures; the Liberals,' who think that reason and goodwill will eventually see that certain ideals are attained; the Radicals,' who think utopian conditions should be achieved by revolutionary means; and finally the Anarchists,' who believe that an ideal state will be achieved by destroying existing structures and leaving people free to act on their own humanity. White found 'affinities' between these three groups of four. For example he said Romantic plots often employed a Formist presentation of events, and embodied an Anarchist ideology.

White correctly points out that explanations of these kinds can be viewed as interpretations. Interpretations are one of several more or less equally well justified accounts of a subject which give meaning and significance to its parts. White did not consider whether one interpretation could be better than another, but F. R. Ankersmit did in his book *Narrative Logic* (1983, Chap. 8), in which he declared that narrative interpretations should aim for maximum scope, not only in the number of events they describe but in the number they illuminate. This seems to be on the right track.

3. Descriptive Theories of Historical Explanation

Hayden White denied that his analysis of historical explanation was exhaustive, but it was far less prescriptive than previous theories, indicating a large range of ways in which historians could account for events in the past. An avowedly descriptive, analytic approach to historical explanations is that provided by McCullagh in *The Truth of History* (1998). McCullagh analyzes forms of explanation common in historical writing to discover both their structure and their conditions of acceptability. He contrasts explanations, which are answers to fairly specific questions guided by the preconceptions implicit in the question, and interpretations, which look for patterns among events with no particular preconception of what that pattern will be. The adequacy of explanations, on this account, is judged by how adequately they answer the questions they are addressing. The questions provide both the logical and conceptual parameters of the required explanation. The questions, he says, can ask *what* happened during a certain event, *how* a certain change took place, *who* a certain person was, *what* a text or symbol meant, *what* the structure of a society or institution was, as well as *why* a certain event occurred.

An interesting outcome of this approach is his analyses of causal explanations in history. McCullagh finds that causes are events which trigger causal powers, or tendencies, and he depicts causal explanations as setting out the active tendencies which, taken together, made probable the event being explained. A complete causal explanation, McCullagh says, will mention all the causes which directly

increased the probability of the outcome, and these could include both reasons for an action and circumstances which made such an action probable. It does not include all events necessary for the outcome, however, as these go back to the beginning of time.

At present there is some uncertainty as to whether the concept of a correct explanation is intelligible, and whether explanations always provide interpretations of the past. The answer depends largely upon our concepts of explanation and interpretation, and there is no doubt that these need further refining.

See also: Explanation, Cognitive Psychology of; Explanation: Conceptions in the Social Sciences; Functional Explanation: Philosophical Aspects; History: Forms of Presentation, Discourses, and Functions; History: Overview; History: Theories and Methods; Narratives and Accounts, in the Social and Behavioral Sciences; Rational Choice Explanation: Philosophical Aspects

Bibliography

- Ankersmit F R 1983 *Narrative Logic. A Semantic Analysis of the Historian's Language*. Martinus Nijhoff, The Hague, The Netherlands
- Braudel F 1966 *La Méditerranée et le Monde Méditerranéen à l'Epoque de Philippe II*, 2nd edn. Librairie Armand Colin, Paris
- Braudel F 1972a History and the social sciences. *Annales: économies, sociétés, civilisations*. In: Burke P (ed.) *Economy and Society in Early Modern Europe. Essays from Annales*. Routledge and Kegan Paul, London
- Braudel F 1972b *The Mediterranean and the Mediterranean World in the Age of Philip II*. Collins, London
- Collingwood R G 1961 *The Idea of History*. Oxford University Press, Oxford, UK
- Culler J 1973 The linguistic basis of structuralism. In: Robey D (ed.) *Structuralism: An Introduction*. Clarendon Press, Oxford, UK
- Dilthey W 1961 *Meaning in History. W. Dilthey's Thoughts on History and Society* [ed. Rickman H P]. George Allen and Unwin, London
- Dosse F 1987 *L'Histoire en miettes: Des 'Annales' à la 'nouvelle histoire'*. Éditions La Découverte, Paris
- Dosse F 1994 *New History in France. The Triumph of the Annales* [trans. Conroy P V, Jr]. University of Illinois Press, Urbana, IL
- Dray W 1957 *Laws and Explanation in History*. Oxford University Press, Oxford, UK
- Dray W 1974 The historical explanation of actions reconsidered. In: Gardiner P (ed.) *The Philosophy of History*. Oxford University Press, Oxford, UK
- Gadamer H-G 1960 *Wahrheit und Methode*. J C B Mohr, Tübingen, Germany
- Gadamer H-G 1986 *Truth and Method*. Crossroad, New York
- Hempel C G 1963 Reasons and covering laws in historical explanation. In: Hook S (ed.) *The Philosophy of History: A Symposium*. New York University Press, New York, pp. 143–63
- Hempel C G 1965 The function of general laws in history. In: Hempel C G (ed.) *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. Free Press, New York

- Lyotard J-F 1979 *La Condition postmoderne: rapport sur le savoir*. Editions de Minuit, Paris
- Lyotard J-F 1984 *The Postmodern Condition: A Report on Knowledge* [trans. Bennington G, Massumi B]. Manchester University Press, Manchester, UK
- Lyotard J-F 1987 The sign of history. In: Attridge D, Bennington G, Young R (eds.) *Post-structuralism and the Question of History*. Cambridge University Press, Cambridge, UK
- Manuel F E 1965 *Shapes of Philosophical History*. Stanford University Press, Stanford, CA
- McCullagh C B 1998 *The Truth of History*. Routledge, London
- Owensby J 1994 *Dilthey and the Narrative of History*. Cornell University Press, Ithaca, NY
- Popper K R 1961 *The Poverty of Historicism*. Routledge and Kegan Paul, London
- Spengler O 1926–8 *The Decline of the West* [trans. Atkinson C F]. Allen and Unwin, London, 2 Vols.
- Toynbee A J 1951–61 *A Study of History*. Oxford University Press, London, 10 Vols.
- White H 1973 *Metahistory: The Historical Imagination in Nineteenth-century Europe*. The Johns Hopkins University Press, Baltimore, MD
- White H 1978 Interpretation in history. In: White H (ed.) *Tropics of Discourse. Essays in Cultural Criticism*. The Johns Hopkins University Press, Baltimore, MD

C. B. McCullagh

Historical Geography

Partly because of its richly diverse and spatially variable traditions, historical geography is not amenable to ready definition. While it is concerned with the study and reconstruction of past geographies and understanding processes of ecological and landscape change, its remit also includes the political, cultural, and economic interactions between past and present. Thus historical geography examines how our interpretations and representations of the past are constructed in contemporary circumstances and shaped by current needs. There are, however, significant differences between societies in the traditions of historical geography and the ways in which it is currently construed, researched, and taught. This article deals less with the history of historical geography than with certain of the prevailing trends in its current status as an academic project. A discussion of the nature of contemporary historical geography is followed by an assessment of three themes, which permit an elaboration of this general context. These are: interconnected historical geographies; differential historical geographies; and the past in the present.

1. The Nature of Historical Geography

Two overlapping perspectives long dominated traditional approaches to historical geography. First was an intensely academic concern, epitomized, for ex-

ample, through the work of H. C. Darby in Britain and Ralph Hall Brown in the USA, with the scholarly reconstruction of past geographies for their own sakes. The second, encapsulated, for instance, in the work of Andrew H. Clark in North America, emphasized vertical themes, the essentially spatial changes and continuities that occur through time and help explain the present. More recently, however, historical geography has been revitalized by the wider concern throughout the social sciences and humanities with the relativity of knowledge. Thus the fundamental assumption informing much current historical geography is that representations of the past are constructed in the present to serve the demands of contemporary societies. The content and methodology of this work, particularly in the anglophone world, also reflects the theoretical shifts in many of the social sciences away from the 'objectivist fallacy' of positivist models of research and knowledge production (Wishart 1997). Conceptualizing the past as a resource for the present also inevitably invokes notions of contestation as various groups compete for 'ownership' of the past. Consequently, much current work in historical geography acknowledges the situated nature of research and the political difficulties of reconciling contested interpretations of the past in multicultural societies.

While this work builds on historical geography's long-standing empirical focus on topics including industrialization, rural settlement and land ownership, agriculture and agrarian change, migration and population patterns, urbanization, and transport networks, it is also moving in new directions. Many contemporary historical geographers, informed by perspectives such as feminism, poststructuralism, antiracism, and postcolonialism, share concerns about questions of power and meaning with other researchers more readily located within the traditional systematics of cultural, economic, political, and social geography. The key themes evident in much recent historical geography include: a more overtly theorized exploration of the material, political, and symbolic dimensions of travel and exploration; imperialism and colonialism; nationhood and state formation; and ideas of nature and environmental change (Graham and Nash 1999, Nash 1999).

The exploration of the philosophy, content, and methodology of historical geography has itself a long tradition (Butlin 1993). This introspection has often been anxious, not least when faced with the positivistic revolution in geography in the late 1960s, a paradigm shift to which historical geography seemed ill equipped. The delineation of the boundaries between historical geography and geographical history has also occasioned much debate, often driven by a perceived need to stake out precise disciplinary territories. For example, Guelke (1997) and Donkin (1997) have expressed their fears at the deconstruction and potential disappearance of historical geography. To them it lacks an academic core, defined particularly through

the ways in which people relate to the earth. Nevertheless, much current historical geography is concerned less with the establishment of precise boundaries within geography, or between geography and history, than in engaging in the arguably more valuable task of creating a body of work that is both geographical and historical. In addition, concerns over the erosion of historical geography by other systematics within the subject, or its dilution through the influences of other disciplines, ignore the interdisciplinary and hybrid nature of both past and present historical geography. Where traditionally it drew on archaeology, economic and social history, ecology, and social theory, it now interacts with feminism, environmental history, post-colonialism, and cultural history.

Thus in his capacity as editor of the *Journal of Historical Geography*, Michael Heffernan (1997, p. 2) argues for an historical geography that is 'a haven of intelligent eclecticism.' He offers 'an alternative and more optimistic prospectus for the future' to those 'who worry that eclecticism breeds conceptual chaos and threatens the intellectual integrity of historical geography as an academic project.' Heffernan sees it as a hybrid discipline open to the widespread contemporary questioning of 'the conventional intellectual categories through which the modern world has been interpreted and conceptualized.'

While this eclecticism is undoubtedly reflected in a more cultural historical geography, that content should be no less concerned with the material shaping of places, spaces, landscapes, and lives, even if the need to ground historical geography in material contexts of power, politics, and economics is sometimes neglected. The past is a political resource and its control is fundamental to issues of hegemony and empowerment in societies. If historical geography is about the interplay between the past and the present, between symbolic and material worlds, it must also be open to the hard-edged realities that, for many people, 'history hurts.'

2. *Interconnected Historical Geographies*

Interconnectedness possesses at least three separate dimensions. First, the history of historical geography itself reflects complex webs of intellectual interconnections, which have produced different outcomes in different places. It is also true, however, that linguistic and intellectual parochialism often limits the potential of these interconnections. Second, historical geography possesses a lengthy tradition of locating regional studies within broader processes operating at macrospatial scales, and of paying attention to both the specificities of the local and their interaction with wider economic, cultural, and political domains and institutional structures. Finally, there are interconnections between places—for example, the modern networks of communication, transport, or trade that

connected Europe and its imperial sphere of influence—but also in the kinds of geographical discourses and practices that have structured the ways in which the relationships between places or regions are understood.

2.1 *Interconnections within Historical Geography*

Ogborn (1999, p. 99) argues that a key dimension to interconnectedness is that of showing how and why issues, ideas, and intellectual interactions ‘cross connect and reshape a kaleidoscopic interdisciplinary terrain.’ Thus cultural and material geographies combine, for example, to produce environmental histories that are also histories of power, colonialism, and imperialism (Wynn 1997). That historical geography is seeking to understand the social or cultural construction of spaces, knowledges, and powers also impacts on the discipline itself, which has been, and remains, a spatially variable process. The contents and methodologies of historical geographies vary both through time and across space, reflecting the current—if transient—trends defining the broader intellectual climates within which they are practiced. Consequently, the subject matter of contemporary historical geography in the anglophone world is dominated by the cultural turn in human geography and by the poststructuralist emphasis on situated knowledges. The early modern and modern periods are preferred to the medieval and the contingent to theories of a past shaped by cyclical processes of historical transformation of varying temporality. Marxian theories of world systems and modes of production that emerged from long-term cycles of history and their periodic crises (e.g., Wallerstein 1974–89) are deeply unfashionable in a discipline that has often embraced Foucault’s dictum that nothing is fundamental with uncritical enthusiasm.

The temporal and spatial variability in historical geography is nowhere better illustrated than in the contrasting perspectives on landscape, historical geography’s oldest and most enduring concept. For example, the evolution of British historical geography has, in many ways, been dominated by efforts to provide alternative directions to the source-driven methodology used by H. C. Darby in his systematic reconstruction of past landscapes. Although open to vertical themes and the cycles of history behind geography, Darby’s methodology invoked an empirical dependence on the inevitably elitist context of cartography and documents, and consequently excluded dimensions of power and culture in the societies and landscapes reclaimed. In some contrast, the French geographer, Paul Vidal de la Blache (1926), emphasized the significance of ordinary people and their environment: to him, the region was not simply a convenient framework, but rather a social reality

(Claval 1984). Vidal’s belief that landscape reflected a harmony between human life and the milieu in which it was lived was also fundamental to the work of the Berkeley geographer, Carl Sauer. For him, the richest source of evidence lay ‘in the material content of the cultural landscape,’ the ‘task of human geography [being] nothing less than comparative study of areally localized cultures’ (Conzen 1993, p. 28).

Vidalian ideas were also crucial to the *géohistoire* of the French *Annales* school, and its concerted attempt to map and explain the complex reality of human life by reference to local and regional studies. They also underpinned the geographical philosophy of Ireland’s most influential geographer, E. Estyn Evans, partly explaining why Irish historical geography evolved along a very different trajectory to that in Britain. This was one in which the meanings of historical place and landscape were seen as fundamental to a people’s genetic and cultural identity, albeit conceptualized in overtly more anthropogeographical terms than through the traits of polyvocality and hybridity that attend identity studies today. As in Sauer’s methodology, the critical evidence was that of the visible and material landscape. Similar ideas also influenced the study of landscape in North American, Scandinavian, and Australasian historical geographies, and clearly retain a contemporary resonance and vibrance, given the hybrid and multi stranded processes that link the discursive and material worlds, which constitute the raw resources of historical geography.

2.2 *Local Particularities and General Processes*

Historical geography has long been concerned with the processes of historical specificity and historical transformation across space and through time. The intellectual attraction of the ‘grand’ structuralist models lies in their identification of general power-related processes, which interact with the specificities of local place to produce an almost infinite chiaroscuro of individual outcomes and experiences. Historical geography becomes the product of this complex hybridity and polyvocality of places and peoples, while the contingent is reconciled—as societies seem to require—with temporal structures that create explanations of change, meanwhile observing a cultural and political need for continuity and social memory. The outcome is spatial and temporal diversity, which historical geography has long sought to reconcile with generality of process through a methodology of regionally based analysis.

This is perhaps most coherently encapsulated in a French *géohistoire* seeking to explain a country defined by diversity in which unity—as elsewhere in Europe—was achieved only by force, civil war, and the creation of ethnic nationalisms. Vidal de la Blache argued that it was this diversity—rooted in the physical en-

vironment—which paradoxically provides France with its identity. The country's unity evolves from a beneficent force of commonality, derived from the ways in which social life combines with—and transcends—its divisions through time. This relationship was encapsulated in the idea of the *pays* (literally an area with its own identity derived, not only from divisions of physical geography, but also from ethnic and linguistic divisions imposed on a region by its history) as the geographical mediation of synthesis and continuity, the product of human interaction with the environment over many centuries.

In his monumental (if unfinished) investigation of the identity of France, Fernand Braudel (1988–90) takes up the same theme. He sees a patchwork France, a jigsaw of regions and *pays* in which the vital thing for every community is to remain other from the next tiny *patrie*. But Braudel, recognizing a similar fragmentation elsewhere, was also critical of Vidal's exceptionalism: 'Yes, France is certainly diverse...it is a diversity which breaks up, divides, and sets one region against another.' But it is not an 'unparalleled' diversity: 'Germany, Italy, Britain, Spain, Poland can all lay claim to diversity' (Braudel 1988–90, p. 669). So too can smaller countries. Pursuing the same symbiosis of physical environment and human society, Evans (1981) believed that even fractious Ulster could seek unity in a diverse past to which all its present inhabitants might subscribe.

Although Braudel and Evans saw the natural world as providing the backdrop to this human diversity, both also recognized the decisive contribution of the economy and social intercourse. Moreover, diversity itself was more apparent because it often reflected what Braudel referred to as *l'histoire événementielle*, short-term episodic history or the history of events. Behind all that, however, lay *l'histoire structurelle*, the long-term changes of *la longue durée*, 'the slow speed of the secular trend, moving scarcely at all' (Braudel 1988–90, p. 678). The real essence of this European conceptualization of diversity lies in the ways in which these different histories, trajectories of time, and social change and geographical particularities have fused in varying ways to produce distinctive places and peoples.

The same theme of region and regionalism—albeit differently expressed and generally lacking the Marxian connotations of *géohistoire*—is also a fundamental feature of North American historical geography. Clark's immensely influential emphasis on historical change was expressed through a methodological approach that was 'explicitly historical, attentive to the economic basis of regions, devoted to field and archive, well integrated with professional history, analytically adventuresome, and book-oriented' (Conzen 1993, p. 61). The outcome, readily apparent in much historical geography in both the USA and Canada, was a heavily empirical emphasis on regional spatial patterns, which often conspicu-

ously failed to ask fundamental questions regarding broader social, political, and economic relationships (Wynn 1993).

2.3 Interconnections Between Places

This final dimension to interconnectedness refers both to the physical linkages between places, but also to the ways in which the relationships between places and regions are understood in geographical discourses and practices. It is concerned with the ways in which technological developments and their exploitation allowed particular countries to achieve at least a transient hegemony in the world system. Hugill (1993) argues that the postmedieval nations that developed and marketed new technologies best were the nations that rose to world power, revolutions in transport and communication technologies actually constituting the moments of transformation from one world economy to another. Globalization is not a recent process, the creation of European imperial empires effectively leading to the integration of large areas of the world into the emerging national space economies of Europe. These global networks and flows of information, capital, people, culture, plants, animals, and objects that linked the world in the past continue to rebound on its present, not least through the persistent patterns of disparity that mock the concept of global free trade.

If imperialism and colonialism were hierarchical, they also represented reciprocal networks of influence and power. European nationalism and the emerging nation-state system were framed within these imperial rivalries and later exported to the former empires. Thus the resistances to imperialism made anti-imperial citizens as well as imperial ones (Ogborn 1997). Moreover, a consideration of the connections between places also involves stressing the ways in which the historical geographies of colonial countries themselves were shaped by colonialism. The commonplace interpretation of the historical geography of colonialism as the imposition of the dominant logic of the colonial political and cultural economy unto non-Western time and territory, requires destabilization (Yeoh 1996). The difficulty in so doing, however, lies in the lack of knowledge about colonized groups, given the asymmetries of the historical record.

One final dimension to the ways in which the interconnections between places help shape geography concerns the linkages between geographical education and the making of citizens. Imperialism and colonialism also entailed political, intellectual and cultural, as well as material, domination. Historical geographers have extended the contention that imperial ideologies remain at the heart of modern Western culture to explore the linkages between the cultures of geography and imperialism. Thus 'knowledges of geography are themselves part of often complex geographies of knowledge' (Ogborn 1997, p. 416). Hence it can be

argued, for example, that both popular and formal geographical education in Britain and notions of Britishness were constructed in direct opposition to imaginative geographies of other places and peoples (Graham and Nash 1999).

3. Differential Historical Geographies

Inevitably, the previous discussion has already touched upon the general argument that despite the interconnected nature of world systems, the processes of social change through time, with which historical geographers are concerned, are still experienced differently by peoples in different social groups and places and articulated in different ways. Moreover, these dissimilar knowledges are articulated at a variety of scales ranging from the local to the global. The first dimension to this issue to be considered here concerns problems of social inclusion and exclusion. Second, however, it is also the case that knowledges themselves are rarely consistent and contemporary historical geography is itself internally differentiated or even contested between different ideological and methodological imperatives.

3.1 Social Inclusion and Exclusion

Narratives of the relationships between places are characteristically defined through processes that privilege particular accounts and peoples while marginalizing or excluding others. Contemporary historical geography recognizes that earlier generations did not sufficiently understand the privileging of Europe in their accounts and is aware of the need to avoid the naturalization of Western models of modernity and development. Again, Conzen (1993) identifies the 'startling silences,' including slavery, the African American experience, and the results of war, in American historical geography. The combined importance given to gender and racism in contemporary social and cultural theory helps explain the polyvocality and hybridity of contemporary historical geography. Just as racism ensured that colonized indigenous peoples were excluded from the narratives of empire, all human geography has traditionally privileged masculine discourses and conceptualized modernity in masculine, middle-class, urban, and Eurocentric terms (Nash 1999). Women—especially rural women—were equated with the authentic and premodern, while, in contrast, the modern was associated with a male-directed logic of rationalization, objectification, and developmental progress. Thus traditionally women have been largely invisible and misrepresented in historical geography, while men have 'monopolized the transmission of history' (Lowenthal 1998, p. 48).

However, it is important not to equate marginalization and exclusion with gender and race alone. A

contemporary historical geography has to be open to the nuances of multiple resistances, and to the enduring importance of class, ethnicity, and nationalism as criteria of identity. The implication of the past in the construction of identity narratives is but one example of the way in which historical geography has become a more pluralist and disputed enterprise since the beginning of the 1970s. There are many different modernities and many diverse dimensions to othering and otherness.

3.2 Historical Geography as a Contested Enterprise

As noted in the discussion regarding the interconnectedness of historical geographies, the result of differentiation may be knowledges that are internally inconsistent and contested between opposed ideological and methodological imperatives. Heffernan's 'intelligent eclecticism' does not preclude disputation, inevitably so because of the privileging of particular perspectives within the geographical discipline itself. If a methodologically diverse historical geography traditionally seemed eclectic, its focus on areal differentiation, modernist notions of development and progress along continuities from past to present, and its reliance on empirical, material evidence drawn from the landscape, did link it to many of the principal strands that defined the emergence of geography as a discipline.

In the 1960s, however, the positivist revolution seemed to have little to offer historical geography. Largely lacking the means to meaningfully quantify data and indulge in model building, its practitioners found little resonance in arguments that geography was a formal deductive science charged to describe and understand the real world. The important exception to this generalization was the reversal of historical geographers' traditional aversion to the urban realm, although this was expressed—as the prevailing *zeitgeist* demanded—through a commitment to the morphological and functional relationships of the historic city and to its modeling. Questions of power and social relationships remained peripheral as they had in the regional studies of the rural world.

Given historical geography's vulnerability to this paradigm shift, Harris's (1978) argument that an understanding of the synthesis of regions, places, and landscapes might come from the development of an historical mind that could encompass experience, memory, imagination, and reflection was a welcome harbinger of the revitalization of historical geography by the cultural turn in geography in the late 1980s and 1990s. It is still important, however, to be aware of the dangers of cultural exceptionalism in what remains a polyvocal enterprise. While the cultural theme was dominant in British historical geography and, if to a lesser extent, throughout the anglophone world for much of the 1990s, its content and methodologies do not define historical geography as a global enterprise,

or even describe the research of many historical geographers within the English-speaking world. The privileging of the cultural perspective at the expense of alternative methodologies and languages disguises the survival of enduring issues in historical geography—frontiers and borderlands, migration, settlement morphology, settlement—the city in particular—as an indicator of cyclical social change and social relations, placenames, and the making of the physical environment.

Again, despite its openness to previously disenfranchised peoples and their narratives, the cultural turn in historical geography is often neglectful of power and economy as they control material lives. To some extent, the denigration of the material world is a result of the extension of the source materials used in researching contemporary imaginative historical geographies to encompass a diverse range of written and artistic representations and texts—diaries, biographies, novels, poetry, and paintings. This otherwise welcome development has, however, led to a denigration of traditional sources of data such as the material physical and human landscape. Ironically, the contemporary emphasis on nuanced qualitative readings of texts and representations is paralleled by the previously unimaginable power of Geographical Information Systems and Remote Sensing methodologies to handle and manipulate huge repositories of statistical historical data.

Historical geography now possesses the technological means of far more effectively addressing its traditional responsibilities of reconstructing past environments, spatial analysis, and regional differentiation. Often, however, these tasks have largely been abandoned to a technologically proficient historical archaeology, which combines historical archival research with excavation and geophysical evidence of past societies in both terrestrial and maritime landscapes. More important, historical geography now also possesses the capacity, but has largely shirked the challenge, of producing and managing integrated masses of geographical information, which would substantially enhance understanding of social change through time but also elucidate the regional and local specificities that underpin the constructs of identity through which peoples project and comprehend their relationships with place. Intelligent eclecticism can encompass the deconstruction of the empirical and the abandonment of the objectivist fallacy, but must remain open to using far more elaborate means of processing of historical data, not least to avoid these being subverted by an apparently enhanced capacity for objectivism.

4. The Past in the Present

Finally, historical geography is also increasingly informed by the ways in which the past is remembered and represented in both formal or official senses and

within popular forms, and the implications, which these have for the present. It is in this respect that the past can be most clearly visualized as a domain of contested meanings, perhaps the most significant way in which contemporary historical geography differs from the archival, 'objective' reclamation of the past that characterized its previous existences. Interpretations of the past are seen as being context bound and power laden, an issue briefly explored here through a consideration of the functions of heritage (Graham et al. 2000).

4.1 Heritage and its Functions

The stories and places of contemporary heritage, and the ways in which they are told and described, often represent the most overt examples of the contemporary politics of the past in the present. Heritage is that part of the past selected in the present for contemporary purposes, be they economic, cultural, political, or social. If heritage knowledge is situated in particular social and intellectual circumstances, it is time-specific and thus its meaning(s) can be altered as texts are re-read in changing times, circumstances, and constructs of place and scale. Consequently, such knowledges are inevitably fields of contestation of meaning and function. As such, the past can be visualized as a resource but simultaneously, several times so. Clearly, it is an economic resource, one exploited everywhere as a primary component of strategies to promote tourism, economic development, and rural and urban regeneration. Heritage—variously defined—is the most important single resource for international tourism although the economic commodification of the past can so trivialize it that arguably the result may be the destruction of the heritage resource which is its *raison d'être*.

But heritage is also a knowledge, a cultural product, and a political resource and thus possesses crucial sociopolitical functions. In particular, the past validates the present by conveying an idea of timeless values and unbroken lineages and through restoring lost or subverted values (Lowenthal 1998). Hence there are archetypal national landscapes, which draw heavily on geographical imagery, memory, and myth. Continuously being transformed, these encapsulate distinct home places of 'imagined communities,' comprising people who are bound by cultural—and more explicitly—political networks, all set within a territorial framework that is defined through whichever traditions are currently acceptable, as much as by its geographical boundary. Such national traditions are narratives that are invented and imposed on space, their legitimacy couched in terms of their relationship to particular representations of the past. So important are these that peoples cut off from their past through migration or even by its destruction in war, often recreate it, or even 'recreate' what could—or should have been there—but never actually was. Inevitably,

therefore, the past as rendered through heritage also promotes the burdens of history, the atrocities, errors, and crimes of the past which are called upon to legitimate the atrocities of the present. Consequently heritage is always a duality—a resource of economic and cultural capital. This is less a dialectic than a continuous tension, these broad domains generally being in conflict with each other.

5. Conclusion

A hybrid and interdisciplinary historical geography has thus widened its interests alongside history itself. As historians increasingly explore popular forms of history today as well as cultural sources for the past, historical geographers have also extended what counts as source material in researching contemporary imaginative historical geographies. These include objects, practices, and discourses of popular culture as well as the past itself. If this makes historical geography more cultural, it should not make it any less concerned with the material shaping of places, spaces, landscapes, and lives, or with issues of empowerment and economy. These current intersections reflect the ways in which historical and cultural geography have overlapped in the past—in the study of landscape for example—just as they do in the present, whatever the labels academics chose to define their research interests. The opposition between an empirically grounded historical geography and a purely qualitative cultural geography is clearly reductive and redundant. Similarly, the contemporary emphasis on the ways in which the past features in the present does not make the work of historical geography any less concerned with happenings of history. Rather, historical geography is about the interplay between the past and the present, between symbolic and material worlds.

See also: Cartography, History of; Geographic Thought, History of; Geography; Historical Change and Human Development; Historical Demography; Historical Sociology; Historiography and Historical Thought; Current Trends; History and the Social Sciences; History: Forms of Presentation, Discourses, and Functions; Postmodernism in Geography; Public Sphere: Eighteenth-century History; Public Sphere: Nineteenth- and Twentieth-century History

Bibliography

- Butlin R A 1993 *Historical Geography: Through the Gates of Space and Time*. Edward Arnold, London
- Braudel F 1988–90 *The Identity of France* (trans. Reynolds S), 2 Vols. Fontana Press, London
- Claval P 1984 The historical dimension of French geography. *Journal of Historical Geography* 10: 229–45
- Conzen M P 1993 The historical impulse in geographical writing about the United States, 1850–1990. In: Conzen M P, Rumney T A, Wynn G (eds.) *A Scholar's Guide to Geographical Writing*

- on the American and Canadian Past*, University of Chicago Press, Chicago, pp. 3–90
- Donkin R A 1997 A 'servant of two masters'? *Journal of Historical Geography* 23: 247–66
- Evans E E 1981 *The Personality of Ireland: Habitat, Heritage and History*, 2nd edn. Blackstaff, Belfast, UK
- Graham B, Nash C (eds.) 1999 *Modern Historical Geographies*. Prentice Hall, Harlow, UK
- Graham B, Ashworth G J, Tunbridge J E 2000 *A Geography of Heritage: Power, Culture and Economy*. Arnold, London and Oxford University Press, New York
- Guelke L 1997 The relations between geography and history reconsidered. *History and Theory* 36: 216–34
- Harris C 1978 The historical mind and the practice of geography. In: Ley D, Samuels M S (eds.) *Humanistic Geography: Prospects and Problems*, Croom Helm, London and Maaroufa Press, Chicago, pp. 123–37
- Heffernan M 1997 The future of historical geography. *Journal of Historical Geography* 23: 1–2
- Hugill P J 1993 *World Trade Since 1431: Geography, Technology, and Capitalism*. John Hopkins University Press, Baltimore and London
- Lowenthal D 1998 *The Heritage Crusade and the Spoils of History*. Cambridge University Press, Cambridge, UK and New York
- Nash C 1999 Historical geographies of modernity. In: Graham B, Nash C (eds.) *Modern Historical Geographies*, Prentice Hall, Harlow, UK, pp. 13–40
- Ogborn M 1997 (Clock)work in historical geography: autumn 1995 to winter 1996. *Progress in Human Geography* 21: 414–23
- Ogborn M 1999 The relations between geography and history: work in historical geography in 1997. *Progress in Human Geography* 23: 97–108
- Vidal de la Blache P 1926 *Principles of Human Geography* [English translation by Bingham M T]. Constable, London and H. Holt and Company, New York
- Wallerstein I 1974–89 *The Modern World-System*, 3 Vols. Academic Press, New York
- Wishart D 1997 The selectivity of historical representation. *Journal of Historical Geography* 23: 111–18
- Wynn G 1993 The writing of Canadian historical geography. In: Conzen M P, Rumney T A, Wynn G (eds.) *A Scholar's Guide to Geographical Writing on the American and Canadian Past*, University of Chicago Press, Chicago, University of Chicago Research Paper no. 235, pp. 91–124
- Wynn G 1997 Remapping Tutira: contours in the environmental history of New Zealand. *Journal of Historical Geography* 23: 418–46
- Yeoh B S A 1996 *Contesting Space: Power Relations and the Urban Built Environment in Colonial Singapore*. Oxford University Press, Kuala Lumpur and New York

B. Graham

Historical Linguistics: Numerical Methods

1. Introduction

The main goal for numerical methods in historical linguistics has always been to answer the question of how closely two languages are related. Applied pairwise to languages, this can be used to reconstruct a tree

for a whole language family (lexicostatistics). More ambitious methods have additionally assigned dates to the splits within those trees (glottochronology). Recently the questions and methods have been extended to involve distant (hence controversial) relationships between languages, so the basic question becomes not just how closely two languages are related, but whether they are related at all. All methods in this area, no matter how sophisticated, rest on the simplest of observations: since languages change over time, two originally related languages gradually become less similar over time. This enables the linguist to infer something about diachrony from synchronic information only, particularly attractive in situations where little is known about the history. The methodological difficulties come in quantifying rates of change, measuring difference, and accounting for the myriad of factors that can interfere with so-called normal change processes (e.g., mutual influence between languages). The attractions of numerical methods in historical linguistics are the same as those attributed to quantitative methods in almost any discipline in the humanities or social sciences: speed, objectivity, replicability, and ability to handle large volumes of data efficiently.

2. Early Methods

The earliest precursor to a numerical method in historical linguistics is a statement made by Robert Gordon Latham (1850, p. 565, see also Hewes in Hymes 1960, p. 338): 'the average rate at which languages change is capable of being approximated' and 'the maximum difference, at a given period, between two or more languages is also capable of being approximated.' He never pursued this, and the next milestone in the development of numerical methods in historical linguistics was a statement by Sapir (1916, p. 76): 'The greater the degree of linguistic differentiation within a stock the greater is the period of time that must be assumed for the development of such differentiation.' Sapir's influence on later linguists was great, so even though he himself never followed this statement up by taking the basic steps of quantifying linguistic differentiation or rate of change, this can be seen as the genesis of the most widely known numerical method in historical linguistics, that of Swadesh (1950, and see below).

The first explicit numerical treatment of genetic proximity of languages was Kroeber and Chrétien (1937), following a method created by two anthropologists, Jan Czekanowski (1928) and Stanislaw Klimek (1935), for use in physical anthropology and ethnography. For each pair of languages, on a list of N features, calculate a (number of features shared by both languages), b (number of features in the first language but not the second), c (number

of features in the second language but not the first), and d (number of features in neither language). This produces a 2×2 contingency table, a well-studied mathematical entity enabling various forms of analysis. In particular, the quantity Nr^2 , where $r = (ad - bc) / \sqrt{(a+b)(c+d)(a+c)(b+d)}$, is χ^2 (one degree of freedom), enabling one to test the statistical significance of the relationship between the two languages. The values of r for all pairs of languages under consideration can become the input to various methods (e.g., hierarchical cluster analysis) for family tree reconstruction. The method is simple but not unproblematic (e.g., construction of the feature-list, difficulties with non-independence of features, interpretation of correlations especially if negative). (For fuller treatment and commentary, see Embleton 1986.)

The second numerical method in historical linguistics is Ross's root-retention model (1950). For N features and m languages, construct an $N \times m$ table in which a cell has a cross if that language exhibits that feature. The question of relationship of languages can then be recast as whether or not the crosses are distributed randomly in the table. For each pair of languages, calculate n_1 (number of features found in the first language is represented by the number of crosses in the first column), n_2 (number of features found in the second language is represented by the number of crosses in the second column), and r (number of features shared by both languages is represented by the number of rows with a cross in both columns). Then the probability for that pair of languages of obtaining the shared number of crosses or more is

$$\sum_{R=r}^N = \frac{\binom{n_1}{R} \binom{N-n_1}{n_2-R}}{\binom{N}{n_2}}$$

for which a normal approximation exists. These probabilities can be tested for significance, and could become the input for various family tree reconstruction methods. Interestingly, Ross's method can be shown to be mathematically equivalent under certain conditions to Kroeber and Chrétien's, and thus subject to the same problems. (For fuller treatment and commentary on this and the method in general, see Embleton 1986.) Although certain elements of these models recur today, both have largely been abandoned, although the problems encountered and lessons learned remain relevant.

Historically, the next method to evolve was Swadesh's, but given its centrality to the field and ongoing relevance, I will deal with that in a separate section. A final numerical method to consider is that of counterindications originated by Gleason (1959), later rediscovered by others. All topologically possible family trees for the languages in question are constructed, each is evaluated for how well it fits against a fixed list of features (every time the tree does not fit the distribution of that feature, a counterindication is

recorded), and the tree with the best score (even if not perfect) is chosen. This method is difficult to implement for two reasons. Even with modern high-speed computing and sophisticated algorithms, the number of topologically possible trees for even a moderately-sized language family quickly becomes unworkably large. The method is also highly vulnerable to the effect of borrowing across languages. A variant of Gleason's method, relying on hierarchical clustering and attempting to estimate branch lengths (i.e., used glottochronologically) is shown in Embleton (1986), along with fuller discussion of the method itself. Taylor et al. (2000), using recent advances and sophisticated techniques from computer science and biological cladistics, are developing a character-based (rather than distance-based) method for family-tree reconstruction with some similarities to Gleason's method. Despite difficulties with dealing with borrowing in realistic amounts, such methods are promising partly because they closely parallel the cognitive processes of skilled practitioners of traditional comparative-historical reconstruction techniques, while being more objective and simultaneously handling vast amounts of data.

3. Swadesh

By far the best known numerical method in historical linguistics is that introduced by Swadesh (e.g., 1950), actualizing Sapir's statement by regarding lexical replacement as analogous to radioactive decay. As with carbon-14 dating, dates could then be provided for branching points in the tree (glottochronology). Take a list of N meanings of supposed basic core vocabulary, found universally in all cultures, expressing basic noncultural concepts, expected to be more resistant to change and borrowing than other vocabulary items. Such a list, now known as a Swadesh list, contains such items as lower numerals, simple kinship terms, body parts, topographical terms, widespread flora and fauna, pronouns (personal, demonstrative, interrogative), verbs denoting basic actions, and naturally occurring noncultural phenomena. Assuming a constant rate of replacement, R , for words representing meanings on this list, one can calculate the time elapsed (time depth) since any pair of languages split using the formula

$$\frac{-1}{2R} = \ln \left(\frac{c}{N} \right)$$

where c is the number of cognates on the Swadesh-list between the two languages compared. By calculating time depths for all pairs of languages under consideration, we can construct a family tree with dates.

Swadesh's method was greeted with widespread enthusiasm and widely applied in the late 1950s, 1960s, and early 1970s, almost with a sense of euphoria about finally having an exact tool for obtaining insights into hitherto unknowable linguistic prehistory. The debate over Swadesh's method quickly became emotional, but regardless of the researcher's subjectivity, a common core of problems can be identified, sometimes with potential solutions.

Difficulties with particular meanings on the Swadesh-list.

(a) Some meanings cannot be translated into the target language. This is fairly rare, and best treated as missing data, reducing N accordingly.

(b) Sometimes there are multiple possible translations in the target language for a meaning on the list. Different researchers might make different but equally valid choices from the range of possibilities ('multiple synonymy'). Solutions include always counting only the most frequent possible translation, choosing at random, always choosing to maximize the cognate count, always choosing to minimize the cognate count, or various fractional scoring systems.

Difficulties with cognacy judgments (remember one is not necessarily dealing with languages where the history, and hence cognacy are well known).

(a) Due to accumulated sound change, cognates may not be recognizable, causing us to overestimate time depth.

(b) Chance similarities may be wrongly interpreted as cognates, causing us to underestimate time depth. In practice, one might hope that this and the preceding cancel out.

(c) Sometimes only one part of a word is cognate to a word in another language ('partial cognation'); how should one score this? Solutions include always counting this as positive, always counting this as negative, or various fractional scoring systems.

Problems with the universality of R , which was shown to vary over time, between languages, and according to meaning. R had been estimated from languages with written historical records as about 14 percent per millennium for the standard 100-word Swadesh-list and about 19 percent per millennium for the standard 200-word Swadesh-list. Not only might the very fact that these languages are written bias the rate (one might expect less replacement in a language with a lengthy written history), but 11 of the 13 pairs used to establish the rate were Indo-European (representing only three of its branches).

Problems with the family tree (Stammbaum) model, generally centering around the splits not being clean (thus giving any precise date is artificial and meaningless). Languages are often in close physical proximity even after they split, mutually influencing one another; tendencies in the two languages predating the split may come to fruition only afterwards. Both these interfere with the model, but are more properly problems with the family tree model, and Swadesh's

adoption of it, rather than problems with Swadesh's method itself.

Borrowings between languages (related or unrelated) can alter replacement rates quite substantially. For some languages, even the Swadesh-list contains a high amount of borrowed vocabulary.

For various reasons (see Embleton 1986 and elsewhere), the use of Swadesh-style lexicostatistics and glottochronology waned by the mid-1970s, although applications lingered on in several language areas, and the methods continued to be refined by applied mathematicians.

4. Sankoff

Most refinements introduced by applied mathematicians involved adding further parameters to the basic Swadesh model. For example, parameters were added to account for drift and for 'chance recurrent cognation' (i.e., when a word was replaced twice, such that the original word was restored), replacement rates were allowed to vary from meaning to meaning, a very crude allowance was made for borrowing, and there was talk of allowing the rates to vary from language to language and from time to time. (Further details, along with some models, can be found in Embleton 1986.) Some of these are useful refinements (e.g., allowing replacement and borrowing rates to vary over time, between languages, and between meanings), but these are very much a mathematician's way of solving the problem, producing quite elegant solutions. But in the real world in which the linguist works, all those parameters have to be estimated—and this is surprisingly difficult even for well studied languages for which we actually have both long and extensive written records. It is also the case that an extraordinary amount of data is required to estimate so many parameters, and this is most often simply impossible. So these elegant theoretical solutions have very little practical import.

However, the single most important development was the incorporation of a geographical dimension into the family tree, thus allowing for borrowing to be largely dealt with. The basic insight here is due to Sankoff (1972), with correction and further refinement by Embleton (see 1986 for the most comprehensive treatment). The resulting set of differential equations required to reconstruct the tree and estimate the branch lengths is complicated, reflecting the increased power of the method (equations are given in Embleton 1986 and elsewhere). The parameters required all have some reasonable chance of being estimated, even for lesser studied families; these are a similarity measure for each language pair, replacement rates for each language, borrowing rates between each language pair (not necessarily symmetric), and knowledge of which languages are in contact with one another.

5. Starostin

Largely independently of the work described above, research on numerical methods in historical linguistics was proceeding in the Soviet Union (later Russia, still later 'in diaspora'). Unlike the impetus for Swadesh's work, which was largely to aid in providing histories and trees for relatively little studied language families (e.g., Amerindian), much of this work gained its impetus from questions of more distant linguistic relationship. This had its effect on the development of the methods, because it was crucial that the methods be able to produce useful results even at much greater time depths, where Swadesh's method typically has difficulty distinguishing the signal (cognacy, true genetic relationship) from the noise (chance resemblance, contact relationship).

Historically the first advance here was due to Dolgopolsky (1986), who capitalized on the fact that different meanings change at different rates. He compiled a list of the 15 most stable meanings, based on examination of 140 languages. One can then calculate the probability of fortuitous coincidence of the words representing those 15 most stable meanings for any language pair; if the probability is significantly less than that expected by mere chance, this can be taken as evidence of relationship. This method is most striking in its extreme simplicity, but can also be related to some more complex and ongoing work, largely carried out in eastern Europe, founded on the observation that several characteristics of a word (e.g., length, number of meanings, age) are intercorrelated with its frequency. Words of more recent origin tend to be less frequent, longer, and have fewer meanings, whereas older words tend to be more frequent, shorter, and have more meanings. The age-frequency relationship in particular is being extensively researched, as there may be a small hard core of ancient short vocabulary which may reveal distant relationships, and may even be able to be harnessed into a lexicostatistical or even glottochronological method.

Another advance has been made by Sergei Starostin, with work first published in the 1980s and continuing to the present, known as root glottochronology or etymostatistics (see Starostin 1999). Compared to standard Swadesh-style glottochronology, the main innovation is the observation that while glottochronology regards any change (however small) in a word's meaning as a loss of that word from that meaning, root glottochronology does not. Thus while standard glottochronology sees English *dog* and German *Hund* in the meaning 'dog' as noncognate, root glottochronology notices the existence of the English word *hound*; this more conservative approach to the maintenance of a language's root stock enables one to work better at greater time depths. Starostin's method also gives more accurate results for known cases where the time depths are more shallow (standard glottochronology tends to give dates which are too recent).

Root glottochronology takes a text, containing a given set of words in the parent language, and then measures the number of these words still extant (in any derivative, with any semantic shift) in the daughter language. It proceeds only with detailed comparative-historical and etymological analysis (eliminating look-alikes and loans), and also can be repeated for more than one text (hence a different set of words), giving greater precision and potentially some estimate of error (see Starostin 1999 for formulas and more detailed explanation).

6. Ringe

Various linguists have used probabilistic methods to assess similarity between vocabularies of two different languages, usually focusing on phonological similarities (an early example is Ross 1950). The basic question is whether the observed degree of similarity between the two vocabularies is greater than what could be expected based on chance. Although the problem is conceptually simple, it quickly becomes mathematically intractable even between just two languages, as one has to wrestle with different phoneme distributions, different phoneme frequencies, how to measure phonetic distance between different phone(me)s, different phonotactic structures, and lengths of roots, quite apart from the questions (familiar from glottochronology) of how much semantic latitude to allow between the words compared. The most recent and sophisticated treatment of this problem is by Ringe (e.g., 1998 and references therein), who attempts to account for many of the above issues, albeit with simplifying assumptions. This research remains active, yet without consensus as to its utility, partly because of the degree of the simplifying assumptions, but also partly because it has mostly been applied in the realm of long-distance relationships, where the emotions over the results can sometimes cloud the objectivity with which the method is viewed.

7. Future Directions

Research in numerical methods in historical linguistics is gradually reentering the mainstream after several decades on the sidelines. The most active areas of current research are improvements to the basic Swadesh model, exploration of the age–frequency relationship, root glottochronology, computer-implemented tree reconstruction using methods borrowed from biological cladistics, and Ringe’s probabilistic methods. The degree of mathematical, statistical, or computational expertise required is extremely high, which is likely to prevent the majority of historical linguists from ever mastering the methods, and consequently render them unable to do more than use results produced by others or participate in research

teams where others control the methods. There seems to be a growing realization, however, that numerical methods, while not providing any magic solutions to any long-standing problems, can make a useful contribution as one of the tools that we use to explore our linguistic history and prehistory.

See also: Archaeology and the History of Languages; Historical Linguistics: Overview; Lexicology and Lexicography

Bibliography

- Czekanowski J 1928 Na Marginesie Recenzji P. K. Moszyńskiego o Książce: Wstęp do Historji Słowian. *Lud*, Series II, Vol. 7. Reprint Lwow, 1928
- Dolgopolsky A 1986 A probabilistic hypothesis concerning the oldest relationships among the language families in northern Eurasia. In: Shevoroshkin V, Markey T (eds.) *Typology, Relationship and Time: A Collection of Papers on Language Change and Relationship*. Karoma, Ann Arbor, MI, pp. 28–50
- Embleton S M 1986 *Statistics in Historical Linguistics*. (*Quantitative Linguistics*, Vol. 30). Brockmeyer, Bochum, Germany
- Gleason H A Jr 1959 Counting and calculating for historical reconstruction. *Anthropological Linguistics* 1: 22–32
- Hymes D H 1960 Lexicostatistics so far. *Current Anthropology* 1: 3–44; 338–45
- Klimek S 1935 The structure of California Indian culture, culture element distributions: 1. *University of California Publications in American Archaeology and Ethnology* 37: 1–70
- Kroeber A L, Chrétien C D 1937 Quantitative classification of Indo-European languages. *Language* 13: 83–103
- Latham R G 1850 *The Natural History of the Varieties of Man*. Van Voorst, London
- Ringe D R Jr 1998 Probabilistic evidence for Indo-Uralic. In: Salmons J C, Joseph B D (eds.) *Nostratic: Sifting the Evidence*. Benjamins, Amsterdam, pp. 153–97
- Ross A S C 1950 Philological probability problems. *Journal of the Royal Statistical Society B* 12: 19–59
- Sankoff D 1972 Reconstructing the history and geography of an evolutionary tree. *American Mathematical Monthly* 79: 596–603
- Sapir E 1916 *Time Perspective in Aboriginal American Culture* (*Memoir 90, Anthropological Series*, Vol. 13). Government Printing Bureau, Ottawa, Canada
- Starostin S 1999 Comparative-historical linguistics and lexicostatistics. In: Renfrew C, McMahon A, Trask L (eds.) *Time Depth in Historical Linguistics*. McDonald Institute for Archaeological Research, Cambridge, UK, Vol. 2, pp. 377–413
- Swadesh M 1950 Salish internal relationships. *International Journal of American Linguistics* 16: 157–67
- Taylor A, Warnow T, Ringe D 2000 Character-based reconstruction of a linguistic cladogram. In: Smith J C, Bentley D (eds.) *Historical Linguistics 1995. Selected Papers from the 12th International Conference on Historical Linguistics, Manchester, August 1995, Vol. 1, General Issues and Non-Germanic Languages*. Benjamins, Amsterdam, pp. 393–408

S. Embleton

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.

6747

Historical Linguistics: Overview

Historical linguistics is the study of language through time, and involves the consideration of changes from some earlier stage towards the present; the construction of earlier, unattested states; and the determination and depiction of relatedness among languages and language groups. Although the main focus of twentieth-century linguistics was synchronic, there is currently increasing interest in historical linguistics, both from linguistic theory and cognate historical disciplines, notably archaeology and genetics.

1. The Scope of Historical Linguistics

Historical linguistics involves the study of languages as they change through time. It includes tracing developments from an earlier stage of a language to a later one: for instance, English 'silly' meant 'blessed' at the time of Chaucer; Latin sentences often had the word order Subject Object Verb, as in *Helena Marcum amat* 'Helena loves Marcus,' whereas its daughter language French typically has Subject Verb Object, as in *Hélène aime Marc*; and many Old English noun and verb endings with final nasal consonants, like *-an*, *-um*, *-on*, dropped gradually about 1,000 years ago. Historical linguists are interested in when, how, and ultimately why changes of this sort happened.

However, this diachronic perspective also operates in the opposite direction, encouraging historical linguists to look back from present-day or recorded languages to earlier, unattested stages in an attempt to reconstruct unrecorded features and characteristics. This may involve filling in gaps in the historical record, for instance in hypothesizing the form a word might have taken in Old English or in Latin, although that word does not actually appear in any surviving manuscript. At greater time depths, features of unattested, hypothetical ancestral languages, like Proto-Indo-European, the probable source of most languages of modern Europe and northern India, may also be reconstructed. Work of this kind also explicitly or implicitly involves consideration of language classification: reconstruction necessarily includes the hypothesis and testing of linguistic family groups, which are taken to be demonstrably descended from a single common source. At present, it is uncontroversial to say that English, French, and Latin are all related, as members of the Indo-European family; but it would be more controversial to contend, as some historical linguists do, that these languages are in turn related to Semitic languages like Arabic, Akkadian, and Hebrew; or that Chinese is related to Chechen and other Caucasian languages.

In this article, the development of historical linguistics will first be considered briefly, and its place within

modern linguistics assessed. Then the two subdisciplines of language change, and linguistic classification and reconstruction, will be examined in turn, with the emphasis on outlining currently lively and productive areas of research. In the final section, some trends common to change and reconstruction will be identified, with their implications for future work in historical linguistics.

2. Historical Linguistics Within Linguistics

In the heyday of traditional work in the late nineteenth century, historical-comparative linguistics was linguistic theory. Its focus and methodology were in line with the strongly historicizing tendency of the day, and with the development of parallel sciences such as geology and evolutionary biology, so that although the foundations for language comparison and study had been laid considerably earlier, the discipline advanced rapidly and attracted considerable scholarly attention in this period (see Davies 1998). The cornerstones of the discipline were set with the classification and reconstruction of Indo-European in particular, and the Comparative Method on which such reconstructions are based (see Fox 1995) is still the foremost tool of the reconstructionist's trade, although many of the specific results obtained have been modified, as knowledge of language change has deepened and developed, and more data have become available. The Comparative Method requires sets of cognate forms in descendant languages: these will be similar in form and meaning, and will show regular, recurrent correspondences of particular sounds across the group of languages compared. We can then hypothesize that these have been derived historically from a single most plausible common source, by the operation of generally regular sound changes. Thus, given English *father*, Irish *athair*, Latin *pater* and Greek *pater*, alongside a large number of similar comparands which equally lack an initial consonant in Irish, and have English /f/ corresponding to Latin /p/, it seems reasonable to hypothesise Proto-Indo-European **pátér*. It is also possible to set out the contexts in which particular sound changes have operated in the daughter languages, and therefore to begin to identify the factors in the surrounding sounds which may have led to these developments (see *Linguistics: Comparative Method*).

Although many of these methodological and theoretical assumptions still hold good in historical linguistics today, the twentieth century has seen a radical shift in perspective in linguistics as a whole. The development of structuralism led to a move away from the study of individual languages and families, and towards identifying language universals, those aspects of human language which recur regardless of genetic linguistic affiliation. Consequently, the historical linguistic concentration on variation as a key to linguistic

relatedness gave way to attempts to identify theoretical limits on variation. Finally, whereas most earlier work had concentrated on vocabulary and phonology, the 'Chomskyan revolution' of the 1960s centered attention on syntax.

However, these rapid and far-reaching changes have not led to the demise of historical linguistics; although it is true that historical work became less of a core part of linguistics for some years in the mid-twentieth century, it is now again one of the liveliest areas of the discipline. This is for three main reasons. First, advances in linguistic theory, while typically based on synchronic study and data, often allow the sources of change to be identified and explained in a way which was not previously possible: much nineteenth-century work, although allowing individual elements to be plausibly reconstructed, was piecemeal and did not trace related changes to a common cause. Explanation comes center stage in modern linguistic work. Second historical data provide an additional means of testing and refining these linguistic theories, which may benefit from a demonstration that they can analyze and incorporate language change. Finally, related disciplines like genetics and archaeology may not always find developments in linguistic theory *per se* particularly useful or enlightening; on the other hand, interactions of data and hypotheses from these areas with those of historical linguistics are promising. All these areas will be considered, with respect to both change and reconstruction, in the following sections.

3. Areas of Debate in Language Change

3.1 Phonology

Much early work on sound change was based on the assumption of relative ease of articulation: that is, it is presumed to be more straightforward for a speaker to utter two similar adjacent sounds, or indeed to make them identical, than to produce a sequence of sounds which require major vocal tract movements between the two. This explanation for the extremely common sound change of assimilation, as seen in Latin *somnum* → Italian *sonno*, was put forward in the last century. However, advances in phonetic technology has allowed historical linguists to pinpoint more accurately the types of sound sequences which are likely to cause such problems, and the advent of experimental phonology has allowed these hypotheses to be tested, not only from the point of view of the speaker's production problems, but also taking the hearer's perception difficulties into account (see *Phonetics: Articulatory*). One of the pioneers in this area, John Ohala (1993), has suggested that certain sound changes like dissimilation, which increases the difference between close or adjacent sounds in the stream of speech, may be provoked by the hearer's

expectations. Thus, a hearer receiving the Proto-Quechumaran word */t'ant'a/ 'bread' might erroneously have assumed that the speaker has produced two ejective [t']s close together by assimilating a second, nonejective stop to the first truly ejective one. Undoing this perceived but not real mistake by the speaker produces the modern Quechua form /t'anta/. Advances in phonological theory may also be relevant here, for instance in helping us understand what it means for sounds a and b to be closer in their constituent features than sounds b and c; or in showing that the apparently disparate environments triggering a particular sound change in fact form a unified class in terms of some deep principle of phonology. In this respect it is encouraging that, while certain phonological theories current in the 1980s, such as Government Phonology and some forms of Declarative Phonology, explicitly ruled historical processes out of their remit, Optimality Theory, a currently popular model which has been developing since the early 1990s, already includes a considerable body of work addressing problems of both change and variation.

3.2 Morphology and Syntax

Morphology has often been left out of historical linguistics as a separate domain of enquiry, being seen as simply an interface: phonological changes, like the loss of final unstressed syllables, may have morphological consequences if case or tense markers are lost, and this in turn may have a knock-on effect on the syntax, changing or fixing word order. However, morphology has more recently been the focus of attempts within Natural Morphology (Wurzel 1989) to identify those features which are favored universally, for instance in terms of ease of production, perception, or acquisition, and the factors underlying that naturalness, as well as assessing why in many cases these universally natural morphological features are subordinated to language-specific requirements which may reduce naturalness. Morphological change is also one component of grammaticalization, the extremely common process whereby independent words become grammatical markers, as in German *Drittel* 'third' and *Viertel* 'quarter,' formed from the numerals *drei* 'three' and *vier* 'four' plus a reduced form of *Teil* 'part'; or the famous French negative *ne ... pas*, where *pas* is derived from Latin *passum* 'a step,' and became an integral part of the negative construction in earlier French, perhaps via gradual bleaching of reinforcing constructions meaning 'I won't go a step.' Grammaticalization has become a major focus of research in linguistic theory (see Hopper and Traugott 1993), introducing many students and researchers to issues in linguistic change (see *Grammaticalization*).

In syntax, historical work has become enormously more sophisticated over the last 30 years, reflecting the

newly ascendant status of syntax in synchronic work and in linguistic theory. Work at first may have been hampered by lack of knowledge of the range of syntactic changes, and uncertainty over quite fundamental questions, such as whether syntactic change can be seen as essentially regular in the same way as many sound changes. These issues are now being addressed by major cross-linguistic surveys (see Harris and Campbell 1995). On a more theoretical level, Lightfoot (1991, 1999) uses historical data and current models of syntactic theory in a mutually informing way, attempting to show that children learn syntax in the way they do because of particular factors in their innate, universal grammar, but that certain superficial changes in syntactic structures may also cause them to learn differently, setting up new structures which will form the basis for subsequent development in their language.

3.3 Semantics

Developments in linguistic theory may also increase understanding of semantic change, which has long been seen as a domain where description is the highest achievable goal. Earlier approaches have attempted to classify changes, but seeing developments as, for instance, extensions, or restrictions of meaning does not necessarily explain why these occur, and linguists have tended to fall back on the old adage that 'every word has its own history.' Developments in pragmatics and in cognitive linguistics may improve prospects here. For instance, Traugott (1982) suggests that many words develop from an external to an internal, subjective sense: thus, English *feel* begins as descriptive of some characteristic external to the speaker, in *It felt soft*, but comes to signal internal states, in *I feel miserable*. Sweetser (1990, p. 2) suggests that linguists must crucially adopt 'a cognitively based theory which takes not the objective "real world," but human perception and understanding of the world to be the basis for the structure of human language': this approach allows recurring semantic changes, such as those involving mind-as-body metaphors of the *grasp*, *cast light on*, and *see* type, to be unified.

3.4 Variation and Change

One of the most productive recent developments in historical linguistics has been the recognition of a connection between variation and change, and the development of sociohistorical research (see *Sociolinguistics*). First, this substantially increases the database of changes available for study, since changes in progress can now be considered alongside the more traditional category of change from earlier, written data to the present day. However, for spoken, colloquial data to be accepted, it is necessary to delimit

and define what constitutes a change in progress, requiring new quantitative techniques for identifying and measuring ongoing change. Innovations, which may provide the seeds of future developments, must now be distinguished from changes, and innovators from early adopters, the speakers who begin to use a new variant and spread it to their own social group (see Milroy 1992, Labov 1994).

This sociohistorical work has also directed attention away in part from internal, systemic factors and towards external factors in the search for explanations for change. Internal factors may belong to the language system itself, or to some other level of the grammar, or involve the innate language faculty which many linguists see as unique to humans; such factors have typically been favored by theoretical linguists in their attempts to explain change. However, socio-linguistic work places great emphasis on the social characteristics of speaker and interlocutor, including age, sex, and social class, and on characteristics of the interaction itself, such as level of formality and topic of exchange. This in turn encourages focus, not only on interaction between speakers, but between varieties and languages. The consequences of language contact have therefore become vitally important in recent historical linguistics, with increasing interest in contact phenomena like borrowing and convergence; the development of pidgins and creoles; and socio-linguistically determined language breakdown and loss in language death (Bickerton 1981, Thomason and Kaufman 1988, Dorian 1989). These are no longer being seen as peripheral and uncommon, but as a central part of language change and germane to a greater understanding of language itself (see *Language Contact*; *Areal linguistics*; *Pidgin and Creole Languages*; *Language Obsolescence*).

4. Areas of Debate in Reconstruction and Classification

4.1 Typology and Reconstruction

In the closely related subfields of classification and reconstruction, matters have advanced considerably since the various nineteenth-century attempts to write a fable in Proto-Indo-European. There is still much interest in what is reconstructed: the reliability and realism of reconstructed systems remain open questions, as does the issue of whether syntax can be reconstructed using the same methods as phonology, or whether early, unattested areas such as word order are beyond recovery (Lass 1997). Innovations in phonological reconstruction continue to occur: for instance, the 'new look' of Proto-Indo-European proposes a set of ejectives alongside the more traditional voiceless and voiced stops, on the basis that the resulting system would be more common cross-

linguistically, even though the alleged ejectives have disappeared, largely without trace, in the daughter languages. This opens a wider debate over the roles of typology and uniformitarianism in reconstruction (see *Linguistic Typology*): that is, to what extent should reconstructions be driven by generalizations that hold for modern-day languages? Is it more important to reconstruct a proto-form or proto-system which accords with the comparative evidence from its daughters, but which might appear skewed when compared with other attested systems, or to produce a balanced, typologically normal system although considerable changes may be required to derive the daughter forms?

4.2 Classification and Methodology

However, classification has recently threatened to overtake reconstruction as the main focus of interest in this area, and here the questions are primarily methodological. The Comparative Method (see *Linguistics: Comparative Method*), with its concentration on regular, repeated correspondences and its incorporation, at least for many practitioners, of reconstruction as a test of the family groups proposed, has been challenged by Greenberg's much simpler method of multilateral comparison, based on straightforward surface inspection of word-lists from many languages. Family groups are then proposed on the basis of etymologies drawn from these lists: forms within an etymology are similar to an extent in form and meaning, but the similarities need not be recurrent, and reconstruction is not necessarily undertaken. These families tend to be considerably larger (and hence potentially older) than those conventionally reconstructed using the Comparative Method. For instance, Greenberg (1987) uses multilateral comparison to propose only three language families for the Americas, including Amerind, a phylum or superfamily containing more than 1,000 languages. Even more controversial connections are drawn by other practitioners of multilateral comparison, for instance between the Sino-Tibetan and Caucasian languages; and there have even been attempts to produce universal etymologies, including forms from all or most language superfamilies, which might suggest forms of Proto-World, the alleged first language of mankind (see Ruhlen 1994).

4.3 Connections with Archaeology and Genetics

Multilateral comparison remains extremely controversial. Its apparent ease of use may be deceptive, not compensating for problems of reliability of data, and potential laxness of the criteria governing what counts as a phonetic or semantic match. Many historical

linguists are reluctant to accept the method or its results, and therefore doubt whether language groupings above the traditional family level, and beyond a time threshold of perhaps 5,000–8,000 years, can be conclusively established.

Long-range comparativists are, consequently, increasingly turning to data from outside linguistics itself in an attempt to validate their work, and exploration of the potential for historical linguistics and related disciplines like archaeology and genetics to produce mutually informing and enforcing data is becoming central. For instance, Greenberg et al. (1986) cite genetic and archaeological data which seem to favour a three-way migration of peoples into the Americas, mirroring Greenberg's tripartite classification of New World linguistic groups. Cavalli-Sforza et al. (1994), in work which caught the popular scientific imagination, analyzed genetic data from 142 indigenous populations worldwide, and argued that the resulting population tree closely matched a linguistic family tree including controversial linguistic phyla.

However, these correlations are themselves problematic. The lack of a one-to-one correspondence between language and archaeological culture means interpretation is all important, and increases the need for data from the fields under comparison to stand independently and to be subject to rigorous validation within their own areas before cross-disciplinary conclusions are drawn. In genetics, in particular, advances in the data-types which can be considered may produce a range of possible population histories rather than a single definitive answer: thus, Cavalli-Sforza et al.'s (1994) picture of the Americas, based on classical analysis of proteins, does not agree with the genetic groupings derived from more recent analysis of mitochondrial DNA and Y chromosomes; these will therefore produce population trees which need not match a particular linguistic tree.

Connections of historical linguistics with genetics and archaeology are also relevant beyond the domain of long-range comparison and influences move in both directions. For instance, Dixon (1997) applies the biological model of punctuated equilibrium to language, arguing that Comparative Method and the associated family tree model may be relevant only to periods of concentrated change and expansion in the histories of language families. At intervening periods of relative equilibrium, language contact and gradual diffusion of features would determine the course of language change. Geneticists and archaeologists are also increasingly seeking a historical linguistic methodology which is truly independent and repeatable, and which might transcend arguments among its practitioners, for comparison with their own data: the development of quantitative techniques of language comparison therefore remains a pressing goal within historical linguistics (see *Historical Linguistics: Numerical Methods*).

4.4 The Evolution of Language

If our attempts to reconstruct might be blocked beyond a certain time-depth on methodological grounds, some progress might be made by starting from the other end: that is, by considering the origin of language in our species. This is a topic which logically belongs with historical linguistics, at least from a nonspecialist point-of-view, but which has traditionally been excluded from the field, being famously banned by the Société de Linguistique de Paris in the 1860s. However, some rehabilitation has taken place during the 1990s, with increasing interchanges of ideas between historical linguists, computational linguists, primatologists, neurobiologists, and psychologists; Hurford, Studdert-Kennedy and Knight (1998), for instance, includes contributions on the possible emergence of phonology and syntax, with some use of computer simulations to show how learning of these new structures might have taken place, as well as work invoking more social and anthropological considerations to explain the rise of human language. Again, this is by definition interdisciplinary research, which raises questions fundamental to historical linguistics (such as how the, probably simple, initial constructions of early language developed into the modern range of variants), and to linguistics in general (such as the limits which might be set on what aspects of language are innately specified and which might be learned, this in turn depending in part on whether the underlying systems responsible for language evolved by natural selection or are in some way maladaptive).

5. Common Trends in Reconstruction and Change

In the sections above, change and reconstruction have been presented separately for convenience. However, these areas are not truly independent, and some common themes can be discerned which may be important for the future of historical linguistics in general.

First, it must be emphasized that change and reconstruction influence one another. For instance, if an early text or unknown language is deciphered, the database usable for drawing conclusions about changes is increased, but the same data may also confirm or refute the results of work on reconstruction.

Second, both subdisciplines need to become reconciled to the development and use of quantitative, computational approaches, and of tools. Work of this kind is probably more advanced in change, where techniques derived from sociolinguistics are already well established in some domains. The need is more pressing, and the issues involved more controversial, in classification.

Finally, it is encouraging to note that historical linguistics is becoming a major focus of research

within linguistics. This means that historical linguists in turn must be willing to work with synchronically-based theories, and to consider what cognate disciplines like genetics and archaeology might require, as well as what they can provide.

See also: Archaeology and the History of Languages; Genes and Languages; Historical Linguistics: Numerical Methods; Languages: Genetic Classification; Linguistics: Internal Reconstruction

Bibliography

- Bickerton D 1981 *Roots of Language*. Karoma, Ann Arbor, MI
Campbell L 1998 *Historical Linguistics: An Introduction*. Edinburgh University Press, Edinburgh, UK
Cavalli-Sforza L L, Menozzi P, Piazza A 1994 *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ
Davies A M 1998 Nineteenth century linguistics. In: Lepschy G (ed.) *History of Linguistics*. Longman, London, Vol. 4
Dixon R M W 1997 *The Rise and Fall of Languages*. Cambridge University Press, Cambridge, UK
Dorian N (ed.) 1989 *Investigating Obsolescence: Studies in Language Contraction and Death*. Cambridge University Press, Cambridge, UK
Fox A 1995 *Linguistic Reconstruction: An Introduction to Theory and Method*. Oxford University Press, Oxford, UK
Greenberg J 1987 *Language in the Americas*. Stanford University Press, Stanford, CA
Greenberg J, Turner C G, Zegura S L 1986 The settlement of the Americas: A comparison of the linguistic, dental and genetic evidence. *Current Anthropology* 27: 477–97
Harris A C, Campbell L 1995 *Historical Syntax in Cross-Linguistic Perspective*. Cambridge University Press, Cambridge, UK
Hopper P J, Traugott E C 1993 *Grammaticalization*. Cambridge University Press, Cambridge, UK
Hurford J R, Studdert-Kennedy M, Knight C (eds.) 1998 *Approaches to the Evolution of Language*. Cambridge University Press, Cambridge, UK
Labov W 1994 *Principles of Linguistic Change. I: Internal Factors*. Blackwell, Oxford, UK
Lass R 1997 *Historical Linguistics and Language Change*. Cambridge University Press, Cambridge, UK
Lightfoot D 1991 *How to Set Parameters: Arguments from Language Change*. MIT Press, Cambridge, MA
Lightfoot D 1999 *The Development of Language: Acquisition, Change and Evolution*. Blackwell, Oxford, UK
McMahon A M S 1994 *Understanding Language Change*. Cambridge University Press, Cambridge, UK
Milroy J 1992 *Linguistic Variation and Change*. Blackwell, Oxford, UK
Ohala J 1993 The phonetics of sound change. In: Jones C (ed.) *Historical Linguistics: Problems and Perspectives*. Longman, London, pp. 237–78
Ruhlen M 1994 *On the Origin of Languages: Studies in Linguistic Taxonomy*. Stanford University Press, Stanford, CA
Sweetser E 1990 *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge University Press, Cambridge, UK
Thomason S G, Kaufman T 1988 *Language Contact, Creolization and Genetic Linguistics*. University of California Press, Berkeley, CA

- Trask R L 1996 *Historical Linguistics*. Arnold, London
- Traugott E C 1982 From prepositional to textual and expressive meanings: Some semantic-pragmatic aspects of grammaticalization. In: Lehmann W P, Malkiel Y (eds.) *Perspectives on Historical Linguistics*. Benjamins, Amsterdam, pp. 245–71
- Wurzel W U 1989 *Inflectional Morphology and Naturalness*. Kluwer, Dordrecht, The Netherlands

A. McMahon

Historical Sociology

In 1844, Auguste Comte's *Lecture on the Positivist Outlook* (*Discours sur l'esprit positif*) proposed the name 'sociology' for the general science of humanity. In Comte's vital vision, the hierarchy of sciences proceeded from mathematics at the base through astronomy, physics, chemistry, and biology to sociology at the summit. The maturing of basic sciences, he declared, now made it possible to construct the capstone science, sociology. In that defining moment for the sociological discipline, Comte's conception of sociology included history, in fact consisted largely of analyzing the development of humanity through historical stages. From that point onward, however, professional history and professional sociology moved in very different directions.

As the historical discipline organized later in the nineteenth century, professional historians came to specialize in particular times and places, for example, modern Latin America or ancient China. This specialization built up clusters of expertise in the sources, languages, and institutions peculiar to different time-place settings, which in turn promoted collective research programs, debate among experts, education of newcomers, and exclusion of nonprofessionals from the historical guild. It also made syntheses of all human history in a Comtean style less attractive to ordinary historians.

Meanwhile, sociologists divided up their own specialties mostly by structures and processes—families, religions, industrialization, crime, and so on—and drew their evidence chiefly from observation of the countries they currently lived in. Sociologists' historical essays divided among broad evolutionary and stage schemes, scattered examples illustrating general arguments, and sketchy prologues to studies of contemporary social phenomena. From these contrasting bases of intellectual organization sprang a number of mutual misunderstandings between historians and sociologists, including the sociologists' common presumption that sociology is an explanatory, generalizing science and history a descriptive, particularizing science fated to provide raw material for sociologists' generalizations. Historians returned the compliment by complaining that the present studied by sociologists

was itself a narrow historical moment with no claim to universal significance, and that to pluck examples from history without thorough knowledge of the relevant sources, languages, and institutions courted intellectual disaster.

Despite increasing separation between the disciplines, to be sure, occasional thinkers since Comte have ventured into grand syntheses that do not obviously belong to a single discipline. The ideas and accomplishments of Karl Marx and Max Weber transcend any single discipline. Grand syntheses addressing major historical questions have continued to appear in the twentieth century, with inquiries by Barrington Moore, Jr., Norbert Elias, Immanuel Wallerstein, and William H. McNeill overflowing disciplinary containers. Such works belong to the common heritage of history and the social sciences rather than to the narrower field of historical sociology.

The idea of a distinct enterprise called historical sociology only formed during the twentieth century, as sociologists themselves began to distinguish sharply between evidence coming from direct observation of the present and from indirect observation of the past. Historical sociologists sought to connect the two through serious investigation of other settings than the ones they lived in. During the twentieth century, historical sociology has differentiated into four rather different pursuits. We might call them *social criticism*, *pattern identification*, *scope extension*, and *process analysis*. All are legitimate, effective forms of sociological analysis. They differ substantially, however, in aims, procedures, and results.

Historical social criticism reconstructs the past on the way to informing human choices in the present and future. It rests on the plausible assumptions that history contains a record of human successes and failures in dealing with problems persisting into the present; that if long-term trends exist, their causes are likely to continue; and that recent history limits what will happen next. Thus sociologists have created competing accounts of capitalism's development as liberation, as oppression, or as something in between. Similarly, sociologists of the family review its history as a means of criticizing current trends and charting alternative futures.

Although the description and criticism of contemporary social conditions preoccupied many nineteenth century sociologists, a large share of nineteenth century sociology consisted of historical social criticism—attempts to make better sense of current dilemmas and general directions of human development by situating the present in long, large social processes. Landmarks of nineteenth century social thought such as Karl Marx's *Capital*, Max Weber's *Economy and Society*, or Ferdinand Tönnies' *Gemeinschaft und Gesellschaft* embedded critiques of contemporary social life in vast historical syntheses. The tradition continues into our own time, as sociologists in the

varied styles of Göran Therborn, Theda Skocpol, Jürgen Habermas, and Orlando Patterson contribute to current social criticism by reworking history. At its best, historical social criticism helps people act wisely and effectively in the present by displaying possible futures and paths to them.

Historical social criticism overlaps with historical pattern identification, which it helped bring into being as a sociological specialty. Pattern identification searches for recurrent structures and sequences across time and space. Before World War II pattern identification often involved natural histories for large social aggregates, for example, Pitirim Sorokin's claim that each civilization passes through Ideational, Idealistic, and Sensate phases. (Sorokin's analyses edged over into social criticism as he argued that western civilization had reached the last, corrupt stage of sensate culture, and that only spiritual renewal could save the West.) Pattern identification also took the form of natural histories for wars, revolutions, social movements, and other smaller-scale phenomena.

After World War II, ideas of development and modernization sometimes followed the natural history model, but pattern-identifying historical sociologists generally moved toward the study of processes within countries or regions such as secularization, industrialization, revolution, democratization, and population growth. Pattern identifiers usually seek to discover necessary and sufficient conditions for large, recurrent social transformations and/or to establish general sequences within those transformations. Much pattern identification consists of lining up comparable cases to determine whether they conform to the proposed pattern. Thus students of state formation as different in approach as Reinhard Bendix, Michael Mann, and Bernard Silberman collect historical evidence on central bureaucracies in multiple countries to see whether state bureaucratization regularly occurs in the course of economic expansion. At its best, historical pattern identification facilitates recognition of recurrent social processes with sufficient accuracy to permit intervention or anticipatory action.

Scope extension applies techniques, models, or generalizations that sociologists have developed in studies of contemporary social life to historical situations. Demographers, for example, have often verified or refined models of fertility change developed initially to explain differences or short-range shifts in contemporary populations by examining longer runs of historical experience. Similarly, network analysts have sometimes extended ideas and techniques from the study of connections within currently observable populations to the reconstruction of changing connections within historical populations. Scope extension to historical cases can remain quite ahistorical or even antihistorical if the analyst disregards the time-space contexts of the social processes under investigation. But it also offers grounds for extensive collaboration between historians and sociologists,

as well as for adaptation of sociological ideas by historians. Historians have, for example, conducted valuable investigations of social mobility in past times, examining evidence and time spans rarely available to their sociological cousins.

Within disciplinary history, in fact, the special fields of social history, family history, demographic history, and economic history have grown up as interfaces with the social sciences, including sociology (Zunz 1985). In these fields, some work actually consists of sociological scope extension or its equivalent in relation to economics and political science. More takes up questions of general concern to historians by means of models, theories, or techniques grounded in social science, for example, using collective biography to trace changes in composition of the English gentry (Stone and Fawtier Stone 1984).

Scope extension has a mirror-image version. It consists of using historical evidence to challenge widespread suppositions of contemporary sociological analysis. The historical work of Viviana Zelizer offers an excellent case in point. Although her later writings about the social meanings of money balance between scope extension and process analysis, her treatments of life insurance and valuation of children in the United States during the nineteenth and twentieth centuries demonstrate that economic transactions commonly portrayed as operating independently of culture—or even to the detriment of culture—actually rest on extensive, changing webs of shared understanding. It took extensive cultural work, for example, to make the insuring of human lives acceptable to Americans. At its best, scope extension accomplishes two distinct objectives: it corrects historical misconceptions as it broadens the domain of social scientific knowledge.

Process analysis examines how social interactions impinge on each other in space and time. Instead of considering space and time as additional variables, it presumes that space-time connections define social processes, and that social processes operate differently as a function of their placement in space and time. Sociologists studying urbanization, for example, have long sought to identify recurrent sequences and geographic arrangements of urban growth (pattern identification) and to determine whether explanations of urban growth that work reasonably well in contemporary capitalist countries also apply in non-capitalist history (scope extension). But they have also tried to determine how, why, and under what conditions large cities host highly differentiated occupations (process analysis). Although theorists have sometimes drawn their ideas on the subject from biological ecology or from the comparison of contemporary populations, this sort of process analysis is intrinsically historical. The analyst has no choice but to locate the processes under examination in time and space, then to ask how time and space affect them. Process analysts typically deny the premise of pattern

identifiers: that large structures and processes repeat themselves in essentially the same form across history. Process analysts search instead for explanations of variation.

A bold example comes from Janet Abu-Lughod's study (1989) of the thirteenth century world system. Reacting to Immanuel Wallerstein's claim that under European capitalist auspices the first economic world system formed during the fifteenth century and has prevailed ever since, Abu-Lughod examined connections among such far-flung cities as Baghdad, Calicut, Canton, and Malacca three centuries earlier. She offered evidence of an active trading system, connected by both land and sea, of which Europe was then a peripheral appendage. Although Abu-Lughod's analysis centers on the same sorts of mechanisms—generation of interaction by trade in high value-for-bulk commodities—as Wallerstein's, it resets the questions. Abu-Lughod forces us to ask how and why a vast Asian-centered world system declined and a European-dominated system arose in its place. At its best, process analysis identifies causal mechanisms of broad scope as well as conditions that affect activation, interaction, and outcomes of those mechanisms (Hedström and Swedberg 1998).

Social criticism, pattern identification, scope extension, and process analysis take somewhat different approaches to historical evidence. Social criticism relies almost entirely on selective use of existing historical accounts, often privileging earlier social criticism and social thought. Thus Jürgen Habermas' *Theory of Communicative Action* reinterprets modern western history (especially that of Germany) via critiques of Karl Marx, Max Weber, Emile Durkheim, Talcott Parsons, and other thinkers, arriving at proposals for a critical theory to serve human liberation. In a more obviously historical vein, Manuel Castells' *The Information Age* reviews changing information systems and identities in the world at large during the later twentieth century in order to identify choices and challenges facing today's humanity. Castells' sources include statistical compilations and his own observations, but beyond general theory and the immediate past they center on works by historians and historical sociologists.

By and large, pattern identifiers draw their evidence from historians' syntheses, extracting comparable elements from the detail. In effect, they apply a uniform questionnaire to distinct historical episodes. Robert Wuthnow's *Communities of Discourse* (1989) asks how ideological innovations take root by comparing the Protestant Reformation, the European Enlightenment, and the formation of Marxist socialism. His implicit questionnaire's broadest headings concern environmental conditions (e.g., pace of economic growth), institutional contexts (e.g., extent of division within ruling elites), and action sequences (e.g., communication of new ideologies to broad audiences). Wuthnow draws evidence under these headings from

wide reading of relevant histories, refashioning the evidence to establish uniformities, for example, the regular emergence not of a unitary belief but of competing ideologies, from which a political selection process singles out one or a few for predominance. Such implicit questionnaires discipline inquiry, but they also run two serious risks: of treating as equivalent events, institutions, and social relations that actually differ fundamentally, and of ignoring the causal connections of events, institutions, and social relations within their actual historical settings.

Among historians, authors of textbooks and of grand narratives often proceed in a manner superficially resembling that of sociology's pattern identifiers; they incorporate other historians' detailed studies into large syntheses. Historians' syntheses, however, typically hang together not through the detection of similar elements in distinct episodes but through the establishment of connections *among* episodes. Historians' connections arise in different ways: from evolution of political or economic systems, from the impact of leaders, from transformation of the dominant culture, or perhaps from critical events such as creation of the first Chinese empires or the Russian revolutions of 1917. But on the whole historical synthesizers, unlike their sociological cousins, insist on connections in space and time.

Within historical sociology, extensions of scope often employ procedures and sources resembling those of historical researchers. Historians make a rough distinction between *primary* and *secondary* sources. If the historical processes under study directly generated the sources, they qualify as primary; administrative correspondence, triumphal arches, artisans' tools, and contemporary portraits can all serve as primary sources. Secondary sources include later efforts to reconstruct what happened, especially efforts by other historians. The distinction must remain blurred, since chronicles, autobiographies, ambassadors' dispatches to their home countries, newspaper accounts, and similar sources occupy a middle ground.

Sociological scope extenders characteristically become familiar with the character of primary sources in their chosen historical eras even when their own analyses rely chiefly on secondary work by specialist historians. Thus historical demographers have frequently compiled and analyzed registers of births, deaths, and marriages for themselves in search of clues concerning determinants of changes in fertility, mortality, or nuptiality (Willigan and Lynch 1982). As historical demography illustrates, sociologists working with a recognized body of historical sources are more likely than their historical brethren to center their analyses on models, methods, or arguments drawn directly from contemporary studies of ostensibly similar phenomena.

Not all sociological scope extenders, however, apply formal models to the past. In common with their cousins in anthropology, for example, sociologists

with historical orientations have often engaged in retrospective ethnography, taking the questions and procedures of local community studies back in time. In a pioneering work of historical sociology, *English Villagers of the Thirteenth Century* (1941), George Caspar Homans reconstructed the organization of different sorts of villages much as contemporary observers of communities were reconstructing the organization of villages in the Americas, Europe, and elsewhere. Homans substituted historical records for the interviews and first hand observation of community ethnographers. Don Kalb (1997) has adopted retrospective ethnography to follow transformations of class, family structure, and politics in industrial North Brabant during the nineteenth and twentieth centuries. Both Homans and Kalb made simultaneous contributions to sociologists' and historians' treatments of their subject matters.

Of our four varieties of historical sociology, process analyses engage sociologists most directly in thinking about history as history, that is, as the embedding of human action in time and space. Here the distinction between primary and secondary sources begins to dissolve, and the sources rarely come ready made. Reconstructing processes by which science changes, for example, engages historical sociologists of science in looking simultaneously at organizational settings, individual biographies, interpersonal networks, contested bodies of thought, and connections among all of them. No simple exportation of contemporary models into the past will accomplish that sort of synthesis.

Process analysts concentrate on connections among events, which means that, whether explicitly or implicitly, they typically construct matrices having multiple dimensions: time, place, type of event, form of connection, and so on. Wherever they find their historical evidence, then, process analysts snip, splice, paste, and rearrange it extensively. John Markoff's *Abolition of Feudalism*, a case in point, recasts local statements of grievances from the France of 1789 into regional and class distributions of demands, then integrates a vast catalog of peasant actions into an analysis of popular mobilization as a dynamic process during the early Revolution. Similarly, Roberto Franzosi pursues a search for causes of industrial conflict by threading his way among aggregate Italian strike data between 1950 and 1980, increasingly refined econometric models, and close examination of shifts in institutional and political settings. In both cases, the analysts ended up showing that when and where political processes occurred significantly affected *how* they occurred, and what results they produced.

As sociology and history now operate, relatively little flow occurs between sociologists' historical social criticism and adjacent fields of historical endeavor. The same is generally true of historical pattern identification. Neither endeavor attracts much attention from historians. Nor do sociologists respond readily to historians' efforts along similar lines. More

ideas flow in both directions across the boundaries of scope extension, but the chief effect of collaboration between sociologists and historians at that interface has been to create a series of specialized subdisciplines: historical demography, demographic history, and their equivalents. As a result, sociologists and historians at large remain woefully ignorant of each other's sources, methods, models, ideas, and discoveries.

In the zone of process analysis, however, we find intense interaction between historians and sociologists. We also find sharp disagreement. Questions of epistemology, ontology, and method align practitioners with competing answers to such questions as 'What is an event?', 'Can we detect causes in history?', 'Do all social processes result from individual choices?', and 'Do organizations really exist?' Debates over such issues roil current discussions of state formation, population change, revolution, capitalist development, and other large processes (see, e.g., Abbott 1994, 1998, Bates et al. 1998, Hamilton 1996, McDonald 1996, Steinmetz 1999). That is as it should be; where relations among historical researchers run harmoniously, we should fear that they have worked out a comfortable but ultimately fruitless division of labor. We can take some satisfaction from noticing that the deep epistemological, ontological, and methodological divisions of our time do not pit historians against sociologists, but cut across the once-forbidding frontier between history and sociology.

See also: Anthropology and History; Comparative History; Comparative Method, in Evolutionary Studies; Comte, Auguste (1798–1857); Historical Archaeology; Historical Demography; Historicism; History and the Social Sciences; Modernization and Modernity in History; Social Evolution: Overview; Spencer, Herbert (1820–1903); Theory: Sociological; Weber, Max (1864–1920)

Bibliography

- Abbott A 1994 History and sociology: the lost synthesis. In: Monkkonen E (ed.) *Engaging the Past. The Uses of History Across the Social Sciences*. Duke University Press, Durham, NC
- Abbott A 1998 The causal devolution. *Sociological Methods & Research*. 27: 148–81
- Abu-Lughod JL 1989 *Before European Hegemony: The World System*. Oxford University Press, New York
- Bates R H, Greif A, Levi M, Rosenthal J-L, Weingast B 1998 *Analytical Narratives*. Princeton University Press, Princeton, NJ
- Bendix R 1964 *Nation-Building and Citizenship: Studies of our Changing Social Order*. Wiley, New York
- Castells MI 1996–98 *The Information Age: Economy, Society, and Culture*. Blackwell, Malden, MA, 3 Vols.
- Comte A 1963 *Discours sur l'esprit positif*. Union Générale d'Éditions, Paris [First published in 1844]

- Elias N 1978, 1982 *The Civilizing Process*. Pantheon, New York, 2 Vols. [German edition first published in 1939]
- Franzosi R 1995 *The Puzzle of Strikes: Class and State Strategies in Postwar Italy*. Cambridge University Press, Cambridge, UK
- Habermas J 1984, 1987 *The Theory of Communicative Action*. Beacon, Boston, MA, 2 Vols. [German original first published in 1981]
- Hamilton R F 1996 *The Social Misconstruction of Reality*. Yale University Press, New Haven, CT
- Hedström P, Swedberg R (eds.) 1998 *Social Mechanisms. An Analytical Approach to Social Theory*. Cambridge University Press, Cambridge, UK
- Homans G C 1941 *English Villagers of the Thirteenth Century*. Harvard University Press, Cambridge, MA
- Kalb D 1997 *Expanding Class. Power and Everyday Politics in Industrial Communities, The Netherlands, 1850–1950*. Duke University Press, Durham, NC
- Mann M 1986, 1993 *The Sources of Social Power I. A History of Power from the Beginning to A.D. 1760; II. The Rise of Classes and Nation-States, 1760–1914*. Cambridge University Press, Cambridge, UK
- Markoff J 1996 *The Abolition of Feudalism: Peasants, Lords, and Legislators in the French Revolution*. Pennsylvania State University Press, University Park, PA
- Marx K 1970–72 *Capital: A Critique of Political Economy*. Lawrence & Wishart, London, 3 Vols. [German version first published in 1867–94]
- McDonald T J (ed.) 1996 *The Historic Turn in the Human Sciences*. University of Michigan Press, Ann Arbor, MI
- McNeill W H 1982 *The Pursuit of Power: Technology, Armed Force and Society Since A.D. 1000*. University of Chicago Press, Chicago
- Moore B Jr 1966 *Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World*. Beacon Press, Boston
- Moore B Jr 1978 *Injustice. The Social Bases of Obedience and Revolt*. M.E. Sharpe, White Plains, NY
- Patterson O 1991 *Freedom in the Making of Western Culture*. Basic Books, New York
- Silberman B S 1993 *Cages of Reason: The Rise of the Rational State in France, Japan, the United States, and Great Britain*. University of Chicago Press, Chicago
- Skocpol T 1979 *States and Social Revolutions. A Comparative Analysis of France, Russia, and China*. Cambridge University Press, Cambridge, UK
- Skocpol T 1992 *Protecting Soldiers and Mothers: The Political Origins of Social Policy in the United States*. Harvard University Press, Cambridge, MA
- Skocpol T (ed.) 1998 *Democracy, Revolution, and History*. Cornell University Press, Ithaca, NY
- Sorokin P A 1962 *Social and Cultural Dynamics*. Bedminster Press, New York, 4 Vols. [First published 1937–41]
- Steinmetz G (ed.) 1999 *State/Culture. State-Formation after the Cultural Turn*. Cornell University Press, Ithaca, NY
- Stone L, Fawtier Stone J C 1984 *An Open Elite? England 1540–1880*. Clarendon Press, Oxford, UK
- Therborn G 1995 *European Modernity and Beyond: The Trajectory of European Societies, 1945–2000*. Sage, London
- Tönnies F 1957 *Community and Society*. Michigan State University Press, East Lansing, MI [German version first published in 1887]
- Wallerstein I M 1974–89 *The Modern World System*. Academic Press, New York, 3 Vols.

- Weber M 1968. In: Roth G, Wittich C (eds.) *Economy and Society: an Outline of Interpretive Sociology*. Bedminster Press, New York, 3 Vols. [German version first published in 1922]
- Willigan J D, Lynch K A 1982 *Sources and Methods of Historical Demography*. Academic Press, New York
- Wuthnow R 1989 *Communities of Discourse: Ideology and Social Structure in the Reformation, the Enlightenment, and European Socialism*. Harvard University Press, Cambridge, MA
- Zelizer V A 1979 *Morals and Markets: The Development of Life Insurance in the United States*. Columbia University Press, New York
- Zelizer V A 1985 *Pricing the Priceless Child: The Changing Social Value of Children*. Basic Books, New York
- Zelizer V A 1994 *The Social Meaning of Money*. Basic Books, New York
- Zunz O (ed.) 1985 *Reliving the Past. The Worlds of Social History*. University of North Carolina Press, Chapel Hill, NC

C. Tilly

Historicism

1. Origin and Meaning of the Term

The term ‘historicism’ comes from the German Historismus, and first appeared in German as far back as the closing years of the eighteenth century. It was only imported into other European languages—French, Italian, English, Spanish, and so on—much later, from the 1930s. In German the term is to be found (albeit only occasionally) in writers such as Novalis and Friedrich Schlegel, and in the nineteenth century, in spite of having an uncertain definition, it came to be used quite widely, assuming a wide variety of meanings, most of them negative. On the whole it was used to refer to the conception of history and society worked out by the German historical school—in its writings on law and (especially) on the economy. Thus for example Carl Menger entitled one of his works ‘Die Irrtümer des Historismus,’ to criticize the reduction of economic theory to economic history, or at least the confusion of the two at the hands of the historical school of economics (see *Political Economy, History of*). A few decades later, Edmund Husserl, in his essay on ‘Die Philosophie als strenge Wissenschaft’ (1911), argued that naturalism and historicism were two parallel forms of thought which reduced philosophy to a Weltanschauung. Even when the word was used more neutrally—as occurred in the early years of the twentieth century, in writers like Ernst Troeltsch or Friedrich Meinecke or, in Italy, Benedetto Croce—its connotation was by no means univocal. It is no accident, therefore, that it was used mainly to refer to movements and theoretical approaches belonging to the preceding centuries.

This use of the word primarily to refer to the past explains why it was employed to refer to different things: (a) the conception of history present in the German historical school; (b) the positions of (often widely differing) writers such as Wilhelm Dilthey, Georg Simmel, Ernst Troeltsch, Friedrich Meinecke, and (to some extent) Max Weber; (c) the conception of history found in Hegel or Marx (this is the use Popper makes of the term); and (d) (a usage confined to Italy) a speculative tradition going back to Vico and Hegel, and leading up to Benedetto Croce. Up until the years between the wars, there were no writers who described themselves as 'historicist,' and up until that time, the word had rather negative connotations. The first attempt to provide a thoroughgoing overhaul and redefinition of historicism as a historical movement was Friedrich Meinecke's *Entstehung des Historismus* (1936, trans. *Historicism*, London 1972), a work which, however, produced more criticism than consensus (see *Historiography and Historical Thought: Modern History (Since the Eighteenth Century)*).

2. Historicism and the Historical School

There is no doubt, however, that historicism is not just a historiographical tendency—even though it has certainly had widespread influence on historians—for it is primarily a general conception of reality, or at least of Man and the human world, a conception whose roots are to be found in late eighteenth-century German culture. (The inclusion of Vico as a historicist is a retrospective operation, and one that is highly controversial.) Historicism in fact emerged in close connection with that reaction against French culture that started to make its voice heard clearly with Johann Gottfried Herder. In *Auch eine Philosophie der Geschichte zur Bildung der Menschheit* (1774), Herder argued against views of history which saw it as the progress of humanity out of a savage state towards civilization, but espoused a developmental perspective which maintained there was an analogy between the life of humankind and the life of an individual. According to Herder the various epochs that came after one another—seen as paralleling the gradual passage from childhood to youth and then virile manhood—correspond to the succession of peoples. Each historical epoch was said to bear the imprint of the national character of a particular people, which was destined, in the same way as humanity as a whole, to run through a life cycle. Herder linked up this organicist perspective with the idea that history was the working out of the design of Providence. Thus each people was seen as carrying out an indispensable task in the history of humanity, imparting in the process the mark of its peculiar nature—its 'spirit'. The idea of 'a people's spirit' (*Volkgeist*) therefore came to mean the essence of a particular people, the

fundamental traits which underlay and united all the forms that people took in the course of its development.

It was precisely this idea which was taken up by the historical school. Whereas Hegelian philosophy of history developed Herder's providential perspective, and tried to show the essential rationality of history, demonstrating that underneath the apparently accidental succession of events was the development of spirit, Friedrich Carl von Savigny interpreted law as the individual creation of a particular people—thus rejecting the natural law tradition. In a similar way, Wilhelm Roscher and other exponents of the historical school among economists saw economic life as the expression of a particular people, and thus a pattern which could not be understood independently from other aspects of that people. Most influentially of all, Leopold von Ranke gave a theoretical framework to romantic historicism in a series of works—from *Geschichten der romanischen und germanischen Völker von 1494 bis 1535* (1824) to his unfinished *Weltgeschichte* (1881–5). Rather than seeing each people as having an essence, Ranke saw each epoch as possessing its own individual character. He believed each age was characterized by dominant 'ideal' tendencies, and that the people who lived in a particular age were shaped by these tendencies. He denied that it was possible to interpret the succession of epochs as having any teleological development, and denied that earlier ages could be seen as laying the way for later ones, or later ones the necessary outcomes of earlier epochs. Each epoch has its own value, and is, as Ranke put it, in an immediate relationship with God. The divine presence in history did not imply, therefore (as Hegel believed) that there was a 'progressive' evolution from one age to another.

A common thread in Romantic historicism was the belief that all knowledge about Man, society and the state must be historically based. Historicism thus put forward an alternative approach to that which had become dominant in the second half of the eighteenth century—exemplified in writers such as Montesquieu or Adam Smith—which took modern natural science as the pattern to follow. Its epistemological model not only implied that historical knowledge was different from knowledge of natural phenomena, but also insisted on the need to bring all disciplines studying Man back to a historical basis. This meant that knowledge of this kind was inevitably 'partial', for it took objects—economic life, say, or law, or the organization of the state—which could only be abstractly isolated from the historical process. Only historical knowledge could lead to understanding of the historical process, it was claimed, or understanding of any epoch in its totality. And only historical knowledge could reveal the deep underlying meaning of events and of the relationship between them. The historicism of the historical school thus remained faithful to the organicist framework it inherited from

Herder. Indeed, it ensured that a shadow of divinity fell upon history—since it was only God who could give a sense to humanity's journey.

3. Modern Historicism

In the last two decades of the nineteenth century a decisive turn took place in historicism. There was an encounter with the neo-Kantian movement and it was felt that what were called the 'sciences of the spirit' or 'sciences of culture' needed to be given an epistemological foundation. This turn started with Dilthey's *Einleitung in die Geisteswissenschaften* (1883). Dilthey positioned himself in the tradition of the historical school, but he abandoned its organicist assumptions. Still more importantly, he rejected any attempt to use history as the basis for a metaphysical perspective which saw history as containing a *deus ex machina*—whether this consisted of a transcendent will (as in Christian ideas) or an immanent one (as in Hegel). Dilthey tried to apply the kind of critique that Kant had made of natural science to the sciences of the spirit. He therefore thought it was necessary to supplement the critique of pure reason with a 'critique of historical reason'—a critique which would 'provide foundations' for the disciplines concerned with Man, society and the state. However, this enlargement of the critical enterprise demanded a shift in approach, according to Dilthey, to permit a non-metaphysical analysis of the historical world.

A kind of historicism thus began to emerge which was different from, and in many ways opposed to, that of the Romantic period. This new historicism was active in two fields—in epistemology and methodology, and in analysis of the structure of the historical world. In both fields work was produced which was often very diverse. Dilthey was more tied than were other new historicist writers to the historical school, and he saw knowledge of history as part of the *Geisteswissenschaften* (seen as including psychology and the various social sciences as well as history). He thus tried to identify what it was about knowledge of this kind that made it different from knowledge in the natural sciences. He saw the natural and human sciences as differing both in their object and in their method. Whereas natural sciences took as their object phenomena which belonged to a world outside human beings, and were rooted in external experience, in the sciences of the spirit the knowing subject and the object belonged to the same world; knowledge thus came from the lived experience knowing subjects had of their own inner states. Dilthey took up Droysen's idea that historical knowledge was 'understanding' (*Verstehen*) rather than explanation (the kind of knowledge which could be achieved in the natural sciences). Other writers, however, such as Wilhelm Windelband and his pupil Heinrich Rickert, who were more directly influenced by the neo-Kantian

movement, took up Ranke's idea that history is a search for individual identities (the identity of an age) and transposed it into the theory that history was a kind of knowledge oriented towards the individual—as against natural science, which is the search for general laws. Still others, such as Simmel, focused on the social sciences and especially sociology, and posed a problem which was analogous to Kant's problem regarding how nature was possible—asking how society was possible.

These various positions are related to different approaches to the structure of the world of history. Dilthey developed the critique of 'historical reason' into a historical critique of reason, and thus came to a theory of historicity which later influenced Martin Heidegger's *Sein und Zeit*. The historical world—that is to say, systems of culture, systems of social organization, historical epochs—was seen by Dilthey as the outcome of a process whereby life was objectified (became 'objective spirit'). Yet life as an attribute specific to human beings was itself a process which evolved in time. Man is thus a historical being, and his life is historically conditioned by the tendencies of a particular epoch. This led Dilthey to deny that there could be any absolute values—a position which was quite different from that of Windelband and Rickert, who took the neo-Kantian movement in the direction of a theory of values. Dilthey's denial of the possibility of absolute values was echoed on the other hand by Simmel, who produced a psychological interpretation of the Kantian *a priori* and asserted that knowledge was just as relative as other forms of life. After 1910 Simmel came to espouse an immanent-type philosophy of life, founded on the dialectic between life, 'more-life' and 'more-than-life.'

In the period between the two world wars, German historicism again changed track and returned to the metaphysical perspectives of Romantic historicism. The work of Oswald Spengler, published just after the First World War, stressed the radically relativistic implications of historicism, and portrayed history as the terrain in which heterogeneous cultures played themselves out—each one carrying its own set of values, which were peculiar only to the people living within the particular culture in question. Spengler was influenced by Goethe and Nietzsche: from Goethe he adopted the idea of the opposition between the 'world of nature' and 'the world of history,' and also the analogy between cultures and organisms, and he drew on Nietzsche to make a distinction between culture and civilization (*Zivilisation*), with civilization being seen as the inevitable decline of any culture (see *Civilization, Concept and History of*). Troeltsch and Meinecke opposed the relativism of Spengler and tried to re-assert the absolute nature of values from within the historical process itself. Troeltsch advocated a 'material' philosophy of history that would create a 'cultural synthesis' between the various national components of European civilization. Meinecke went

back to Ranke's work and investigated the relationship between the structure of the historical process and God. For both writers, the fact that values were situated in history did not imply that they could not be transcendent, and thus escape becoming outmoded by the course of historical development or becoming.

4. *Beyond Historicism*

In the first two decades of the twentieth century Max Weber took another path, going 'beyond historicism' as Wolfgang Mommsen put it (see *Weber, Max (1864–1920)*). Weber broke decisively with Romantic historicism, took up Menger's critique of the historical school in economics and rejected the organicist assumptions of so much historicism. Weber did not draw on Dilthey—who was still linked to the historical school—but rather on Rickert's distinction between historical knowledge and natural science. For Weber, however, historical knowledge consisted of a set of disciplines that needed not only the explanation of historical phenomena but also what he called nomological knowledge. The focus on individual phenomena therefore in no way meant that general concepts and laws could not be used: these are, indeed, essential if events and the relations between events are to be explained. These general concepts and laws differed from those of natural science in that they took the form of ideal-types. However it was they that made up the basis of 'pure' economics and sociology—these being seen as generalizing disciplines which tended to distinguish them from historical research proper. Weber's position was similar to, but more rigorous than, Simmel's. Both wished to restore the relationship with the social sciences which the emphasis on the individual nature of historical phenomena (the inheritance of Romantic historicism) seemed to imply was impossible.

Weber also broke with Rickert's theory of values on another terrain—that of the interpretation of the status of values. Rickert had sought to ground historical knowledge in universal values, which were transcendent and independent of the course of history. For Weber, on the other hand, values were the criteria for choosing from the infinite range of empirical data—and thus a criterion whereby a historical object was constituted. Far from being absolute, therefore, values possessed a validity that coincided with the possibility of their finding concrete form in human action. Weber also denied that values were organized in systems, as Rickert believed, for values are multiple, as are the spheres of values. Between the various value spheres, and between individual values within each sphere, there was 'collision' of values. This is true even within the sphere of ethics, which is dominated by the split between ethics of intention or conviction (*Gesinnungsethik*) and ethics of responsibility (*Verantwortungsethik*). Given that there is a plurality of conflicting values, Weber argued

that people have to choose; and choosing certain values inevitably means rejecting others. For Weber, history was thus the stage on which people positioned themselves vis-à-vis values, adopting certain values as criteria in their own action, thus putting them into effect.

5. *Historicism outside German culture*

Like that of Troeltsch and Meinecke, Weber's thought was cast in the framework of modern historicism. This was not the case for Croce, who was in intellectual dialogue with Hegel's philosophy of history—although he also put forward a 'reform' of the Hegelian dialectic. Croce denied the existence of a nature existing outside spirit—so even ideas are included in spirit. So reality as a whole was identified with spirit—and this, in turn, was identified with history. In epistemological terms, knowledge was seen as historical knowledge. The sciences (whether abstract sciences such as mathematics or political economy, or empirical sciences, such as the natural sciences) were denied any cognitive validity: Croce argued they simply had an 'economic' function. Croce termed his thought 'absolute historicism'—that is to say, it asserted the thoroughly historical character of reality and knowledge as a whole. Croce's thought was absolute in another, perhaps deeper, sense. For Croce saw history as the 'unfolding' of a first principle, spirit, which becomes real by moving from one to another of its eternal forms—whether art, logic, economic activity or ethical action. He saw individuals as—at best—simple instruments of these processes. History for Croce was about spirit, not about individuals, even though it did deal with individual's 'works'.

Croce worked out the principles of his 'absolute historicism' in the years following the late 1920s—so at a time when modern German historicism had virtually come to an end. The distance between the two types of historicism became clear in the debate which followed publication of Meinecke's *Entstehung des Historismus*. This debate made it clear that Croce and the German historicists thought of historicism in different terms, and also had differing interpretations of the history of historicism. Croce saw his historicism as opposing not only the 'abstract rationalism' of Enlightenment culture, but also the irrationalism which he saw as being present in the historical school, and also in Meinecke. Meinecke saw historicism as originating in the passage from the natural law tradition to Romantic thought, and thought of its greatest triumphs as being the works of Goethe and Ranke. Croce, in contrast, was critical of Ranke's approach to history, which he saw as being unproblematic, and he imagined a direct line of inheritance running from Hegel to his 'absolute historicism' (although the genealogy was complicated by Vico—the national progenitor).

In this same period, historicism (especially German historicism) started to spread to the Latin and the English-speaking countries. This produced some significant accounts and re-statements, e.g., at the hands of José Ortega y Gasset in Spain, or Raymond Aron in France. Croce's historicism (although not fully distinguished from Dilthey's) found an echo in Britain, in Robin G. Collingwood's work. However, by this time, historicism was almost at an end. In 1944 Karl Popper (see *Popper, Karl Raimund* (1902–94)) published a book with the significant title, *The Poverty of Historicism*, which took up the criticisms which neo-positivists (particularly Otto Neurath) had made of the traditional distinction between historical knowledge and natural science. After the war historicism was subjected to head-on attack from various sides (e.g., in the critiques of Leo Strauss and Karl Löwith). In the second half of the twentieth century it has been rare for writers to admit to describe their philosophical positions as 'historicist', even though some have taken up themes and forms of analysis which are similar.

6. From Historicism to 'Historical' Social Science

At the same time as its influence in philosophy was declining, historicism became important again for historians, especially in Germany. Central was the reference to Max Weber and his insistence on the need to link historical research and the social sciences, and on the need for history to employ concepts and theories developed by the social sciences. Instead of making the traditional distinction (and contraposition) between historical disciplines and the natural sciences, or that between historical understanding and 'systematic' knowledge of society, Weber favored a strong link between the latter two kinds of knowledge; this approach was taken up by historians such as Hans-Ulrich Wehler, Jürgen Kocka, Wolfgang J. Mommsen and others. In a 1969 essay Wehler proposed a definition of history as 'historical social science' and argued that history ought to be seen as one of the social sciences, yet at the same time have its specificity recognized. He thus emphasized historiography's affinity to economics, sociology or political science, in so far as it necessarily drew on theories, and took most of them from other social sciences.

One of the consequences of this current of thought was that it opened the way for a return to Marxism: not Marxism as a general theory of history, but as a set of theories which could be tested in the interpretation of specific historical processes. In other words, Marxist concepts were used in a 'Weberian' fashion, as ideal types. Theories thus had a purely functional value and could not guarantee the validity of research. This brought out the problem (stressed especially by Kocka) of how to check the findings which theoretically guided research led to, and the related problem of

how the validity of theories could be assessed. Kocka's solution was that it is not theories which legitimize findings, but rather findings which demonstrate the heuristic efficaciousness of theories.

These works left aside many of historicism's classic problems, and rejected many of its methodological stances, such as its opposition between explanation and understanding. Their aim was to show that historical writing had its scientificity, albeit scientificity of a special, 'sui generis' kind. They had nothing in common with the revival of Dilthey in hermeneutics championed by Hans-Georg Gadamer. And they were firmly opposed to the notion of history as 'narrative,' proposed by Hayden White and many other American scholars (in work that, not surprisingly, also often embraced a hermeneutic perspective) (see *History: Theories and Methods; Bourgeoisie/Middle Classes, History of*). Nonetheless, this did not stop writers like Jörn Rüsen recognizing the narrative character of historical writing, and seeing this as the feature which marks history out from other social sciences. But Rüsen does not deny either the need for theories nor the scientific character of historical research: on the contrary, he aims to show that the narrative structure of historiography is perfectly compatible with the use of general theories, and does not absolve historians from the duty to empirically check the results of their research.

Historicism has thus not so much been extended as 'superseded.' For not only does 'historical social science' reject Ranke's ideal that the historian should be purely 'objective'. It goes so far as to claim that history can and should be critical (or 'emancipating' in Habermas' terms): that is to say, critical vis-à-vis traditional history-writing and its methodological principles, including the separation between history of ideas and social history, but also critical towards traditional historiographical interpretations of Germany's past. This has led to a significant shift in attitudes towards the Enlightenment (see *Enlightenment*). As we have seen, historicism was born partly out of polemic with Enlightenment culture, which it considered anti-historical. In contrast, Horst Walter Blanke has recently tried to link 'historical social science' up to Enlightenment historiography and the Enlightenment conception of history. This marks a sharp break with the traditional conceptions and categories of historicism—especially (but not exclusively) those of Romantic historicism. Thus, historicism did not only decline as a general philosophical approach, but also as a paradigm of historical research and writing.

See also: Hegel, Georg Wilhelm Friedrich (1770–1831); Historiography and Historical Thought: Current Trends; History and the Social Sciences; History: Overview; Simmel, Georg (1858–1918); Weberian Social Thought, History Of

Bibliography

- Antoni C 1959 *From History to Sociology. The Transition in German Historical Thought*. Wayne State University Press, Detroit, MI
- Aron R 1938 *Essai sur la théorie de l'histoire dans l'Allemagne contemporaine*. J Vrin, Paris (2nd edn., 1950 entitled *La philosophie critique de l'histoire. Essai sur une théorie allemande de l'histoire*)
- Blanke H W 1991 *Historiographieggeschichte als Historik*. Fromman-Holzboog, Stuttgart-Bad Cannstatt, Germany
- Engel-Janosi F 1944 *The Growth of German Historicism*. Johns Hopkins University Press, Baltimore
- Hughes H S 1958 *Consciousness and Society. The Reorientation of European Social Thought, 1890–1930*, 1st edn. Knopf, New York
- Iggers G G 1968 *The German Conception of Historical Thought from Herder to the Present. The National Tradition of History*. Wesleyan University Press, Middletown, CT (orig. edn. *Deutsche Geschichtswissenschaft*, DTV, Munich, Germany)
- Iggers G G 1984 *New Directions in European Historiography*, rev. edn. Wesleyan University Press, Middletown, CT
- Kon I S 1964 *Die Geschichtsphilosophie des 20. Jahrhunderts—Kritischer Abriss*. Akademie Verlag, Berlin
- Lee D E, Beck R N 1953–4 The meaning of historicism. *American Historical Review* 59: 568–77
- Rand C G 1964 Two meanings of historicism in the writings of Dilthey, Troeltsch and Meinecke. *Journal of the History of Ideas* 25: 503–18
- Rossi P 1956 *Lo storicismo tedesco contemporaneo*. Einaudi, Turin, Italy (new edn., 1994 *Edizioni di Comunità*, Milan)
- Rossi P 1960 *Storia e storicismo nella filosofia contemporanea*. Lerici, Milan (new edn., 1991, *Il Saggiatore*, Milan)
- Rossi P 1989 'Historismus' und 'storicismo': zwei Denktraditionen. In: Esch A, Petersen J (eds.) *Geschichte und Geschichtswissenschaft in der Kultur Italiens und Deutschland*. Max Niemeyer, Tübingen, Germany
- Rothaker E 1960 Das Wort Historismus. *Zeitschrift für deutsche Wortforschung* 16: 3–6
- Rüsen J 1993 *Konfigurationen des Historismus*. Suhrkamp, Frankfurt, Germany
- Tessitore F 1991 *Introduzione allo storicismo*. Laterza, Rome and Bari, Italy
- Wittkau A 1992 *Historismus*. Vandenhoeck & Ruprecht, Göttingen, Germany

P. Rossi

Historiography and Historical Thought: Christian Tradition

Ever since the beginnings of orderly reflections on the topic of history in Christianity and the first literary attempts to write history (by the author of *The Gospel According to St. Luke*, c. 81–96 AD, cf. Luke 1, 1–4 and Eusebius of Caesarea, c. 264–340 AD, *Ecclesiastical History*), four basic problems have been characteristic of all such attempts. Their uninterrupted

significance from the ancient world up to the present allows development of a history of problem areas of Christian historical thought. All Christian historical thought and also all Christian historiography can be characterized by basic decisions concerning the following four problem areas: (a) the classification as a specific inside or outside perspective on Christianity (Sect. 1); (b) the respective understanding of the position of one's interpretation on a range between historical and theological standards (Sect. 2); (c) specific mechanisms of selection (Sect. 3), and (d) a small number of models by which the past is reconstructed see (Sect. 4).

1. The Problem of Inside Perspective and Outside Perspective

A Christian understanding of history describes Christianity from an inside perspective; but in many cases such descriptions can, to some extent be regarded as outside perspectives. This is because, together with an outside perspective on Christianity, e.g., by non-Christians, there are also outside perspectives of Christian minorities on Christian majorities, and vice versa. Such perspectives have a special forming influence on the understanding of history: The *Ecclesiastical History* by Theodoret of Kyrrhos (c. 393–466 AD), one of the most definitive works of Christian historians of the second generation after Eusebius, regards historical developments as formed by the debates between orthodox church ('orthodoxy') and apostate heresy. The *Ecclesiastical History* by his contemporary Philostorgius (c. 368–425 AD) has the same perspective, however with the opposite judgment: the theological tradition of the Alexandrine presbyter Arius (c. 260–336 AD), which has influenced theological thought mainly in Great Britain up to the present, is regarded as true orthodoxy by Philostorgius but as apostate heresy by Theodoret. Besides these, there have always existed more or less critical descriptions from an outside perspective by non-Christians (e.g., by the Roman historian Ammianus Marcellinus, c. 330–395 AD, in the heyday of Christian historiography in late antiquity) in which Christian historians were tackled implicitly or explicitly.

It was only in modern Europe that Christian historians have made determined attempts towards a reduction of inside-perspective elements in Christian historiography. This is, of course, mainly with respect to the narrow horizon of thought formed according to the self-awareness of the three Christian denominations that arose in Central Europe after the religious wars of the sixteenth and seventeenth centuries (the Religious Peace of Augsburg 1555 and the Peace of Westphalia 1648). At the beginning of the modern era, reflecting denominational antagonism, the Lutheran, Reformed, and Roman Catholic interpretations of

history entered into competition. The protagonist of the overcoming of this denominational approach to the history of Christianity was Gottfried Arnold (1666–1707; Professor of History in Giessen 1697–1698). He tried to write a ‘*unpartheiische*,’ which is a non-denominational church history. In the twentieth century, following the institutionalization of the ecumenical movement, which has tried to overcome the denominational splittings of the sixteenth century, the call for a general ‘ecumenical church history’ (Ernst Benz) arose. In a widespread late twentieth-century attempt to conceptualize the alternatives, ‘church history’ as opposed to ‘history of Christianity’ has been used. ‘Church history’ describes a perspective orientated towards the Christian denominations, and ‘history of Christianity’ describes a trans-denominational perspective.

2. Christian Interpretation of History—Historical and Theological Standards

A Christian interpretation of history experiences a tension between historiography (or history) on the one hand and Christian theology on the other. This tension has existed from the very beginning in so far as theology (which in this context means the theory of the Christian faith) naturally favors certain theoretical options when reconstructing a history, for example, its assumption of a god acting in history, which other interpretations of history may hold but are by no means bound to hold. Such specific theoretical options always have methodical consequences; this is what the religious philosopher Ernst Troeltsch (1865–1923) showed in his clear analysis and what has been called ‘historical or dogmatic method’ (Troeltsch 1922, pp. 729–753). But one must not overestimate the ‘theological’ elements of Christian interpretations of history before the modern era: from the fourth century they had as a matter of course been orientated towards the literal and methodical standards of the interpretation of history of their respective times. This theoretical background was broken only very partially by specific religious or theological concepts. The first author of an ecclesiastical history, Eusebius of Caesarea, already followed the forms and standards of the historiography of his time in his historical writing, as can be seen in the prologue (Eusebius, *historia ecclesiastica* I 1,3f). Consequently, his account leaves out central questions of a theological interpretation of history. Because pagan ancient historiography very rarely explicitly treated the difficult relation between divine acting in history through fate (*θεοτεῖχη/theoteiche* or *fatum*) and human freedom, Eusebius, too, never picked out the exact relation between man’s freedom and God’s acting in history as a central theme.

The fundamental problem of the relation between an interpretation of history according to modern

standards on the one hand and specific theological premises on the other hand intensified in modern Europe. In the course of the modern secularization of historical thought, under the premise of *etsi Deus non daretur* (God is not necessary), new determinants of history (such as fate, chance, reasons of state, and ethics) were established in place of the Christian approach to a history regarded as being under the law of God. In consequence, the problem of how—under these conditions—a theological view of history was possible, increased. Most Christian interpretations of history after the Enlightenment have, however, tried to take up the problem, often through a religious or theological interpretation of the philosophical framework of the historiographic theory. The Berlin theologian Friedrich Schleiermacher (1768–1834) for example made the following point in his notes of the years 1796–1799: ‘*Die Historie ist immer religiös und die Religion muß ihrer Natur nach historisch sein*’ (‘history is always religious and religion must be historical by nature,’ Schleiermacher 1984, p. 25). At the same time the possibility of a Christian interpretation of the world, which was lost during the period of Enlightenment, was meant to be restored by means of such a religious charging up of history.

3. The Mechanisms of Selection of Christian Interpretation of History

All interpretation of history and, furthermore, all historical writing, is characterized by acts of selection from a huge amount of material. Within the framework of a Christian interpretation of history, first the basic choice of subject must be made: to date, it has been heavily disputed whether a Christian interpretation of history merely ought to deal with certain fields of general history such as Christianity, the history of certain denominations or ecclesiastical groups, or a particular history formed in a particular way by a transcendental power (‘*Heilsgeschichte*’, ‘salvation history’), or whether a Christian interpretation of history ought rather to come up with a distinctive Christian perspective on history as a whole called ‘world history’ since the end of the eighteenth century. The controversies over whether the chosen ideological perspective necessarily leads to a concentration on an ecclesiastical history, or on a history of Christianity, or alternatively on the wider framework of a universal history, has always been an issue in Christian historical writing.

The first Christian author of a historical monograph, Eusebius, chose to concentrate on an ecclesiastical history without any detailed consideration of the political history of his time. A universal history, now lost, was presented by the Greek presbyter Philippus of Side less than a hundred years later, before AD 439. The Christian interpretation of history

that has had the strongest influence up to the modern era, and which was fundamental in the middle ages, was written by the North-African theologian Augustine (AD 354–430). It takes the perspective of a universal history: after the city of Rome had been captured by the Goths in AD 410, both pagan and Christian interpretations of history, which until then had their focus in the Empire, broke down because of the obvious collapse of the Roman Empire. As a consequence, the bishop was bound to develop his own distinctive concept of a universal history which would be able to convey a new sense of history. Augustine presented it with his work *De Civitate Dei*, 'The City of God' (AD 413–426; for details see Sect. 4).

While the Middle Ages mainly expressed concepts of universal history based on Augustine's, e.g., by Otto of Freising (c. 1112–1158), the traditional Christian universal—historical interpretation broke down during the Enlightenment. Theological categories, such as 'God's revelation' or 'God's acting', which had been systematic premises of the Christian interpretation of a universal history, and had consequently formed a unity of transcendental and earthly reality, became problematic. After the attempts of the nineteenth century to recreate such a concept of unity on a new basis, and thus to once more present a Christian interpretation of a universal history (in very different ways by Schleiermacher, Hegel, and Ranke), the twentieth century progressively withdrew from such theological interpretations of history. Christian interpretation and presentation of history in the twentieth century were more or less confined to the specific context of Christianity and church, even if the theoretical basis was sometimes broader. That was the systematic reason for the modern criticism of concepts which attempted to create a close correlation between the history of human development and God's acts of salvation, by way of 'salvation history', the theological interpretation of certain incidents.

Along with the fundamental question of whether a Christian interpretation of history must be understood universally, or be restricted to a history of Christian institutions, there are other mechanisms of selection taken from general historiography which are of importance to a Christian interpretation of history. Historians of the ancient world and the middle ages used, as important mechanisms of selection, the questions of the meaning of an incident, of its benefit for future generations, or of its documentation in the respective privileged sources, and so forth. The bias thus created is obvious. However, in general, the critical sense with which historians before the modern era treated their own modes of selection is underestimated. Such reflections were usually made in the prologue of historiographical works.

The vehement calling into question of traditional mechanisms of selection in the general history of the twentieth century (see e.g., the criticism of any

historical interest with a biased concept of histories of ideas or politics, the criticism of the concentration on Europe, and of the neglect of feminist research) reached the historiography of Christian denominations with a certain delay; its consequences on a general Christian historical awareness only gradually become apparent.

4. The Basic Models of a Christian Interpretation of History

Christian conceptions of history up to the modern era usually have reconstructed it on the basis of a small number of comparatively simple models. Eusebius describes the history of the church according to a model of a process of constant 'advances,' something which is in strict contrast to the historiography of his times with its stress on cyclic processes and the decadence of traditional Roman values. The history of his own religion teleologically leads to the Christian empire of Constantine. Augustine breaks with Eusebius' model of a progress with his concept of two cities (*civitates*) wandering through the ages. He differentiates between the *civitas Dei*, the 'citizenry of God' as an earthly community of the heavenly citizens, and the *civitas diaboli*, 'the citizenry of the devil.' Of course, the Roman state and the community of those who do not live under God and towards him are not identical in the sense of the state as such being the *civitas diaboli*. But Augustine takes up the traditions of radical Christian criticism of the state, and with his anthropologically based historiographic model gives a clear analysis of the social deficiencies of his time. The European Reformation breaks with this conception. It mainly follows the models of decadence developed by the medieval church reformers, which can also be found with humanistic authors: the promising beginnings of the first Christians had decayed in a middle period which are only now brought to light again. This conception was monumental, and had a strong anti-Catholic impetus in the so-called 'Magdeburger Zenturien'—a team effort by protestant theologians (Basel 1559–1574) led by the strict Lutheran Matthias Flacius Illyricus (1520–1575), which has been known under that name since the eighteenth century, and which has remained a fragment. Along with a comparatively simple modeling, the work is characterized by the use of an enormous number of sources. Cardinal Caesar Baronius (1538–1607) reacted to this challenge with his *Annales Ecclesiastici*. But this catholic reply, in terms of method only, mirrored the protestant work, and it tried to answer the reproach of having broken with the truth with a proof of uninterrupted continuity from the age of the early Christian church.

The model of decadence gained special influence (e.g., in Goethe's work); Arnold applied as a basis—despite his programmatic orientation towards an

impartial standpoint. Arnold explained the history of Christianity according to a model of progressive decadence, beginning after the pure earliest church. Despite his persistent interest in counterevidence to official protestant ecclesiastical historiography, he held to his conception of decadence, and to a structural principle of the denominational history of Christianity. Such simple basic models for the reconstruction of the history of Christianity were by no means, or first, only questioned by the European Enlightenment. Augustine's criticism of Eusebius' conception of progress in *De Civitate Dei* is an example, because it was motivated by an empirical diagnosis of society and by critical work on biblical texts. Humanist historians of the Reformation must also be mentioned, particularly Philipp Melanchthon (1497–1560) and Johannes Sleidanus (1505–1556). In his short world history, and in a work on contemporary history entitled *Denkwürdigkeiten, Betreffend Staat und Religion Unter Kaiser Karl V*, strict models of decadence are missing. Sleidanus always endeavored to describe religious changes with regard to imperial power and state power; he can be regarded as one of the progenitors of a strong orientation towards political history.

Along with the orientation towards simple conceptions of progress or decadence, the history of Christian historiography, and the Christian awareness of history, has called into question over-simplified models up to the present. In Ranke's formula, for example, each epoch is immediate to God; divisions into periods are good indicators for the respective underlying models. In this respect, denominational Christian interpretation of history tends to differ from the standards of general history; protestant conceptions for example speak of a 'period of Reformation' as a threshold between the middle ages and the modern era.

The four basic problems outlined above have characterized Christian concepts of history from its beginnings, and they show that the crucial changes in Christian interpretation of history, caused by the increasing influence of the methodical standards of modern historiography and finally by the victory of historicism, must not be overlooked in theology, either. The victory of historicism has had a significant consequence on academic theology as a whole, because the historicization of all theological disciplines was demanded (among others by Troeltsch), and was partially carried out. It is equally clear and almost undisputed that the general 'crisis of historicism' also had an impact on theological form of Christian interpretation of history, something which had already been diagnosed by contemporary church historians like Karl Heussi (1877–1961). Mainly in continental Protestant theology, this general crisis was accompanied by a theological criticism of general historicism. This development is now called an 'anti-historical revolution' (Graf 1988, pp. 377–405), a term which points to its consequences: the removal of

church history as a leading subject of academic theology.

In the twentieth century there was an interesting revitalization of apocalyptic thought on a large scale, the relation of which to the 'crisis of historicism' is not immediately clear. This form of Christian theology, which goes back to a movement in Hellenistic Judaism, and which had already left deep marks on certain biblical texts, is characterized by a distinctive awareness of history. The present time is interpreted as the 'last age,' and forms part of a certain fixed scheme of the course of history, whilst there is no historical interest in the identification of certain 'signs of the times.' In the ancient world, the formation of a Christian awareness of history was accompanied by a drastic loss of meaning of apocalyptic elements for Christianity. The eruptive revival of apocalyptic trends in the middle ages, the early modern era, and the present has in most cases been accompanied by a protest against Christian interpretations of history orientated towards the standards of the respective time.

5. Current Problems and Questions of Christian Interpretation of History

The systematization of current schools of Christian interpretation of history meets with the difficulty of the actual presence of pluralism (Nowak 1997, pp. 3–12), perceived a 'new confusion.' Such concepts are present in the institutional fields of denominational church history as well as independent of these. The Christian inside perspective, considered to embody prejudice in some accounts, presents a challenge to other approaches to the problems of Christian historicism; they have not, however, discarded it. This shows the limited significance of the presently popular alternatives of a 'church history' and a 'history of Christianity.' In both cases, attempts to take an outside perspective are, more or less admittedly, wrapped in an inside perspective.

Within the horizon of the standards of history at the beginning of the third millennium on the one hand, and of Christian theology on the other, almost all schools have tried to achieve careful de-theologization. This, however, is rarely picked out as a central theme. It is only very occasionally that a decidedly theological contribution to the general discussion on history is demanded (Herms 1997, pp. 305–30); a 'resort to pragmatism' can be noted (Nowak 1997, p. 12). Even if some historians still demand a concept of universal history in Christian terms as regards the selection of subject matter, most publications actually present contributions to a discipline of general history. That such presentations are in most cases characterized by the leading assumptions of universal history (and theology) is obvious to any sensitive examination; but

it is only rarely admitted or made explicit. The presently vigorous call for an approach to church history, or history of Christianity, as historical-cultural studies also implies certain universal-historical structures in the shape of an '*histoire totale*'.

The critical calling into question of historiographical mechanisms of selection and of modeling, which mainly arose in the twentieth century, has increasingly been accepted by the institutionalized Christian interpretation of history. But it is very rare that it is understood as one of the consequences of fundamental decisions made in the course of the European Reformation movement which have created a high level of ideology-critical potential.

These four basic problems of Christian interpretations of history cannot be solved by means of isolated special theological hermeneutics. They call for an original contribution to the general discussion of history and of the interpretation of history. This is especially true in view of the latest developments in 'writing history in the postmodern era': can the concept of a literal sense as the regulatory concept of the process of interpretation really be given up in the course of the deconstruction of its own narrative? does not a Christian interpretation of history which does away with the 'grand narratives', and the 'master processes' dissolve itself? is the notion of 'God's acting' dispensable in a Christian interpretation of history? Here, a new and interesting challenge is put to the long-neglected field of a theory of Christian interpretation of history. It is possible that a theoretically renewed, clear-cut, Christian interpretation of history helps to preserve in a positive sense two functions of the Christian interpretation of history under the conditions of the postmodern era: the function of forming society, and the function of critically analyzing society. But with all this, the traditional objections to a simplistically designed history of salvation remain important.

See also: Historiography and Historical Thought: Classical Period (Especially Greece and Rome); Historiography and Historical Thought: Modern History (Since the Eighteenth Century); History: Theories and Methods

Bibliography

- Graf F W 1988 Die "antihistorische Revolution" in der protestantischen Theologie der zwanziger Jahre. In: Rohls J, Wenz G (eds.) *Vernunft des Glaubens. Festschrift W. Pannenberg*. Vandenhoeck & Ruprecht, Göttingen, Germany, pp. 377–405
- Hermes E 1997 Theologische Geschichtsschreibung. *Kirchliche Zeitgeschichte* 10: 305–30
- Markschies C 1995 *Arbeitsbuch Kirchengeschichte*. JCB Mohr, Tübingen

- Markschies C 1998 Kirchengeschichte. In: Goertz H-J (ed.) *Geschichte. Ein Grundkurs*. Rowohlt, Reinbek bei Hamburg, Germany, pp. 408–22
- Meinhold P 1967 *Geschichte der kirchlichen Historiographie*. Karl Alber, Freiburg, Germany
- Nowak K 1997 Wie theologisch ist die Kirchengeschichte? *Theologische Literaturzeitung* 122: 3–12
- Schleiermacher F 1984 *Schriften aus der Berliner Zeit*. Meckenstock G (ed). W. de Gruyter, Berlin
- Troeltsch E 1922 Ueber historische und dogmatische Methode in der Theologie. In: *Zur religiösen Lage, Religionsphilosophie und Ethik*. JCB Mohr, Tübingen, pp. 729–53

C. Markschies

Historiography and Historical Thought: Classical Period (especially Greece and Rome)

Historiography, the rational reconstruction of the past in the medium of a prose that claims to literary quality, was developed in Greece during the fifth century BC. Its emergence reflects the peculiar intellectual and political developments of this age. The high cultures of Egypt and Mesopotamia had known lists of rulers, their foundations, festivals, and wars that represented the official version according to which the monarch inevitably would crush enemies. There was no need and no chance to develop a conception of the past in which the course of events would be ascribed to the contending efforts of multiple actors, the result of which was unpredictable and thus needed to be reconstructed and narrated in each case. The historical books of the Old Testament, in a certain sense, took account of the contingency of human affairs but at the same time attributed the outcome of events to God's intervention.

1. The Origins of Greek Historiography

After the development of literacy during the eighth century BC, Greek commemoration of a distant past had materialized in the epic tradition (Homer, Hesiod). Its combination of 'historical' and 'mythical' elements became subject to criticism in an age when Ionian sixth century philosophy (as represented by the Milesians Thales, and Anaximander) had attempted at a 'scientific' explanation of the physical world and of human society. About 500 BC Hecataeus of Miletus tried to reconstruct the genealogical facts out of the mythological traditions. He wrote a pioneering work of geography which reflected the expansion of the geographical and historical horizon of the Greeks as a result of their colonization (since c. 750 BC). This

included their expansion on the shores of the Mediterranean and Black Seas, their voyages of discoveries, and the information that the Eastern Greeks (as subjects of the vast Persian Empire) had acquired about parts of Asia. Historiography, in a narrower sense, developed out of the experience of the Persian Wars. However, the realization that the great victories of the Greeks in 490 and 480–479 BC had been achieved, so to speak, against all probability apparently developed only with a certain distance to the events themselves.

Herodotus (c. 485—after 425 BC; a widely-traveled native of Halicarnassus in Asia Minor) opened his account of the Persian Wars with the statement that he wanted to present the results of his research (*historie*) in order to commemorate the deeds of both Greek and Barbarians and to establish the cause of their conflict. *Historie* is a general term for a rational inquiry. It was only later (especially with Aristotle's *Poetics*) that it became the label for the literary genre which Herodotus had created. Herodotus pointed out that his study was the work of an individual and that he felt obliged to an impartial approach. Greek historiography was developed by free intellectuals and did not start as officially committed accounts. Apparently, especially exiles like Herodotus—and later *inter alia* Thucydides, Xenophon or Polybius—were able to take on the position of an (at least, subjective) neutral observer. Herodotus did not only provide the first work of an *histoire événementielle* based on stories which might have been generated two generations before his own time; he also continued the tradition of geographical and ethnographical accounts by incorporating digressions on the various peoples (from the Egyptians to the Scythians) under Persian control. Here he relied on local traditions and autopsy and sometimes reported stories, the authenticity of which he himself doubted.

Herodotus' younger contemporary Thucydides of Athens (c. 460—c. 400 BC) created a quite different model of historiography. He deliberately distanced himself from Herodotus (though without naming him) since he thought Herodotus to have shown too much confidence in his sources. Thucydides' statements on his methodological principles are (with the notable exception of Polybius) almost without parallel in ancient historiography. He chose as his subject the Peloponnesian War between Athens and Sparta and their respective allies (431–404 BC; Thucydides' account, however, is incomplete and breaks off at 411 BC). According to his own words he had started to take notices already with the outbreak of the war since he had anticipated the unprecedented dimensions of this war which affected the whole Greek world. Later, he used his time in exile (the consequence of his defeat as an Athenian general in 424) to collect all available facts by interrogating protagonists and eyewitnesses of both sides. Being confident that he was able to reconstruct the objective course of events he did not

name his sources or quote varying, or even conflicting, versions by different witnesses but digested all information into an authoritative account of 'how it had really happened' (an aside of his which was later to be taken up by Ranke). He wanted to differentiate the long-term causes, which led to a climate of mutual distrust between the two powers, from the immediate occasions, which made them take military action. War and civil war were seen by Thucydides as revealing the *condition humaine*, love of power, ambition, greed, and fear as constituents of people's invariable nature; thus the thoroughgoing analysis of human action should provide insights of everlasting value.

Already Herodotus had incorporated some speeches into his narrative in order to explain the motives behind the decisions of protagonists but Thucydides went much further by systematically using this stylistic device based on the techniques of rhetoric that were taught by contemporary Sophists. As Thucydides explained, his methodical principle was 'to make the speakers say, what, in my opinion, was called for by each situation while keeping as closely as possible to the general sense of the words that were actually used.' Those speeches served to present opposing arguments (in pairs of speeches ascribed to antagonists) and to discuss the political, moral, and juridical implications of certain decisions or to express the self-image of a society (especially in the 'Funeral Speech' attributed to Pericles). The unity of style, the high level of abstraction, the implicit cross-references, and the shortness of the speeches point out that they were primarily a medium to carry the historian's interpretation. Since Thucydides, the inclusion of more or less fictitious speeches was a standard feature of ancient historiography. Thucydides made two exceptions to his principle of writing a strictly contemporary history. He gave an account of the development of early Greek history in order to prove that due to the level of its material culture, a war of such dimensions like the Peloponnesian War would have been impossible. That was a sort of conjectural history based on theories of societal development and on the interpretation of cultural survivals. He also gave a selective description of the 50-year period between the end of the Persian Wars and the outbreak of the Peloponnesian War in which he stressed the rise of Athenian power and the distrust it fostered with the Spartans. This account also served to bridge the gap between Herodotus' narrative and his own.

Thucydides' model of historiography differed from the Herodotean one by eliminating the features of cultural history and ethnographical discourse. But both corresponded with respect to concentrating on large-scale wars, which involved great parts of the Greek world, and with regard to treating matters of constitutional history and internal politics of the states concerned only insofar as they were relevant for the course of events. Their different attitudes to the period

of time that could be reconstructed by drawing on oral traditions notwithstanding, they both primarily relied on oral sources. As a rule, down to the Christian historians of Late Antiquity ancient historiographers did not consult documentary evidence systematically like inscriptions or state papers in archives and did not engage seriously in 'antiquarian' research on constitutional, legal, and religious institutions. That is why they could not think of replacing the narratives of former historians by a new account based on other better sources. Thus Thucydides was a continuator of Herodotus, Xenophon, and other fourth century historians who started their Greek histories at the point where Thucydides' narrative had broken off. Historians who wanted to treat former times again, cannibalized (often in combination with harsh criticism) the accounts of their predecessors.

2. The Variety of Greek Historiographical Production

Thucydides had shaped the trend-setting model of military and political history that was followed by a great number of Greek writers including the historians of Sicily. However, his successors, embellished the rhetorical elements or amplified scenes with emotive appeals to readers. Throughout antiquity, Herodotus enjoyed a dubious reputation as being at the same time the father of history but also a narrator of fabulous stories (in Cicero's famous phrase). The broader Herodotean approach, however, was not totally outdone; it was taken up again by the historians of Alexander the Great and survived in later works with strong geographical and ethnographical interests like that of Posidonius (c. 135–c. 50 BC) and in the specialized geographical works from Eratosthenes (third century BC) to Strabo (Augustan age).

Besides the mainstream of military and political history, a number of specialized historical genres developed since the late fifth and during the fourth centuries. These included: universal history (Ephoros of Cyme, fourth century BC, whose—later lost—work was used by Diodorus Siculus, first century BC); foundation histories, local chronicles, and constitutional histories (e.g., on Athens); cultural histories (Dicaearchus, pupil of Aristotle, late fourth century BC); biographies (late fifth century BC sketches later to be used by Plutarch, c. 45–c. 120 AD; and Xenophon's life of the Spartan king Agesilaus II, fourth century BC).

During the Hellenistic Age, an enormous number of historical works of various kinds have been produced by nearly a thousand historians known to us by name. This tremendous increase of historical production reflected the political reorganization of the Mediterranean world by the establishment of the vast Hellenistic empires (and later the rise of Rome to a world power). It was fostered by the patronage of Hellenistic

monarchs and the establishment of libraries and scholarly institutions, especially in Alexandria and Pergamum. Almost all of this extremely rich production has, however, got lost and is only known from quotations and references in later works. For example, the various reports on Alexander's campaigns by court historiographers, officials, and generals only emerge with the compilation of Diodorus; comprehensive accounts of Alexander's history based on the works of his companions date from the second and third centuries AD (Arrian and Curtius Rufus).

Polybius (c. 200–c. 120 BC) is the only Hellenistic historian of whose historical work a substantial amount has survived. Polybius, a military of the Achaean Confederacy (on the Peloponnese) was one of 1,000 leading Achaeans who were deported to Italy after 167 BC. He came into close contact with prominent figures of the Roman nobility, especially Scipio Aemilianus (the destroyer of Carthage). The subject of his historical work was Rome's rise to world dominion in the time from the Punic Wars to the destruction of the Macedonian state (167 BC). Polybius calls it a 'pragmatic history'; it focuses on political and military actions and is especially based on eye-witness reports and first-hand knowledge of the theaters of war. His intentions and explicit methodological reflections (combined with polemic against several former historians) imply a conscious return to the Thucydidean model of historiography. But Polybius understood Rome's military success as having depended on its uniquely stable, and at the same time flexible, political system. In order to explain this point to his Greek readers he incorporated a description of the Roman constitution and military system and thus gave a systematical treatment of institutions otherwise not to be found in ancient historiography. (His analysis of Rome's 'mixed constitution' should become one of the most important subjects in the reception of ancient political theory since the Renaissance.)

3. Roman Republican Historiography

Roman history was first made the subject of historiographical works by Greek authors like the Sicilian Timaeus of Tauromenium (c. 350–c. 260 BC) and other Greek writers who were especially interested in the stories about the foundation of Rome. Latin literature only developed since the mid-third century BC as a product of the reception of Greek culture. Certain records of the past, however, had been preserved since early times: public documents like laws, decrees of the senate and international treaties, and family records of the aristocracy. And there were the *annales* kept by the pontiffs in which the names of consuls (the eponymous magistrates) were listed, triumphs and religious prodigies were noted, and some

miscellaneous information was collected. About 200 BC, the senator Fabius Pictor provided the first history of Rome written by a Roman yet in Greek. Drawing on the models of Hellenistic historiography Fabius Pictor covered the time from the founding of the city until 200 BC. He probably gave elaborate accounts of the early, and of the contemporary periods and only a sketchy one of the time in between, a proportion that should become typical of many later writers. He was followed by other historiographers writing in Greek until the mid-second century. The older Cato (234–149 BC) was presumably the first to write a historical work in Latin. His *Origines* applied the interest in foundation history not only to the earliest history of Rome but also to the emergence of the cities in Italy and paid particular attention to the unification of Italy. His treatment of political history focused on the justification of his own policy and thus provided the model for the *ex parte* tendency of Roman contemporary historiography. Roman historiography was ethnocentric and it was the product of the leisure-time activity of senators. Their intentions were to demonstrate Rome's cultural equality with the Greeks, to justify her domination over the Mediterranean world, to praise the wisdom and success of the ruling aristocracy, and to take sides with respect to contemporary conflicts. From the late second century onwards, a number of Latin historiographical works were composed which followed in their structure the pontifical *annales* (which were published in a full version in 120 BC) and gave a year-by-year account from the origins of the city. They apparently embellished their descriptions of the early republic and especially the 'Struggle of the Orders' between patricians and plebeians by projecting back features of the internal conflicts that had become almost endemic since the tribunate of Tiberius Gracchus (133 BC). Some historians seem to have taken into account certain results of serious antiquarian research, others had no scruples and they invented 'documents.' (It was the Greek author Dionysius of Halicarnassus who, living in Augustan Rome, showed the greatest interest in evidence of antiquarian nature.) All these Latin 'annalistic' works (as well as the earlier historiographical attempts previously mentioned) are only indirectly known. They were used by Livy (c. 59 BC–AD 17) who was the first Roman historian without a record of political and military experience of his own. His history covered the time from the foundation of the city to 9 BC. Out of his 142 books, only 35, which reach down to 167 BC (with *lacunae*) have survived. Livy often struggled with the contradictory accounts of his predecessors and the quality of his own descriptions obviously depends on the quality of the sources available to him. He did not invent his own versions of the course of events but his work shows the problematic nature of a history totally made on the basis of other, often overtly partisan, historical works that were mostly accepted at face value.

The Thucydidean model of contemporary history was imitated by Sallust (86–35 BC) who after the failure of his political career wrote monographs on the conspiracy of Catiline (63 BC) and on the Roman Wars against the Numidian king Jugurtha (111–106 BC). He interpreted the political crisis of the Late Republic as the consequence of a moral decline following from the abandonment of traditional values. (No other contemporary work on the Late Republic has survived; a comprehensive account of this period is only available in works by later authors such as Appian, second century, and Cassius Dio, third century, who both wrote in Greek.)

4. Roman Imperial Historiography

Writing contemporary history might become a dangerous job under the Emperors as the fate of the historian Cremutius Cordus demonstrates. Under the reign of Tiberius (in AD 25) he was indicted for *lese majesty* because of the republican sympathies he had shown in his work on the Civil War period (he committed suicide; his work was burnt). The short sketch by his contemporary Velleius Paterculus, a former military, covered Roman history from the very beginnings to his own days in only two books and culminated in the panegyric of the emperor Tiberius.

The historiography of Tacitus (AD c. 55–c. 120) reflects a new climate of intellectual freedom during the reigns of Nerva and Trajan in comparison to the experience with the tyrannical régime of Domitian (under whom Tacitus had started his public career). His (only partly preserved) works, the *Histories* and the *Annals*, cover the time from AD 69–96 and AD 14–68, respectively. Tacitus kept the annalistic form and he paid special attention to the proceedings of the senate. But the form was often at odds with the contents of the works since the policies of the rulers could not easily be fitted to a chronological scheme that had been adequate for a senatorial régime with yearly changing consuls at the head. Tacitus showed certain nostalgia for republican liberty without indulging into illusions about a return to a republican constitution. Though claiming to write *sine ira et studio* his personal judgments on rulers (especially the gloomy picture of Tiberius) should serve to uphold the mirror of an honest ruler and his comments on senatorial proceedings reflected his standards for dignified behavior under a monarchy. Tacitus' work with its stress on the psychology of emperors and senators marks the final point of Latin senatorial historiography. (However, later there was an annalistic history of Rome from the beginnings to AD 229 written in Greek by the senator Cassius Dio.)

Apart from short summaries of Roman history, during the second century serious historiography gave way to imperial biographies. Suetonius (AD c. 70–c. 140, secretary for the library and subsequently for the imperial correspondence under Hadrian) gave the life

of 12 rulers from Caesar to Domitian. They were arranged in a more topical than strictly chronological manner and drew on archival materials as well as traditions stemming from court gossip. Modeled on Suetonius but of a highly fictitious character (with numerous references to imaginary sources) was the *Historia Augusta*, a series of biographies of emperors for the time from Hadrian (AD 117–138) to the accession of Diocletian (AD 285) which was probably produced in the late fourth century by one or more anonymous writers who took on the names of bogus authors.

5. *The Impact of Christianity*

The religious tensions of Late Antiquity due to the rise of Christianity did only partly affect pagan historiography which continued to concentrate on short summaries of Roman history. The general restraint from religious polemic holds also true for the work of Ammianus Marcellinus (c. 330–395), a Greek from Syrian Antioch and former military. His history, written in Latin, is otherwise a great exception since he embarked on a continuation of Tacitus that treated the time from Nerva's accession to the throne (96) to the emperor Valens' disastrous battle of Adrianople against the Goths in 378 (only the exhaustive parts on the years from 353 onwards have survived).

Christian authors did not enter into a competition with pagan ones on the traditional fields of political and military historiography but they developed models of historical writing that should become the most important innovations until Renaissance times. Since the third century they compiled chronicles of universal history which synchronized the data of biblical, Greek, and Roman history. The chronicle—in Greek—by Eusebius (c. 260–339, bishop of Caesarea, adviser, and later biographer of Constantine the Great), its Latin translation and continuation (until 378) by Hieronymus and several successor works provided the model for a view of world history since the creation. Eusebius, again, created with his *Church History* a new historiographical genre not only with concern to the subject, the history of the Christian 'nation,' but in formal respects as well by his dispensing with speeches, his giving bibliographical surveys, and his verbatim quoting from letters, official documents, and other sources. The awareness that their theological approach might make their accounts challengeable led Eusebius and later *inter alia* Socrates and Sozomenus (both early fifth century) to prove thus the accuracy of their factual reports.

6. *Concluding Remarks*

Given the complexity and variety of ancient historiography it is almost impossible to summarize its impact on the later Western tradition, especially since the Renaissance, in a few sentences. The most im-

portant legacy of Greek historians was the development of a critical attitude to the past, the invention of methods that made it possible to discern between facts and fiction. The shortcomings of their attempts in view of modern eyes should not obscure this unique contribution to the development of the Western mind. The adoption of historiography by the Romans provided the model for constructing the past in the shape of national history. The Christian writers invented ecclesiastical historiography as a subject of its own and a model for universal chronicles and histories starting from the creation. Since the Renaissance, Thucydides and Tacitus were taken as the masters of political history, Polybius as the model military historian. With the growing interest in Oriental cultures Herodotus gained a new reputation as a cultural historian. Ancient historiography, which had covered republican city-states as well as world monarchies, was as a whole considered as teaching political and moral lessons of everlasting value (and thus was also of great importance for political theory). The tradition that historiography and antiquarian studies were two quite different subjects survived until the late eighteenth century. There were a vast number of commentaries on ancient texts, collections of inscriptions and coins, learned dissertations on the constitutional, legal, military, and religious institutions of the ancient world but nobody thought of replacing the ancient historiographical works by writing anew narrative histories of antiquity. Gibbon was one of the first to do this and to integrate the results of antiquarian research but he still avoided the period covered comprehensively by ancient historiography and thus started his own narrative only with the second century AD. The doubts on the reliability of ancient historiography, especially of the Roman annalistic tradition, that had been developed in learned dissertations since the seventeenth century led only in the early nineteenth century, starting with Niebuhr, to compose a modern historiography of antiquity based on source criticism and systematic researches on the monumental remains of the ancient world.

See also: Antiquity, History of; Aristotle (384–322 BC); Christianity Origins: Primitive and 'Western' History; Classical Archaeology; Historiography and Historical Thought: Christian Tradition; History: Forms of Presentation, Discourses, and Functions; History: Overview; History: Theories and Methods

Bibliography

- Dorey T A (ed.) 1966 *Latin Historians*. Routledge and Kegan Paul, London
Fornara C W 1983 *The Nature of History in Greece and Rome*. University of California Press, Berkeley, CA
Hornblower S (ed.) 1994 *Greek Historiography*. Clarendon Press, Oxford, UK
Meier C 1975 *Geschichte (II. Antike). Geschichtliche Grundbegriffe* 2: 595–610

- Meister K 1990 *Die griechische Geschichtsschreibung*. Kohlhammer, Stuttgart, Germany
- Mellor R 1988 Roman historiography and biography. In: Grant M, Kitzinger R (eds.) *Civilization of the Ancient Mediterranean*. Charles Scribner's, New York, Vol. III, pp. 1541–62
- Momigliano A D 1955–1992 *Contributi alla storia degli studi classici e del mondo antico* (9 Vols.). Edizioni di storia e letteratura, Rome
- Momigliano A D 1978 Greek historiography. *History & Theory* 27: 1–28
- Momigliano A D 1990 *The Classical Foundations of Modern Historiography*. University of California Press, Berkeley, CA
- Strasburger H 1977 Umblick im Trümmerfeld der griechischen Geschichtsschreibung. In: *Historiographia Antiqua. Commemationes Lovanienses in honorem W. Peremans septuagenarii editae*. Leuven University Press, Leuven, Belgium, pp. 3–52
- Usher S 1988 Greek historiography and biography In: Grant M, Kitzinger R (eds.) *Civilization of the Ancient Mediterranean*. Charles Scribner's, New York, Vol. III, pp. 1525–40
- Winkelman F 1991 Historiographie. *Reallexikon für Antike und Christentum* 15: 724–66
- Woodman A J 1988 *Rhetoric in Classical Historiography*. Croom Helm, London

W. Nippel

Historiography and Historical Thought: Current Trends

1. Introduction: Dominance of Social Science Models of History around 1970

The years 1970–2000 have seen important reorientations in historical thought and historical writing. These changes are closely related to fundamental transformations in the conditions of modern life.

Briefly comparing the intellectual and historiographical situation in the year 1970 with that of 2000, we can distinguish certain dominant trends then and now. In neither case, however, is it possible to reduce historical thought to simple common denominators. There are no paradigms in the Kuhnian sense that determine historiographical practice but a great plurality of outlooks that are perhaps even more numerous now than they were then. Nevertheless, in 1970 most historians were still convinced that history was a science, albeit a social science, engaged in recreating a real past as it had actually occurred. Today this confidence in the objectivity of science, including historical inquiry, has come under severe attack, as have the notions still dominant in the early 1970s of a coherent process of history. To be certain, faith in objective truth, in historical inquiry, and confidence in history as the ongoing development of civilization had been undermined as early as the second half of the nineteenth century, e.g., by Jacob

Burckhardt and Friedrich Nietzsche, and even in the United States, which remained a citadel of belief in science and progress, by leading historians such as Carl Becker and Charles Beard in the 1930s.

Historical science, as it was generally understood in the Western world around 1970, viewed history as a social science. It replaced an earlier orientation sometimes labeled 'historism' (which had been dominant in much of the nineteenth and well into the twentieth century), that had argued that history as a *Geistes-* or *Kulturwissenschaft* required a logic of inquiry different from that of the hard sciences, namely one which takes into account the elements of meaning and intention in human life which cannot be reduced to abstract generalizations. These conceptions of historical study that viewed history as a special sort of *Geisteswissenschaft* were increasingly replaced in the twentieth century by conceptions of history as a 'social science.' The new social science orientations went in several directions discussed below, but they shared certain assumptions. They criticized the older historiography for narrating a history of events, primarily political, focusing on leading personalities in positions of power, and called for a history that proceeded analytically, seeking to examine the social context in which events occurred. While the older historiography relied heavily on philology as a tool for the examination and understanding of texts to fathom human intentions, the new history turned to the neighboring social sciences, sociology, economics, and psychology, and in France also geography, which analyzed social structures and processes and stressed the interdisciplinary nature of historical studies. They strove for a history that was more comprehensive than that of the conventional narrative political historiography. They also wanted to include broader segments of the population into the historical accounts.

But it was only after 1945 that forms of social science history emerged as a dominant model for historical studies. Again it is difficult to reduce social science-oriented history to a common denominator; it ranged from the studies of the material and biological bases of historical life in the French *Annales* in the 1960s, to the new historical demography in France and Great Britain, to quantitative urban history and electoral behavior in the United States, to the formulation of quantifiable models of economic cycles in the United States, Great Britain, and France, to a 'historical social science' in West Germany strongly influenced by Max Weber. Nevertheless, despite their differences, they all agreed that history had more kinship to the social sciences than to the humanities. This assumed, as the earlier social science-oriented history had also done, that there was a real historical past that constituted the object of historical studies. It sought to make historical method conform to the methodologies of other social sciences. Hence, it generally preferred different kinds of sources than did the older historiography; for example, hard economic

and demographic data. This made Geoffrey Barraclough comment in a survey written in the 1970s for UNESCO on trends in recent historical studies that:

the search for quantity is beyond all doubt the most powerful of the new trends in history, the factor above all others which distinguishes historical attitudes in the 1970s from historical attitudes in the 1930s (Barraclough 1978, p. 39).

These historians tended to sacrifice individual human beings to structures and processes thus largely ignoring the individuals who compose history.

A key to almost all forms of social science history, with the possible exception of many of the historians in the *Annales* tradition, was the notion of history as an ongoing directional process. This was true of the various advocates of history as an empirical, analytical social science as it was for the different shades of Marxists and for the German school of 'Historical Social Science.' A broadly shared concept was that of modernization. It went back to nineteenth-century positivism and assumed that history, or at least the history of the modern Western world, was marked by an ongoing process towards greater scientific enlightenment followed by technological progress and by the concomitant development of a 'modern' society. While most forms of social science history tended to downgrade political history, the German school of 'historical social science' (Wehler, Kocka) sought to locate politics in the context of social structures and processes to understand how it was possible for National Socialism to gain power in Germany.

2. Challenges to Social Science Approaches to History: Postmodernism

This widely held confidence in social science and modernization was radically challenged in the 1970s and 1980s, most vocally by postmodernist philosophers and literary critics, but more quietly also by historians, particularly cultural historians. While it is difficult to reduce these critics to a common denominator, they agree in repudiating two assumptions that they consider to be central to the modern historical outlook. The first concerns the belief in the coherence and continuity of world history, the invention of a 'master' or 'metanarrative' (Lyotard) which sees history as an onward development to the modern Western world with its promise of enlightenment, emancipation, and well-being for the world as a whole. Practitioners of the established historiography, critics charged, had overlooked the close relationship between Western progress and imperialism abroad and the consolidation of power relations based on race, gender, and class at home taken for granted in their view of history. But history, they argued, had no coherence, thus eliminating the possibility of a grand narrative and opening the way to innumerable separate histories.

Closely related to this critical view was the rejection of the possibility of objective historical inquiry into the past. According to Peter Novick the idea and ideal of objectivity lay at the very center of the historical profession as it was constituted with its 'commitment to the reality of the past, and to truth as correspondence to that reality' and consequently 'the separation of history and fiction' (1988, pp. 3–4). This commitment, Hayden White noted, rests on the 'illusion that there is a past out there that is directly reflected in the texts' (1987, p. 209). Postmodernists according to Keith Jenkins seek to free modern historical consciousness from the notion that behind appearances there lies a solid, foundational, real world. 'In fact history now appears to be just one more foundationless, positioned expression in a world of foundationless, positioned expressions' (Jenkins 1997, p. 6).

For the postmodernists history consists of texts that are linguistic constructs with no direct relation to reality. It is language that creates reality. Jacques Derrida went further in denying that there is anything beyond the text (*il n'y a pas de hors-texte*) (see Megill 1985, p. 266). Moreover, for Derrida the text has no stable signification, but is open to an infinity of significations because there is no Archimedean point from which a clear meaning can be assigned. The text, moreover, according to Foucault is seen not only independently of its relation to the external world, but also independently of its author. Derrida criticized what he considered to be the 'logocentrism' of the classical Western philosophic tradition since Socrates that rested on the error that reality could be reduced to rational concepts, ignoring that these concepts reflected systems of domination. The idea that reality was not given but was constructed through language was further developed in feminist theory. Gender, for Joan Scott, is not found as such in nature but is socially constructed through language, texts, and cultural symbols that reflect or challenge power relationships. Thus the deconstruction of language is a necessary tool for revealing and dismantling these unequal power relationships.

All this leads to epistemological relativism. Perhaps the most influential proponent of such relativism as it relates to historiography has been White, who argued that historical thought is 'the captive of the linguistic mode in which it seeks to grasp the outline of objects inhabiting its field of perception' (1973, p. xi). Consequently it is not the purportedly objective investigations of the historian into a real subject matter which lead to knowledge about history but rather the knowledge at which the historian arrives is conditioned by the linguistic mode in which she/he operates. Professional historiography for White generates no more objective knowledge of the past than does speculative philosophy of history or the historical novel. Any attempt to go beyond the facts to a coherent narrative is a 'poetic act' so that the 'best

grounds for choosing one perspective on history rather than another are aesthetic or moral rather than epistemological' (1973, p. xii). 'Viewed simply as verbal artifacts,' he notes, 'histories and novels are indistinguishable from one another' (1978, p. 122). Similarly, 'the opposition between myth and history ... is as problematical as it is untenable' (1978, p. 83).

This negation of a distinction between history and fiction is at the core of the French literary critic Roland Barthes's essay—*Le Discours de l'histoire* (*Discourse of History*). For Ankersmit, while 'science was the alpha and omega of the modernists' (1995, p. 166), for postmodernism 'the metaphorical dimension of historiography is more powerful than the literal or factual dimensions ... criticizing metaphors is indeed an activity which is just as pointless as it is tasteless. Only metaphors "refute" metaphors' (p. 180). Hence research is irrelevant because 'the "real past" does not enter into historiography except rhetorically' (Jenkins 1995, p. 18).

3. *The Turn to Cultural History and the Return to Narrative*

Few practicing historians would agree with Barthes, Derrida, White, Ankersmit, or Jenkins that there are no criteria of objectivity. They would continue to hold that historical works in contrast to historical novels must proceed from research that requires standards of objectivity shared by the scholarly community. Postmodernism had reflected a general mood that was shared by many historians. Part of this mood was a general skepticism regarding the nature and quality of modern civilization and its science; the recognition that the historical science of the West, whether in its historicist or its social science form, had legitimized relationships of power in which a middle and upper-class elite had been privileged at the expense of the broad masses of the population, including women, and of the nonWestern world. Thus, by the 1970s the old scientific ethos of the social science historians was shaken. In 1979 in a now famous article, *The Revival of Narrative*, Lawrence Stone noted that in the 1970s a change had occurred in the way history was viewed and written. The belief, central to social science history, that 'a coherent scientific explanation of change in the past' (1979, p. 19) was possible, was widely rejected. In its place there emerged an interest in the most varied aspects of human existence, accompanied by the conviction that 'the culture of the group, and even the will of the individual, are at least as important causal agents of change as the impersonal forces of material output and demographic growth' (p. 9). Yet, although Stone emphatically rejected the illusion of 'coherent scientific explanation,' he nowhere suggested that historical narrative should surrender its claim to rational inquiry and realistic reconstruction.

4. *The Impact of Cultural Marxism on the Writing of History*

The new directions which historical studies took must be understood within the context of the political climate of the 1960s and after, such as the emergence of a women's movement, the rise of ethnic consciousness, the struggle for racial equality, anti-colonialism, and environmentalism which all pointed at the inadequacies of the social and economic order of an industrialized, modernizing Western world and the social science which served it. The transformations within Marxist theory and historiography provide important indicators for the change of attitude among historians. Marxism itself was in a deep crisis as indicated in the Soviet power block by the suppression of the Prague Spring and by the rejection of Marxist orthodoxy in the West.

But the significance of a non-dogmatic Marxism on historical thought in Western Europe as well as in Japan, India, and Latin America should not be underestimated. Marxism–Leninism was of little significance there as compared to Western forms of Marxism informed by Antonio Gramsci, Georg Lukács, and the Critical Theory of the Frankfurt School. Western Marxism preserved the idea of class conflict, the role of domination as a major force in social relations, but gave up the idea of a necessary development to a classless society.

The reorientation of Marxism began in the 1920s and involved an examination of the traditional economic interpretation of history. Gramsci's writings in Mussolini's prisons in the 1920s and 30s, which became known in the West only in the 1960s, explained the defeat of the communist workers movement in Italy in terms which involved factors insufficiently developed in traditional Marxist philosophies of history. For him the strength of capitalism rested ultimately not exclusively on its economic strength but on its cultural hegemony over the working classes. Lukács already in 1923 had seen in the section on 'The Fetishism of Commodities' in *Capital* the core of Karl Marx' critique of economic conceptions of capitalism. The empiricism that dominated scientific and social science thought in his opinion failed to understand the broader context of society and culture. A similar note was pursued in the 1930s and 1940s by the 'critical theorists' of the Frankfurt School (Horkheimer, Adorno, Marcuse) who, proceeding from a critique of capitalism that they argued had turned culture into an industry in the service of economic ends, strove for a science which went beyond quantifiable data to the qualitative aspects of society and culture. They nevertheless gave up Lukács' historical optimism.

Perhaps no book had as great an impact on the reorientation of social history, at least in the English-speaking world but also beyond, as E. P. Thompson's *The Making of the English Working Class* (1963).

Thompson maintained the concept of class and the notion that 'the class experience is largely determined by the productive relations into which men are born—or enter involuntarily' (p. 9). But he stressed that class must be seen not as 'structure,' or even 'category' but as something which in fact happens in human relationships (p. 9). The English working class was thus not primarily the product of economic forces, as it had been in the *Communist Manifesto* or in structuralist, functionalist sociology, but an active agent, coming from a specific cultural tradition, which made itself as much as it was made. Hence, his reliance on qualitative elements that make up a culture such as literature, art, folklore, and symbolism. Moreover, Thompson emphatically rejected the orthodox Marxist notion that sees the past as a step to the future.

Yet Thompson's historiographical approach soon came under attack from cultural Marxists who argued that he had not broken sufficiently with classical Marxist notions of class. Even if he saw class in terms of culture, they charged, this culture still focused on an industrial working class in which those not directly linked to the industrial work process played little of a role. The *History Workshop: A Journal of Socialist Historians*, founded in 1976 in Great Britain—which in 1982 changed its subtitle to *A Journal of Socialist and Feminist Historians* and in 1995 dropped the subtitle altogether—tried to overcome this by expanding the concept of work in a changing society and including women in the family and the workplace.

From the perspectives of much post colonialist history of the 1980s and 1990s, such as put forward by Latin American *Dependencia* Theory or Indian Subaltern Studies, Marxism operated with a concept of historical development, akin to that of capitalist writing, for which modernization was understood as Westernization. Edward Said in *Orientalism* (1978) argued that Western scholarship had invented an Arab culture that rested not on historical knowledge but on preconceptions, pitting an enlightened West against a backward Muslim East which needed to be Westernized. Postcolonialist writings integrating feminist, psychoanalytical, and Marxist theories argued that categories of class, ethnicity, and gender must be seen in their interaction. They held that a Marxist conception of class conflict misses the analysis of power to which Foucault had pointed as an infinitely complex network of micro powers that permeate every aspect of social life and culture.

5. From Macro- to Microhistory

While few social historians could be identified as postmodernists or Marxists, certain central ideas of postmodernism and of cultural Marxism impacted a great deal of writing in social history which increasingly moved in the direction of cultural history. The idea of the unity or direction of history gave way to the

concentration on small units, on the local and the marginal.

While this same orientation destroyed the credibility of orthodox Marxism, the Marxist commitment to the oppressed, even if oppression was now understood differently, occupied a great deal of the new culturally oriented social history whose subjects were most often 'the lives and feelings and behaviour of the poor and obscure rather than the great and powerful' (Stone 1979, p. 19). Anthropology and semiotics replaced sociology and economics as the key auxiliary sciences of social history. But anthropology underwent a sea change. While classical anthropology, e.g., that of Bronislaw Malinowski in the 1920s and 1930s, had tended to a static analysis of 'primitive' and 'timeless' societies, anthropologists had now become increasingly historical in outlook, recognizing that all, even the most 'primitive' cultures, evolve and that modern societies as well are proper subjects of anthropological study.

At the same time, historians increasingly turned to an interpretative, cultural anthropology. As the cultural anthropologist, Clifford Geertz, phrased it:

Believing with Max Weber that man is an animal suspended in webs of significance he himself has spun, I take culture to be those webs, and the analysis of it to be therefore not an experimental science in search of laws but an interpretive one in search of meaning (Geertz 1973, p. 5).

In Germany, this anthropological turn in the 1980s took the form of *Alltagsgeschichte*, which sought to reconstruct the experiences and emotions of common people (Hans Medick, Alf Lüdtke, Lutz Niethammer) in earlier preindustrial as well as in modern industrial times. In Italy, historians (Carlo Ginzburg, Eduardo Grendi, Carlo Poni, Giovanni Levi) around the journal *Quaderni Storici*, began in a series of books and articles to give a firm theoretical foundation to a microhistory (*microstoria*) which reduces historical research to identifiable individuals and from these creates a history from which the net of relationships of these individuals would emerge. In place of the thousands of quantifiable data on which the older social history constructed a history of mentalities, the microhistorians wanted to recapture a fuller picture of individual experiences using transcripts and inquisitorial records in which the voices of the subjects were recorded. Outside of Italy, Emmanuel Le Roy Ladurie in *Montaillou* (1975) used such records to recapture the lives and emotions of individual villagers in a fourteenth-century village while Natalie Davis in *The Return of Martin Guerre* (1983), using judicial records, focused on gender in sixteenth-century rural France.

Nevertheless, the new orientations toward microhistory did not necessarily mean a break with the social sciences. Foucault had rejected the idea that theories can be verified because in his view the standards of verification came from a social scientific discipline which represented institutions of power;

Geertz had argued that any attempt to approach a historical subject matter with theoretically oriented questions distorts this subject and has to be replaced by 'thick description' which immerses itself untheoretically into an alien world. This, in effect, invalidated traditional forms of historical inquiry. But neither Ginzburg nor Levi were willing to go so far. For them, *microstoria* does not reject the empirical social sciences *in toto*, but stresses the methodological need for testing their constructs against existing reality on a small scale. Levi's concern in *L'eredità immateriale: Carriere di un esorsista nel Piemonte del Seicento* (1983) (*Inheriting Power: The Story of an Exorcist*) is social scientific insofar as he wishes to test established hypotheses on a local level. His central concern is patterns of power relationships in the village. In studying land purchases, Levi questions the extent to which impersonal market forces and the development of the modern state machinery determine these power relationships and argues that the decisive element in the understanding of the peasant world is 'the preservation or transmission of intangible or symbolic goods: power and prestige' (back cover of English paperback edition). Here Levi was much closer to Pierre Bourdieu's conception of 'symbolic capital' than to Geertz' 'thick description.'

The gap between social and cultural history began to close markedly in the 1980s and 1990s. On the one hand, there were important works in microhistory emphasizing anthropological aspects of culture (Medick, David Sabean). Despite the reservations these historians expressed in regard to the systematic social sciences, they relied heavily in their work on economic and demographic data and made use of computer analyses. On the other hand, historians coming from a social science orientation recognized the significant role of the cultural dimension in society. The work that the German school of 'Historical Social Science' had undertaken in the social analysis of politics now increasingly gave consideration to the interaction of society and culture. An example of this is the major comparative project on the nineteenth-century European bourgeoisie (*Bürgertum*), headed by Jürgen Kocka, which defined the latter as much in terms of culture and education as in terms of economic status. The journal *Geschichte und Gesellschaft* became an important forum for the presentation and discussion of these tendencies.

The economic analysis of class also began to play a decreasing role in the study of the revolutionary movements since the eighteenth century. By the 1970s not only the economic interpretations of the French Revolution by Marxists such as those of Georges Lefebvre and Albert Soboul, but also of their non-Marxist critics such as Alfred Cobban and George Taylor, were replaced by new viewpoints (Régine Robin, François Furet, Lynn Hunt) which placed greater stress on culture and language. As Lynn Hunt explained in the introduction to her *Politics, Culture,*

and Class in the French Revolution (1984), this work conceived in 1976 began as 'a social history of Revolutionary politics' but 'increasingly turned into a cultural analysis in which the political structures ... became but one part of the story' (p. xi).

Linguistic analysis plays an important role in William Sewell's analysis of the 1848 Revolution in France, in Gareth Stedman-Jones' study of British Chartism, and Thomas Childress' examination of electoral propaganda in the Weimar Republic. None of them, however, would share Foucault's radical position that 'reality does not exist, that only language exists' (quoted in Berman 1988, p. 183). Rather, they would agree with Carroll Smith-Rosenberg that 'while linguistic differences structure society, social differences structure language' (Spiegel 1990, p. 84).

6. The Transformation of Political History

With the move to the social sciences in the 1950s and 1960s and to the cultural sciences since the 1970s, the significance of political history has been downgraded. Various forms of social science history, as early as the *Annales* before 1945, reacted against the focus on political history in conventional historiography. Moreover, the new cultural history that sought to move away from conceptions of centralized power also consciously neglected the state. Yet this led to a certain blindness regarding the realities of the present world. The rapid changes in international relations since the collapse of communism in 1989–1991 have made the significance of politics for society and culture apparent. Recent historical theory with its emphasis on the pluralism of historical narratives has assigned the process of modernization and globalization to the sphere of macronarratives that it rejects. But one cannot escape the reality of processes of modernization, however complex and varied they may be, as significant aspects of modern historical development.

The history of politics, of international relations, of peace and war, and the study of political regimes in the twentieth century continued to occupy an important place in the historiography of the last third of the twentieth century, notwithstanding the new emphases on cultural history. And political history continued to work closely with documentary sources.

Nevertheless, as seen above in recent histories of the French Revolution, greater emphasis was given to cultural factors and to language. Something similar has happened in the studies of the Holocaust. Earlier studies had seen the destruction of the Jews as a vast and complex administrative process, as Raoul Hilberg had described it in *The Destruction of European Jews* (1961), carried out from their desks by bureaucrats like Adolf Eichmann acting on orders from above. The long debate between the 'Intentionalists' and 'Functionalists' centered on the question whether the genocide had its roots in the conscious decisions of the

Nazi leadership or developed without a clear plan in the course of the war. Both approaches failed to deal concretely with the Holocaust as it involved perpetrators and victims on the local level. More recent studies such as Christopher Browning's *Ordinary Men: Police Reserve Battalion 101 and the Final Solution in Poland* (1993) have sought to study the killing through the concrete participation of individual perpetrators.

7. Conclusion

The concerted attack on ideas of objectivity and historical method on the part of postmodernist critics has by no means resulted in the abandonment of serious historical study based on research into the sources. Although the borderline between scholarship and literature has become more fluid, as for example in Simon Schama's *Dead Certainties: Unwarranted Speculations* (1991) which points at the tremendous problems in telling a truthful history, few historians, whether in the areas of political, social, or cultural history, would abandon the commitment to honestly reconstructing the past. At the same time, the confidence of nineteenth-century professional historians in the possibility of telling definitely *wie es eigentlich gewesen* has yielded to the recognition that the same sources and the same set of events lend themselves to a variety of interpretations, but that these interpretations do not necessarily disprove one another but cast different perspectives on a subject. The years 1970–2000 have seen the collapse of illusions dear to modern historical thought until then: the linear progression of Western history and the accumulation of historical knowledge. The scope of historical study has been expanded immensely. The concentration on centers of political power and macroeconomic and macrosocial processes has been supplemented by a new interest in many aspects of life and culture which previously had not entered the imagination of historians, or at least of professional historians, and which require special research strategies. These two scales, the macro and the micro, by no means exclude each other, and in fact, as we have indicated, political and social historians have become increasingly aware of the role of culture, and cultural historians have become aware of the social and political context in which collective human existence takes place.

See also: Gender History; Historicism; Subaltern History

Bibliography

- Ankersmit F 1995 *History and Tropology. The Rise and Fall of Metaphor*. University of California Press, Berkeley, CA
Appleby J, Hunt L, Jacob M 1994 *Telling the Truth about History*. Norton, New York
Barraclough G 1978 *Trends in History*. Holmes & Meier, New York

- Berman A 1988 *From the New Criticism to Deconstruction. The Reception of Structuralism and Post-Structuralism*. University of Illinois Press, Urbana, IL
Geertz C 1973 *The Interpretations of Cultures*. Basic Books, New York
Green A, Troup K 1999 *The Houses of History. A Critical Reader in Twentieth-century History and Theory*. New York University Press, New York
Hunt L 1984 *Politics, Culture and Class in the French Revolution*. University of California Press, Berkeley, CA
Hunt L (ed.) 1989 *New Cultural History*. University of California Press, Berkeley, CA
Iggers G G 1997 *Historiography in the Twentieth Century. From Scientific Objectivity to the Postmodern Challenge*. Wesleyan University Press, Hanover, NH
Jenkins K 1995 *On 'What is History?'*. Routledge, London
Jenkins K (ed.) 1997 *The Postmodern History Reader*. Routledge, London
Lyotard J F 1979 *La Condition postmoderne: rapport sur le savoir*. Editions de Minuit, Paris. (English edn. 1984 *The Postmodern Condition: A Report on Knowledge*. University of Minnesota Press, Minneapolis, MN)
Megill A 1985 *Prophets of Extremity. Nietzsche, Heidegger, Foucault, Derrida*. University of California Press, Berkeley, CA
Novick P 1988 *That Noble Dream. The 'Objectivity Question' and the American Historical Profession*. Harvard University Press, Cambridge, MA
Scott J 1988 *Gender and the Politics of History*. Columbia University Press, New York
Spiegel G 1990 History, historicism, and the social logic of the text in the Middle Ages. *Speculum* 65: 59–86
Stone L 1979 The revival of narrative: Reflections on a new old history. *Past and Present* 85: 3–24
Thompson E P 1963 *The Making of the English Working Class*. Vintage Books, New York
White H 1973 *Metahistory. The Historical Imagination in Nineteenth-century Europe*. Johns Hopkins University Press, Baltimore, MD
White H 1978 *Tropics of Discourse. Essays in Cultural Criticism*. Johns Hopkins University Press, Baltimore, MD
White H 1987 *The Content of the Form, Narrative Discourse and Historical Representation*. Johns Hopkins University Press, Baltimore, MD
Windschuttle K 1997 *The Killing of History: How a Discipline is Being Murdered by Literary Critics and Social Theorists*. Free Press, New York
Wolfe P 1997 History and imperialism: A century of theory from Marx to postcolonialism. *American Historical Review* 102: 388–420

G. G. Iggers

Historiography and Historical Thought: East Asia

1. Features of East Asian Historiography

The art of recalling the past has developed very differently in Europe and East Asia. Whereas, since the time of Herodotus and Thucydides, Western

historiography has been centered on histories written by individuals for individuals, since ancient times East Asian historiography has developed a historical culture pivoting on public historiography. The distinctive characteristic of East Asian historiography could be described as publicly sponsored history, a tradition that continues today as a *basso continuo*.

'East Asia' as a historic-cultural concept, comprises present-day China, Japan, Korea, and Vietnam, where history developed around China, which overwhelmed the peripheral countries with its physical size, and political and cultural ascendancy. This historic East Asia has been termed 'the Chinese world order' or a system of international relations with China at the center. The system reached Korea and Vietnam in the early Han period (202 BC–9 AD), and Japan in the first century AD.

This Sino-centric system also formed the common cultural heritage of East Asia: the classical Chinese language, and Chinese writing system; the Confucian social and familial order; Buddhism through Chinese translations; the legal and administrative systems, and the art of historiography. The peripheral states of the East Asian world might also be called Sino-graphic cultures, in the sense that they used classical Chinese as a common language, which is, in a sense, comparable with the use of Latin in mediaeval and early modern Europe.

2. Role and Purpose of Historiography

For the most part, the task of historiographical compilation in East Asia was traditionally a state-run project. The 'official history' produced by that compilation, along with materials collected for the purpose, constituted the core of East Asian historiography. It could be said that historiography was the primary cultural undertaking in East Asia. This is in contrast to the cultures of Europe, India, and Islam, where the concentration of cultural power has not been fixed in history.

For 2,000 years, Chinese historiography centered on the 'official history' compiled by each successive dynasty as a state enterprise. Later generations positioned the *Shiji* (*The Records of the Grand Historian*) by Si-ma Qian (145 – ca. 86 BC) as the first official history and since then, 24 official histories have been compiled. A characteristic of these official histories is that they have an encyclopedic tinge; that is, the body of the work originated by Si-ma Qian brought an entire culture, its politics, economics, society, culture, technology, etc., into one unified structure. History was written as a means of comprehensively describing such a world system.

Historiography in East Asia is perhaps equivalent to such 'primary cultural undertakings' as Biblical commentary and the *Corpus Iuris Iustianianus* in the

West, the *Laws of Manu* in India, and the *Koran* in the Muslim world.

The purpose of writing history in East Asia was based on the Chinese philosophical premise that historical facts were the only certain and immutable reality. Chinese metaphysics was not premised on a revealed religion based on the existence of a unique, almighty God, like Judaism, Christianity, and Islam. Rather, it took the world as an ever-changing phenomenon, as presented in the *Yijing* (*The Book of Changes*). Therefore, it sought immutable reality in history, because human beings could not alter that which had already happened. This belief took history as its axis in Chinese civilization.

This culture of history later spread throughout East Asia in tandem with Confucianism, creating a common historical culture throughout East Asia. The proclamation of this philosophy may be found in Confucius' statement that, 'All the empty words I want to write down are neither so clear nor so startling as seeing their meaning in action' (*Shihji*).

The Chinese had developed their own way of creating an immutable past. In China and Korea, it was standard practice that, once the state's historical compilation bureau had completed compiling the official history of the previous dynasty, the bureau destroyed all the sources it had collected. This was to prevent the revision or rewriting of the official history, for once it was published by the government, the history itself took on the character of a sacred text. The most certain way to endow the official history with the imprimatur of authority was to destroy the sources on which it had been based. During the Korean Chosŏn dynasty (1392–1910), for example, the source materials were burned after use. When that was done, the account embodied in the official histories became the facts of history.

In this fashion, East Asian cultures preserved the ideal that history was the sole immutable basis for human judgment. The biographies that comprise over half the material in the official histories maintained this tradition of objective narration in their own way. In the biographies, as in other sections of the official history, they first set forth what they believed to be 'fact,' and following that, the historians added their own evaluation. This vast corpus of biographies forces us to consider why the historians believed the biography to be a necessary part of a history. It is because in a culture that lacks a unitary supreme being, the records of the lives of eminent individuals are the only true sacred texts.

3. Historical Objectivity

The tradition was maintained of seeking objectivity in historical narration in the lives of the historians who wrote it even more than the narration itself. We can

find an episode that presents the historian as a lofty figure in the life of a fifteenth-century Korean historian.

It is recorded that in 1431, as the compilation of the *T'aejong sillok* (the *Veritable Records of King T'aejong*) was nearing completion, his successor King Sejong (1418–1450) asked the compilers to show him their work in advance. 'In the previous dynasty, every monarch personally reviewed the veritable records of his predecessor; but King T'aejong did not review the Veritable Records of King T'aejo.' Sejong's senior ministers replied that, 'If Your Majesty were to review [the work in progress], later monarchs would surely revise [the historians' work]. Then, [future] historians would suspect that the monarch might look at the draft, and they would inevitably fail to record the facts completely. Then, how would we transmit [facts] faithfully for the future?'

In traditional East Asia, the role of the past was to serve as a normative history. This forms an interesting contrast to Western historical practices, where it evolved as a cognitive discipline. That is, in the West, the historian found his *raison d'être* in rewriting the past. The discipline of history developed as a competition among the interpretations and approaches of different historians.

'History' in modern East Asia is generally parsed with the two-character compound Chinese word *lishi* (Japanese, *rekishi*; Korean, *yoksa*), but until the middle of the nineteenth century it was expressed in the single-character term *shi* (Japanese, *shi*; Korean, *sa*). This word *shi* itself originally denoted the historian himself, and only later, by analogy, did it come to refer to the product of the historian's craft, the written work of history itself. This is particularly important when we note that in the languages of Europe, it was precisely the reverse: The word 'historian' was formed to refer to the maker of the written 'history.'

It is a notable characteristic of historical consciousness in East Asia that it tends to focus far more on the attitude of the historian, than on actual works of history themselves.

4. Development of East Asian Historiography

4.1 China

In China, Sima Qian substantially established the tradition of Chinese historiography by inventing his own format for compilation. He succeeded his father as the grand historian of the Han emperor, and with his access to court archives, he completed the *Shiji* (The *Record of the Grand Historian*) in 130 volumes. It covers a history of three thousands years from the (legendary) Huang-di dynasty to his own period, and covers China and Central and South Asia. The format consists of five sections: 'Benji,' or imperial annals (12 volumes); 'Biao,' a chronology of the successive

dynasties (10 volumes); 'Shu,' treatises on economy, technology, astronomy, etc., (eight volumes); 'Shi-jia,' annals of feudal states and nobles (30 volumes); and 'Liezhuàn,' biographies of famous persons and foreign peoples (70 volumes). Sima's format is encyclopedic in nature, systematizing the world in a single, unified structure; his historiographical format dominated East Asian historiography until the mid-nineteenth century.

Ban Gu (32–92 BC) followed Sima Qian's work, but limited himself to the history of a single dynasty, which he completed as the *Hanshu* (History of the [former] Han dynasty). His practice, that each Chinese dynasty compiled the history of its predecessor, became standard for all of the following official histories. It was in the Tang dynasty (618–906) that historical culture truly flourished, as history became an independent subject in the Chinese classification of learning. Historians were closely linked with governmental bureaucracy and their team method of compilation started in this period and continued until the twentieth century.

A new format for compiling historiography was proposed by Sima Guang (1019–1086) in his 294-volume *Zizhi tongjian* (The *Comprehensive Mirror for Aid in Government*) in 1084. It was an extensive historiography that employed the method of chronological order to help readers find certain historical events through dynasties. It covers sixteen dynasties of 1,362 years and includes information from dynastic histories, chronicles, biographies, novels, etc. It is necessary to refer here also to the Hanxue (the Han School), the development of which established historical study based on textual criticism of historical documents in the Qing dynasty. Qian Daxin (1728–1804) was one of the great figures among them, whose *Nianershi kaoyi* (Verification of the text of the 22 histories) was an examination of characters and language, dates and places, genealogies and institutions in the text of the 22 official histories.

4.2 Japan

The idea of official history spread from China to the other East Asian countries that compiled their own official histories, modeled on the Chinese format and written in classical Chinese, not in their native languages.

The first Japanese state historiography was the *Nihon shoki* (The *Chronicle of Japan*, 720) compiled by Prince Toneri, and modeled after the Chinese format. It covers the period from the Age of Gods to the end of the reign of Empress Jitô (686–697), arranged in chronological order, in annalistic form. Toneri used history as a forum to establish Japanese national identity and consciousness. This is in marked contrast to the *Kojiki* (Record of Ancient Matters, 712) by Ôno Yasumaro, which was also a historiography based on a transcription of oral tradition.

This tradition continued until the late nineteenth century. The *Honchō tsugan* (*Comprehensive Mirror of Japan*, 1644–1670), compiled by Hayashi Razan and his son Hayashi Gahō, was also in the same line of official history sponsored by the Tokugawa shogunate, and the *Dai Nihonshi* (*the History of Great Japan*, compiled 1657–1906) was compiled under the support of the Mito clan.

Apart from the historiographical tradition, it is important to add such auxiliary sciences as bibliography, classics, textual criticism, and philology, which had developed in Japan to the same levels as in the West by the nineteenth century. These originated, in Japanese tradition, from the *Chronicle of Japan* (720), of handing down variant texts or conflicting readings and traditions intact to posterity. In this important respect, China differed markedly from Japan. In China, history was transmitted as an authentic, authoritative text, which admitted no variant. This difference is richly illustrated by the work of Hanawa Hokiichi (1746–1821). His *Gunsho ruijū* (*Classified collection of Japanese classics and documents*), begun in 1779 and still in progress, first appeared in 1819, with 1,270 titles in 530 volumes; another 2,103 titles in 1,150 volumes appeared in 1822. The objective was not to establish the authoritative text and obliterate the others, but to make available all the variant texts, which later generations might judge. This collection compares favorably with such European endeavours as the *Monumenta Germaniae Historica*, the *Rerum Italicarum Scriptores*, etc.

The Meiji government, established in 1868, which was responsible for the restoration of the emperor as the focus of sovereignty, and led the subsequent modernization of Japan, also continued the tradition of ‘official histories,’ founding the Office for the Compilation of National History. The aim here was to create an official history of Japan, but for various reasons the objective was amended to that of compiling historical documents, and the original idea of compiling a national history was abandoned. The work of the Office continues today in its successor, the Historiographical Institute, now part of Tokyo University. Moreover, virtually every prefecture, district, and locality in Japan, even down to the level of individual villages, continues the tradition of official history at the local level today, compiling local historical sources, and publishing local histories. Similarly, many individual government bureaus, as well as corporations, universities, and other corporate entities, regularly commission their own histories, extending the model of ‘official history’ to the private realm.

4.3 Korea

It was around the fourth century that Korea founded its historiographical tradition, though the epoch-making historiography did not appear until the twelfth

century. Kim Pu-sik (1075–1151) compiled the *Samguk sagi* (*History of the Three Kingdoms*) in 1145, the oldest extant history of Korea, covering from antiquity to the decline of unified Silla in 935. It follows the format of Sima Quan’s official history as is shown by the name *sagi* (the Korean pronunciation of *Shiji*), and he intended to provide moral lessons for posterity. Iryōn (1206–1289), a high-rank Korean Buddhist monk, compiled the *Samguk yusa* (*Memorabilia of the three Kingdoms*) in 1281, which is the only extant history of Korea before the Chosōn dynasty (1392–1910), together with the *Samguk sagi*. This work is unique in the sense that he wrote it privately, including myths, tales, and social issues, and even refers to the sources he used.

So Kō-jōng (1420–1488), compiled the *Tongguk t’onggam* (*Comprehensive Mirror of Korea*) as a Korean official history in 1485, modeled on Sima Guang’s *Zizhi tongjian*. It covers the period from ancient times to the end of the Koryō dynasty in 1392. This work was compiled under the influence of neo-Confucianism, and reflects the moralistic purpose of praising good and condemning evil. An Chōng-bok (1712–1791) likewise published his *Tongsa kangmok* (*Outline of Korean History*), 20 volumes, in 1778, a chronological treatment of Korean history from Tan’gun, the legendary founder of Korea to the decline of Koryō in 1392.

4.4 Vietnam

In Vietnam, just as in Japan and Korea, some official histories were published including *Dai Viet Su Ky Toan Thu* (*Complete Book of the Records of Vietnam*) in 1479, which is based on the style of the *Zizhi tongjian* of Sima Guang.

5. Historiology in East Asia

The study of historical theory in East Asia began with Liu Zhiji’s *Shitong* (*Comprehensive Historiography*), written in China in the eighth century (708). It was the first thorough treatise in Chinese—or any other language, for that matter—on historical criticism, which also constituted a history of Chinese historians and a theory of historiography, and deals with such questions as selection and criticism of historical materials, historical objectivity, judgment in history, historical causality, etc. The work was a constant reference for Chinese historians until the nineteenth century, and gave rise to the meta-historical tradition in Chinese historiography.

Liu’s theory of history was further developed by Zheng Qiao (1104–?1162) in his *Tongzhi* (*Comprehensive Monographs*, 200 volumes). In the introduction to the *Tongzhi*, he proposed his core idea of ‘synthesis’ and ‘inter-relatedness and causality,’ through which he argued the importance of the continuity of his-

torical development. Based on this theory he completed the voluminous *Tongzhi*, including the 'Twenty Summaries,' his noteworthy encyclopedic chapters.

Zhang Xuecheng (1738–1801) wrote his *Wenshi tongyi* (*General Survey of Literature and History*), a work on the theory and history of historiography (only published in 1920, long after his death). The first chapter starts with 'The Six Classics are all history,' in which he expressed his philosophy of history, that the very idea of the Sacred (*tao*) can only be known through men's actions (*shi*), which are recorded in the *Six Classics*.

It is thought provoking to realize how few works of historical criticism were written outside China, in comparison with the voluminous amounts of historical works. It signifies that traditional East Asian historiography was normative in nature, in the sense that it required little discussion of the nature of history. It is also necessary to refer to the fact that there were relatively few works of metaphysical interpretation of the course of history—save perhaps Wang Fuzhi's (1619–1692) discussion of historical logic and historical tendency in work published only after his death. East Asians valued detailed precision in history, rather than the philosophical interpretations that were developed in Europe. Some exceptions, informed by Buddhism, were found in medieval Japan. Jien (1155–1225), a prominent Japanese Buddhist cleric, wrote his self-deprecatingly titled *Gukanshō* (*Notes of the Views of a Fool*, 1220), in which he proposed his philosophy of history with the term *dōri*, an idea of historical inevitability, through which he explained historical processes in terms of the state of development and proposed his own periodization. Similarly, a century later Kitabatake Chikafusa proposed in his *Jinnō Shōtōki* (*Chronicle of the True Descent of the Divine Sovereigns*, 1339) that Japan is a land founded by the gods and maintained by the blessings of an unbroken lineage of divinely descended emperors.

6. The Rise of Modern Historiography in East Asia

Western concepts of 'civilization and enlightenment' were introduced systematically to East Asia beginning in the mid-nineteenth century had a transformative effect on notions of history and historiography as well. The movement toward 'modern' and 'scientific' history in East Asia began with a variety of intellectual experiments among Japanese scholars in the 1870s and 1880s. As Korean and Chinese intellectuals ventured to Japan toward the end of the nineteenth century, 'modern' Western notions and methods of historical inquiry and explanation gradually filtered across the entire East Asian region.

It is worth special mention that Western cultures and technologies including historiography were

introduced through translations of European and American works into Japanese. Japanese intellectuals created their own neologisms to translate Western terms by combining two or three Chinese characters together. Those newly invented terms in Chinese characters were later exported to China and Korea, and in consequence, East Asian scholars now share most academic terms in the liberal arts and sciences.

The very fact that such translations were feasible, and successfully received, demonstrates the stage that Japanese historiography had achieved. By introducing Western historiography through translation, rather than keeping it as the preserve of a few language specialists, Western history and historiography were transformed into common intellectual property for all East Asian historians, and, in particular, made accessible to those who worked on their national histories.

As far as historiography is concerned, there already existed a traditional Japanese historiography, and the auxiliary sciences made possible the introduction and importation of Western historiography. Native pioneers in the introduction of Western historiography, historians like Shigeno Yasutsugu (1827–1910) and Kume Kunitake (1839–1931), were initially trained in *Kōshōgaku* tradition (Chinese, *kaozhengxue*: textual criticism, which originated from the so-called 'Han School' in the Qing period), and were thus well prepared to accept the empirical methods of Leopold von Ranke. They were particularly receptive to his notions of a rigorously critical approach to sources and the avoidance of value judgments, when his student Ludwig Riess, who was invited to Tokyo Imperial University as a professor of history in 1887, introduced Ranke's ideas to Japan.

The acquisition and acceptance of European historical theory was started by Japanese translations of leading contemporary European works on historical method and theory, such as those by J. G. Droysen, E. Bernheim and C. V. Langlois and C. Seignobos. Under the strong influence of Bernheim, Tsuboi Kumezō published his *Shigaku kenkyūhō* (*Methods of Historical Research*, 1903) in Japan, and Liang Qichao (1873–1929) published the *Zhongguo lishi yanjiufa* (*Methods of Chinese Historical Research*, 1922) in China.

'To study history' in traditional East Asia had meant 'to become conversant with the official Chinese histories,' and the introduction of Western ideas of historiography opened new horizons for students of history, who began to inquire, 'what are the historical facts,' in the sense of going beyond the historical narrative to examine the underlying facts themselves. It was an indication of the opening of modern historical scholarship in East Asia, in which the introduction of Western learning played an important role.

A typical example of this consciousness is found in Shigeno Yasutsugu, nicknamed 'Doctor Obliteration.'

Applying the European historical method to Japanese history, he re-examined the *Taiheiki* (*Chronicle of the Great Peace*, 1370–1371), and proved that the fourteenth-century military commander Kojima Takanori never existed.

In China, Gu Jiegang (1893–1980) was one of the most eminent figures, whose *Gushibian* (*Critiques of Ancient History*, seven volumes, 1926–1941) contributed greatly to the modernization of Chinese historiography by taking a finely honed critical scalpel to the world of ancient Chinese legend, which had long been accepted as fact. His school was named ‘ancient doubters.’ His method was rooted in the Qing dynasty (1644–1911) school of ancient doubters; however, it was at the same time backed with the new European historical method.

Yi Pyŏng-do (1896–1989) was the first Korean historian to have received training in the modern historical method. He stressed critical and objective methods of treating historical sources in his *Han’guk-sa taegwan* (*Grand Survey of Korean History*, 1963).

With the introduction of Western culture to Japan, history became re-institutionalized from the mid-nineteenth century onwards, from the governmental section to the newly established Imperial University. However, it should not be forgotten that the introduction of Western historiography reduced East Asian historiography from its former position as a state enterprise to merely a single academic field within the university. Individual historians started to have their own ‘specialty’ in order to survive in this new ‘university’ setting, and in order to distinguish themselves from eminent amateur historians, their researches were published as monographs in the newly established historical journals.

On the one hand, in the West, the ‘professionalization’ of historical research resulting in the establishment of history as an independent university academic field was a great achievement for this institutionally neglected discipline. In East Asia, on the other hand, the introduction of modern Western historical research heralded the end of the prestigious traditional East Asian historiography, which had aimed at a comprehensive description of the entire world. Since then, the subsequent evolution of East Asian historiography has developed largely along the lines of ‘modern’ historiography, in the same basic methodological and theoretical terms as the rest of the ‘advanced’ world.

7. Development of Modern Historiography

Since modern historiography was introduced to East Asia, its development varied according to the situation in each country.

In China, along with Gu Jiegang’s ancient doubters, newly discovered archaeological results urged a new

perspective of ancient history on Chinese historians. In particular, the discovery of oracle-bone inscriptions laid the foundations for the new Chinese ancient history led by such historians as Luo Zhenyu (1866–1940) and Wang Guowei (1877–1927). These new works were later synthesized with Marxist materialism by Guo Moruo (1892–1978). His *Zongguo gudai shehui yanjiu* (*Study of Ancient Chinese Society*, 1930) is the first Marxist work on ancient Chinese society. Thereafter, especially with the establishment of the People’s Republic of China in 1949, Marxism became dominant among Chinese historians. Traditional historiographical achievements were mostly replaced by reverse interpretations, a trend which was highlighted in the period of the Cultural Revolution (1966–1976). After the 1990s, the worldwide information revolution and the introduction of the latest historical methods of social history are changing Chinese historiography, opening it to the rest of the world.

In Japan, soon after the introduction of modern European historiography, the emperor-centered state promulgated a new political orthodoxy, and began to restrict intellectual freedom, especially in the study of ancient Japanese history. Since the 1930s, social and economic history began to apply the Marxist framework to Japanese history, mostly to the study of Japanese capitalism. After World War II, Marxism emerged as the dominant historiographical school, and it was only in the 1970s that Japanese historians began to accept the idea of social history as their latest historical method from France, Germany, and England. Since the late 1980s, many Marxist historians quietly shifted from socioeconomic history to social history.

8. Metamorphoses of Public History

Currently, in East Asia, the aura of official history still floats on the surface of peoples’ collective historiographical consciousness. The past is still seen as the trustworthiest ‘mirror on mankind.’ What is being sought from official historiography is not ‘a’ history, but ‘the’ history, the reliability of which, people undoubtedly believe, is guaranteed by public editorship.

In China, 24 official dynastic histories have been published and the tradition of public history continues to the present day. The *Qingshigao* (*Draft of the History of the Qing Dynasty*) was completed in 1927, and was finally published in Beijing in 1977. This applies to local history too; the publication of prefectural histories sponsored by local governments began in the Ming dynasty (1368–1644), and continues even in contemporary China.

In Japan, this idea is seen in school history textbooks and local histories. The Japanese Ministry of Edu-

cation vets all school textbooks, the idea of which seems a reincarnation of official history as far as history textbooks are concerned. Japanese schools teach students 'what' rather than 'why.' They cram students' heads with a fixed course of history – Japanese history, world history, whatever – without much attempt to have students recognize that there are differing viewpoints. Forty-two Japanese prefectures out of 47 have compiled their own prefectural histories – some of which have been compiled two or three times over the last 120 years – and there is hardly a city, town, or village anywhere in Japan that has not done likewise. These local histories are premised on the traditional historiographical notion that there is a historical truth that is embodied in authoritative narratives.

Korea also continued to chronicle an official history and this tradition is alive in present-day South Korea. The 25-volume *History of the Republic of Korea*, whose publication was completed in 1979 by the National History Compilation Committee (established in 1948), is a state-produced history of the homeland.

Since the late nineteenth century, the practice of history has undergone professionalization and specialization on a worldwide scale, and East Asia too has been drawn into this whirlpool. Since then the style of historiographical practice in East Asia has, in most respects, been the same as that in Western countries. The work of the historian, as an individual, is to write many research papers and books. However, many historians, as far as historians working on their national history are concerned, are also at the same time involved in publicly sponsored historiography; countless university-based Japanese historians, for example, serve on the editorial boards of local and prefectural history projects. We can thus observe the continuing 'coexistence' rather than 'conflict' of the two differing traditions of historical study.

To put it another way, East Asian historians simultaneously inhabit two worlds of historiography, one a world of Rankean (or traditional East Asian) factual objectivity and another world of post-Rankean interpretations 'varied and opposing voices.'

See also: Area and International Studies: Development in Southeast Asia; China: Sociocultural Aspects; East Asia, Religions of; East Asian Studies: Politics; East Asian Studies: Society; Japan: Sociocultural Aspects; Korea: Sociocultural Aspects; South Asia: Sociocultural Aspects

Bibliography

- Beasley W G, Pulleyblank E G (eds.) 1961 *Historians of China and Japan*. Oxford University Press, London
de Bary W T, Wing-tsit Chan, Watson B (eds.) 1960 *Sources of Chinese Tradition*. Columbia University Press, New York

- Han Yu-shan 1955 *Elements of Chinese Historiography*. W H Hewley, Hollywood, California
Ku Chieh-kang (Gu Jiegang) 1931 *The Autobiography of a Chinese Historian*. Trans. Hummel A W, E J Brill Ltd., Leyden
Maruyama M 1974 *Studies in the Intellectual History in Tokugawa Japan*. Tokyo University Press, Tokyo
Nivison D S 1966 *The Life and Thought of Chang Hsueh-ch'eng*. Stanford University Press, Stanford, CA
Schneider L 1971 *A Ku Chieh-kang and China's New History*. University of California Press, Berkeley, CA
Watson B 1958 *Ssu-ma Ch'ien: Grand Historian of China*. Columbia University Press, New York

M. Sato

Historiography and Historical Thought: Indigenous Cultures in the Americas

1. Terminology

History is information handed down from generation to generation. Since information can be inherited in the form of natural objects as well as by culture itself, a basic distinction needs to be drawn between natural history and the history of civilization. The way people develop consciousness of their history is by taking information from their natural environment, their cultural artifacts, and their memory, which is then collectively conceptualized by means of interpreting, reconsidering, and agreeing on its meaning. Historical concepts resulting from this always reflect the people's experience of their natural environment as well as their specific cultural achievements. In exceptional cases a historical concept can be limited to the individual. However, historical research is more concerned with collective concepts, since it is characteristic of human societies that mainly those dominant concepts are handed down. Thus each society cultivates a particular historical concept, which again needs to be handed down in order to become and remain effective. All cultures are alike in inheriting their historical concepts in the form of oral (stories and myths) and figural traditions (monuments). In addition to this, civilizations often use writing (books and inscriptions) as the preferred means for handing down historical concepts to future generations.

2. America

The only people of pre-colonial America to develop a fully functioning writing system were the Lowland Maya. They used the written records to maintain knowledge of a wide thematic range and thus stored and conserved an enormous amount of cultural

information from which they formed their own distinctive historical concept. Rather incomplete writing systems (picture-writing) were characteristic of their northern and western neighbors—that is in historical order the Epi-Olmecs, the Zapotecs, the Mixtecs, and the Aztecs. Taken together, these cultures form the core of the cultural area Mesoamerica, characterized by similar, and in many respects uniform, conceptions of history, and therefore can be treated to all purposes of this article as a unit.

On their first contacts with indigenous peoples of the Americas, the European conquerors came across various forms of picture writing that was used to note down historical events. This even occurred in some areas of North America, predominantly in the eastern woodlands and on the plains and prairies west of the Mississippi. South America was in this respect quite the reverse, even though pictorial representations in painting (Moche, Nasca), weaving (Paracas, Nasca), and rock engraving (Nasca) were highly elaborated.

The description in Sect. 3 is mainly restricted to Mesoamerica, which is distinguished from all other cultural areas by its emergent writing systems.

3. *Components of Mesoamerican History*

The Mesoamerican *Weltbild* shows three basic components, inasmuch as the world is conceived of as consisting in discrete and ordered entities, often arranged in cycles. Connections and explanations of meaning beyond the historical facts make use of analogy, i.e., interpreting something that is identical or similar in form to be also identical in substance or related in a causal way. Another method is the reification of originally mental constructs and their material (picture-written) form of expression. This is particularly significant if information is handed down over a long time and cannot be clarified independently from written or pictorial sources. These and other methods of interpretation were operating in historical traditions, which is why they contain to a great extent patterns, topics, and stereotypes extraneous to the historical facts.

Mesoamerican cultures have also developed methods of dating. A cyclical and public calendar based on the solar year of 365 days, in addition to a ritual 260-day almanac for private and personal use that is likewise cyclical and parallels the solar calendar, are the foremost methods of dating historical events. The Lowland Maya included in their historical writing the calculation of temporary celestial phenomena like the lunar, Venus and Mars cycles, and observed the positions of planets in constellations of fixed stars; comparable to the Zodiacs of the Mediterranean cultures. They even managed the complex arithmetic of interrelations between some of these celestial cycles. Yet they accomplished only the arithmetic calculation

of it, but never presented a geometrical model of the world's system corresponding to the Ptolemaic and Copernican models.

4. *Historical Topics*

By means of chronology Mesoamerican people created histories (theogonies and myths) of anthropomorphic and theriomorphic deities. This is conveyed in detail particularly for the Aztecs through the *Legend of the Suns* (Lehmann 1938) and the famous calendar stone. Mesoamerican people believed in the existence of four or five successive world ages, each with duration of a few thousand years. Each age is quite distinct with different creatures inhabiting a world, which is then destroyed in cataclysm by water, fire, and air (wind)—the reader may note the analogy to the ancient Greek theory of the elements. A new creation marks the beginning of each world age. Not even the contemporary world age of the Aztecs around 1500 was considered to be of unlimited duration. Within the chronological framework of this age man appears relatively late, and more detailed chronological units (years, year bundles of 52 years) apply to his history. Yet Mesoamerican people connect themselves occasionally, and in a speculative way with the world of their deities anteceding the human epoch. This holds especially true for the ruling dynasties of the Maya who very often consider themselves to be descendants of the gods of preceding world ages (Palenque). Such a Mayan world age lasts exactly 1,872,000 days, which is about 5000 years. The duration of the world ages in Central Mexico is not exactly known, since their chronology is technically less developed. The interrelation of the people with their pantheon becomes, furthermore, apparent in the tutelary deities of each tribal group. These are commonly organized in city-states during Postclassic times: Quetzalcoatl for the Toltecs (capital: Tollan) around AD 900–1100 and for the Chololtecs (capital: Cholula) around AD 1000–1500, Tohil for the Quiché (capital: Utatlán) around AD 1000–1500, and Huitzilopochtli for the Aztecs (capital: Mexico) around AD 1350–1500. In this manner origin-related, trans-historical interrelations take form in a similar way as they are known for the civilizations in the Old World (Sumer, Ancient Egypt, Babylon, Jewish culture of the Old Testament).

Tribes with their leaders, the ruling dynasties arising from them during the process of getting sedentary and their tutelary deities are the main actors of Mesoamerican history. Rulers are described with their most important *rites de passage* (birth, inauguration, marriage, procreation of offspring, death), martial triumphs, and—to a lesser extent—civilizing deeds (introduction of cultivated plants or technical skills). There exists the idea among Mesoamerican tribes that they originated from a distinct place, which is the tree

of Apoala for the Mixtecs, the Seven Caves or the island of Aztlan in the center of a lake for the Central Mexicans and the Totonacs; and this origin is placed in an apparently rather historical not very remote time; according to the Aztec main tradition this happened in the year AD 1064. In some areas, the version of an origin in an even more historical setting, namely the city of Tollan can be encountered (Cakchiquel). This Tollan was the capital of the famous Toltecs, who were actually a historical people, admired for their skills and accomplishments. Migration myths, apart from being descriptions of historical events, which are to a certain extent factual—as far as can be concluded from language history and geography—seem to have functioned as a means of political legitimization. The successfully migrating and thus surviving tribal group legitimizes its right to rule over the traversed territories and the cities within, or at least to exert indirect political control in case the areas in question are too distant for factual dominance. After settling down, topics such as marriage alliances, conspiracies, insidious martial acts, and heroic self-sacrifice appear in the form of historical episodes. It is in these later times, when major wars between the different Mesoamerican polities become a subject of history. In Central Mexico it is learned that the Mexicans throw off the yoke of Tepanec rule by Azcapotzalco and gain completely independent status, prestige and property of lands. In the Southern Maya Lowlands the city-states of Calakmul and Tikal are at war with each other for hegemony of the region; and the Guatemalan Highlands are the scene of frequent warfare between the Mam, Quiché, and Cakchiquel. Since the range of historical accounts rarely transgresses the own polity and is immediate neighbors and never exceeds the limits of the Mesoamerican core area, 'world history' from the precolonial Indian perspective is very limited. Nevertheless this compares, albeit on a more narrow basis, with antique and medieval concepts of the world in Near Eastern and European cultures. No statements revealing knowledge or speculation about countries and peoples beyond this horizon can be found, even though archaeological investigations indicate external contacts in the form of trade relations with peoples far outside Mesoamerica, namely with the North American Southwest, with Panama, and with Ecuador. This seemingly odd and willful limitation of the historical horizon is characteristic for many early states and for most chiefdoms worldwide.

It is men who predominate the historical records as well as the actual political scene, and these are generally members of the nobility as all Mesoamerican societies were ranked. The historical records give special emphasis to particularly successful individuals: such as to ruler '8 Deer Tiger Claw' among the Mixtecs or to the rulers 'Pakal' of Palenque and 'Kakupakal' of Chichén Itzá among the Lowland Maya. In addition to politically active members of the ruling nobility,

historical records of the Maya and the Aztec convey a few names of artists (painters, sculptors, poets, and performing artists), thus demonstrating to some extent interest in cultural and art history.

Taking into account that political history is predominantly the subject of men, the reader's attention must be drawn to the occasional significance of women among the Lowland Maya, especially in the city-states of Palenque and Yaxchilán around AD 600/700. These women presumably never equaled men in rank, but gained prominence in their patrilineal societies only in order to substitute for a male heir and to uphold dynastic continuity. It remains unclear whether this role involved real implementation of power or whether it was confined to representational duties, although—in keeping with ethnological theory of gender studies—a rather representative and symbolic function is suggested. Apart from this, noble women appear as objects in marriage alliances that serve to enlarge the ruling territory and become instruments of political provocation (loot, rape, and sacrificial death).

5. *Indian Reflections on History*

It is impossible to give a definite description of historical thinking among the Mesoamerican Indians, since no explicit statements related to this can be found in the historical sources. It is, however, manifest that man is controlled and directed by deities to whom, he for his part, is compelled to pay tribute in the form of prayers, offerings, and cult (there are 18 major monthly feasts per year) in order to maintain and ensure the world's existence. This is most evident in the Aztec's dogmatic obligation of invigorating the sun on its course with sacrificed human hearts. This compulsion to human sacrifice led to those sacrificial excesses that were later witnessed by the Spanish conquerors. Prior to the conquest, several Indian philosophers had already criticized and perceived the practice of human sacrifice as incorrect by expressing their desire for the idealized Toltec times (around AD 1000) to return, when people sacrificed flowers and animals instead of human beings.

The general Mesoamerican outlook on life is pessimistic and fatalistic, which manifests itself especially in the divinatory rituals of the Maya, Zapotec, and Aztecs where negative prophecies are outnumbered by the indifferent or positive ones. The collective fatalistic behavior of the colonial Yucatec Maya and—two centuries later—of the Itzá concerning their own political fate might serve as an example of this common way of thinking. However, it needs to be taken into account that the Yucatec case could well be the result of traumatic experiences during the Spanish conquest; since only colonial sources give information on this topic, it is difficult to decide, whether fatalism is a traditional or an adopted behavior. Human life is

considered to be hard and fateful (like in Old Jewish faith), controlled by deities, and improvement can only be gained by religious manipulation. Consolatory is hence the Aztec concept of the otherworld, where sweet paradise awaits those who endure dangerous labor and hardship on earth, favoring warriors and childbearing women. Yet we lack a theologically differentiated comprehension of the Aztec idea of an afterlife, since this is not further described in the historical accounts. The Lowland Maya left more written and pictorial documents regarding this topic, although these have not been systematically analyzed thus far.

Rarely, reflections on antecedent cultures can be found among the descriptive historical traditions of the Central Mexican Aztecs. The major city Teotihuacan, which had been abandoned prior to Aztec times, is taken by myth as the place of origin of the gods. Earlier cultures are remembered by conserving their artifacts and imitating their products. The idealization of such a past culture—the Toltecs—has been mentioned earlier.

6. Traditions and Sources

The numerous monuments from lasting material (stone), although not yet completely understood, are an important source for historical thinking. They date back as far as 1000 BC: Beginning with sculptures portraying Olmec rulers and slabs illustrating Zapotec conquests (so called *danzantes*). The first real inscriptions date from the first two centuries of the Christian era, called the Epi-Olmec period. Especially noteworthy is the stela from La Mojarra, although its long inscription presumably representing the Proto-Zoquean language of the Isthmus region is not fully deciphered yet.

Vase paintings of the Maya with their pictorial and partially narrative depictions and hieroglyphic explanations are one of the chief sources for reconstructing the Mesoamerican classic from AD 300–800 AD. These vessels show painted scenes from royal palaces, royal visits, payment of tribute, or the hunting of deer and birds. All illustrations do probably depict scenes from the otherworld. Since daily life can be assumed to have served as the *calque* for the concept of this otherworld, it is possible to reverse this relation in deducing every-day life of the Maya from the vase paintings. Vase paintings have been similarly exploited for reconstructing daily life of the South-American culture of Moche, which was contemporaneous but unrelated to the Maya.

Further sources of historical records are manuscripts with historical contents, which—due to their fragility—generally do not date back too far. About a dozen precolonial manuscripts and perhaps a hundred copies and extracts compiled during early colonial times have been preserved. Most are of Central

Mexican (*Codex Huitzilopochtli*) and Mixtec (*Codex Nuttall*) origin.

Christian missionaries introduced the writing with the Latin alphabet in the early colonial times (AD 1520–1650). Written documents of this type are especially numerous in Central Mexico. Basically European in form, and to some extent even in content, the object of these written sources is almost completely related to the colonial situation (lawsuits dealing with inheritances, properties, and privileges). The writings obviously connect pre-Spanish traditions of picture writing with oral traditions and form a new integrated genre. In the Guatemalan highlands (*Popol Vuh*, *Annals of the Cakchiquels*), as well as with the Yucatec Maya (Books of *Chilam Balam*) these sources are not only less numerous but show less European influence. From other Mesoamerican regions only a few historical documents have been preserved, written in the Spanish language by Spaniards who drew on remembered information and interviews. Thus Gonzalo Fernández de Oviedo y Valdés reported on the Nicarao who lived at the southern periphery of Mesoamerica; the Franciscan missionary Juan de Torquemada conveyed comparable traditions of the Pipil; and the accounts of the Mixtecs by Francisco de Burgoa were comprehensive but a little concise.

A major European element in sources, which have been written in the indigenous languages but with Latin letters, is the synthesis regarding world history and natural history. This synthesis can be noted, for instance, in a general chronology drawn from the uncoordinated lists recording royal succession in individual Indian city-states as presented by the *Annals of Quauhtitlan* and the works of Domingo Chimalpahin. As suggested by its subdivision into 12 books, the topics, and the illustrations, the *Historia General* of the Franciscan friar Bernardino de Sahagún obviously resembles medieval European natural history and its antique origins (Pliny the Elder).

7. Problems of Research

Without explanations from the creators themselves, it remains difficult to understand form and symbolism of the great range of historical objects and sources, even though researchers have presented seemingly comprehensive and highly sophisticated interpretations (in particular Eduard Seler around 1924–5). Hieroglyphic captions of Mayan vase paintings, which have been subject to serious research since the 1980s, are—although seemingly well understood—overly standardized (e.g., ‘this is the beaker for cacao belonging to X’) and, therefore, still meager in cultural content and fuzzy in context. Most inscriptions that designate singular objects like ‘earring’ or ‘[sack full of] beans’ are short and do neither contribute to a differentiated understanding of the depicted scene, since they only name the object, nor do they help much in a broader

comprehension of local history, since they only mention the owner or receiver of said object.

Although theoretically achievable, the decipherment of pictorial and hieroglyphic writing systems (Zapotecs, Epi-Olmecs, Maya, Mixtecs, Aztecs, Tetzocans) has not been completed yet and not all documents have been thoroughly studied. The research strategy regarding these documents assumes a uniform paradigm underlying all different regional forms of pictorial and syllabic writing (Maya, Mixtec, Aztec, Tetzocan, etc.). Nevertheless, painstaking and detailed analysis of each document, as well as methodological reasoning in decipherment and interpretation are required. Investigation should strategically begin with thematical complexes that are especially revealing because their cultural contexts are well understood, such as burials at major archaeological sites containing inscribed goods.

Texts from early colonial times (until about 1650) that have been written down in the indigenous languages by means of the Latin alphabet are easier to approach, although they are still not very well understood either. Most of these documents originated with the Aztecs, some of them with the Yucatec and Highland Maya, while there are only very few from other language groups and regions. Whereas the more technical problems of translation can be met by increasing language competence, the main problem of ethnohistory (the technical term for the discipline of history that analyses sources in the writing-systems and languages of the people studied) is to grasp concepts and metaphors, etc. from the *Weltbild* of the Indians. These are then interpreted in terms of European historical thinking, and their authenticity elucidated with regard to events that have actually come to pass.

Devoting more time to research of chronological problems is most likely rewarding, since Mesoamerican chronology is not only of high accuracy but also omnipresent in historical records. For the time of the Maya Classic from AD 300–900 there is not a single year without several exactly datable historical accounts. Not close by region but by time, Central Mexican historiography from approximately AD 600 until the Spanish conquest is structured with about the same amount of dates. However, due to the calendrical system, Central Mexican dates are chronologically ambiguous, since they recur every 52 years. The interrelations of dates in the different calendar systems (of Mixtec and Aztec and between Aztec and Lowland Maya), and the comparison of dates with chronological reconstructions in archaeology, and the absolute dating by way of historical astronomy, making use of permanent celestial phenomena, some day will render it possible to compile a wide-ranging historical chronological chart for the Mesoamerican core area.

Because of Mesoamerica's high civilizing and political complexity as well as its dynamics, the analysis of overarching political systems and their changes is a

very important field of research. A great amount of effort has already gone into acquiring knowledge on the Post-classic times in Central Mexico, Oaxaca, and Northern-Yucatan; the Lowland Maya Classic is currently subject to politico-historical analysis (Nikolai Grube and Simon Martin). No historical events can be reconstructed for the Formative Period, in which the later Mesoamerican states emerged, since it lacks sizable written records. Yet the abundant archaeological findings of this period, including monumental sculpture and buildings, serve for assessing general theories on the emergence of the state. Such theories have made use of the respective Indian cultures, mostly Aztec and Inca, since Adolph Bandelier (1880) classified the Aztecs within the scope of an evolutionary scheme, followed over 100 years later by Stefan Breuer's (1990) more sophisticated analysis. American cultures are particularly relevant for this kind of research because it is assumed that they developed independently of Old World civilizations.

8. *Survey of Other American Regions*

With its politically tight organization and its technical achievements (metallurgy, road system) the South American cultural area of the Central Andes, which was last dominated by the Incas, can be perceived as superior to Mesoamerica. The study of historical concepts is nevertheless here rendered more difficult due to a lack of pre-colonial written records. All records of Indian history originated in early colonial times, are based on oral traditions and the system of knotted cords (*quipu*) only, and have been written mostly by Spaniards (Juan de Betanzos, 1551) or mestizos (Garcilaso de la Vega). They thus show the categories and forms of the European conception of history. This holds true also for writings that are sometimes considered as 'ancient Indian' such as the preciously illustrated *Nueva Corónica y Buen Gobierno* of Guaman Poma de Ayala, which is formally European in its illustrations and text, even though with mainly Indian contents. The anonymous historical drama *Ollantay* is according to its form and subject rather similar to a baroque theater play from Europe than being an expression for Indian thinking, albeit written in a stylistically pure form of *Quechua*, the language of the Inca empire, which makes it appear very Indian. The general assertion alone that something like an 'Andean Historical Concept' exists is not helpful. If more substantial information on the Indian historical concept of this cultural area is to be revealed, progress can be achieved only by concentrating on less spectacular administrative sources (*visitas*) and by diagnostic separation of European and ancient Indian components in all sources. It should also content itself with subjects of manageable size in respect to thematical scope and regional breadth, before proceeding to generalizations.

In conception and structure similar to the Mesoamerican historical tradition are the eighteenth- and nineteenth-century historical traditions of the Lenape (Algonkin language family) and the Indians of the Great Plains in North America. The Lenape (*Walam Olum*, the authenticity is controversial) and other related groups, e.g., Creek, cultivated an oral and possibly picture-written tradition of the tribe's migrations led by certain chiefs; while the Plains Indians depicted individual events from the group's daily life following the year's seasons (*Winter Count*).

See also: Central America: Sociocultural Aspects

Bibliography

- Adorno R 1989 Die visuelle Vermittlung zwischen Oralität und Literalität in einer Anden-Chronik. In: Birgit Scharlau (ed.) *Bild-Wort-Schrift*. Gunter Narr Verlag, Tübingen, Germany, pp. 1–17
- Barrera V, Rendon A, Rendón S 1948 *El libro de los libros de Chilam Balam*. Fondo de Cultura Económica, México [abridged in later editions]
- Breuer S 1990 *Der archaische Staat. Zur Soziologie charismatischer Herrschaft*. D. Reimer Verlag, Berlin
- Brinton D G (ed.) 1885 *The Annals of the Cakchiquels*. (Brinton's Library of Aboriginal American Literature, 6). Philadelphia 1885 [reprinted: AMS Press, New York]
- Crónica M 2000 *Quellenwerke zur alten Geschichte Amerikas*, 14. Gebrüder. Mann Verlag, Berlin
- Eschmann A 1976 *Das religiöse Geschichtsbild der Azteken*. (Indiana, Beiheft 4). Gebrüder Mann Verlag, Berlin
- Ethnohistory 1972–5 In: *Handbook of Middle American Indians*. University of Texas Press, Austin, TX, Vols. 12–15
- Guaman P de A F 1936 [1615] *Nueva Corónica y Buen Gobierno*. (Codex péruvien illustré). (Travaux et mémoires de l'Institut d'Ethnologie, 23). [reprinted: 1968] Institut d'Ethnologie, Paris
- Lehmann W (ed.) 1938 *Die Geschichte der Königreiche von Colhuacan und Mexico*. (Quellenwerke zur alten Geschichte Amerikas, 1). W. Kohlhammer, Stuttgart [rev. edn. by Gerdt Kutscher 1974]
- Marcus J 1992 *Mesoamerican Writing Systems. Propaganda, Myth, and History in Four Ancient Civilizations*. Princeton University Press, Princeton, NJ
- Middendorf E W (ed.) 1890 *Ollanta. Ein Drama der Keshuasprache*. Leipzig, Germany
- Prem H J 1989 *Geschichte Alt-Amerikas*. (Oldenbourg Grundriss der Geschichte, 23). R. Oldenbourg Verlag, Munich, Germany [English edition in preparation]
- Riese B 1997 Zeitstrukturen in Mesoamerika. In: Müller K E, Rüsen J (eds.) *Historische Sinnbildung*. Rowohlt, Reinbeck bei Hamburg, pp. 240–50
- Scharlau B, Münzel M *Qellqay. Mündliche Kultur und Schrifttradition bei Indianern Lateinamerikas*. Campus, Frankfurt am Main, Germany
- Schultze J L (ed.) 1944 *Popol Vuh*. (Quellenwerke zur alten Geschichte Amerikas, 2). W. Kohlhammer Verlag, Stuttgart & Berlin, Germany [reprinted: 1975]
- Seler E 1924–5 *Gesammelte Abhandlungen zur amerikanischen Altertums- und Sprachkunde*. 5 Bände. Berlin [reprinted with additional Index-volume: Akademische Druck- und Verlagsanstalt, Graz 1960–7]

Taylor G 1999 *Ritos y tradiciones de Huarochiri*. (Travaux de l'Institut Français d'Etudes Andines, 116). Paris

Voegelin C (ed.) 1954 *Walam Olum or Red Score*. Indianapolis

B. Riese

Historiography and Historical Thought: Islamic Tradition

'Islamic historiography' is a useful but somewhat misleading term commonly applied to the historical writing produced in the predominantly Muslim societies of Asia and Africa between the rise of Islam (post-622 CE) and the mid-nineteenth century. In contrast, history written after that point may have a distinctive regional or cultural orientation, but increasingly it falls within the post-Enlightenment historiographic discourse established and dominated by Western Europe. The term is useful because it allows us to discern the degree to which the enormous historical literature produced within Muslim-dominated societies and polities embodies an unbroken (albeit complex and constantly changing) tradition of debate over the meaning and interpretation of the past. It is misleading insofar as it beguiles us into thinking that this literature was produced entirely by Muslims, or that it is exclusively grounded in Islamic doctrine and focused on religious issues. In fact non-Muslims produced important works (admittedly a small proportion of the total) within the 'Islamic' historiographic tradition. More importantly, much historical writing by Muslims focuses on decidedly nonsacred topics and presents these topics in a manner that owes little to Islamic religious concepts and concerns.

1. History as a Form of Knowledge and a Profession

Quite apart from the ambiguity of the term, Islamic historiography represents an intriguing paradox. Historical works were produced in great profusion from at least the mid-eighth century on, and it is evident from the number and distribution of manuscripts that these works were widely read. Technical terms for historical writing emerged very early and remained in constant use thereafter: *ta'rikh* (literally 'dating,' hence 'chronicle'), *akhbar* ('recitals of events'), and less commonly, *sirah* ('biography'). '*Ta'rikh*,' the most common term, referred not only to chronologically organized works but to historical writing of every kind, including alphabetically arranged biographical dictionaries. However, even though historical writing permeated medieval Islamic culture and was universally seen as a distinct form of knowledge and literary practice, it never achieved the status of a 'science' (Arabic, '*ilm*')—i.e., it never became a formal

subject of study and training in the scholarly curriculum. Within the normative classification of knowledge constructed by medieval Islamic culture, history remained always on the margins of scholarship and literature. For religious specialists it was ancillary to the core disciplines of legal studies (*fiqh*), Prophetic tradition (*hadith*), and Qur'anic exegesis (*tafsir*). Secular writers might well regard history as *magister vitae*, a trove of practical wisdom, but only rarely did they regard it as the foundation for a systematic investigation of politics and society. Even in the humble roles of handmaid and schoolmaster, however, history still generated many works of great literary and methodological sophistication. The political (and occasionally social and economic) analyses of the best Muslim historians yield nothing to their counterparts in the ancient, medieval, and Renaissance West. All this suggests that history played a much more fundamental role in defining Islamic self-understanding than that culture's spokesmen knew how to articulate. History showed Muslims how God was working his purposes out in the deeds and sufferings of the community of believers; in particular, it measured the fidelity (or lack thereof) of the believers to the revelation brought by Muhammad. History challenged Muslims to measure their progress on the path to salvation or perdition as the inevitable Day of Judgment approached.

Historians were in effect self-taught amateurs, bureaucrats, or religious scholars who composed *ta'rikh* in their spare time. They learned how to do it not through formal study with established masters, but simply by reading the works of other historians and adapting the forms and approaches of these writings to their own purposes. Some histories were written at the behest of rulers anxious that their deeds be properly memorialized for future generations, but even these do not represent a true professionalization of history. We do find official court historiographers under the Ottomans between the sixteenth and mid-nineteenth centuries, but for the most part the job was handed over on an *ad hoc* basis to any member of the ruler's entourage who was thought to possess sufficient political acumen and literary skill.

Two outstanding scholars did attempt to make history into a generally recognized formal science, but in spite of their great prestige and influence, neither achieved his goal. The Qur'anic exegete, jurist, and historian Abu Ja'far al-Tabari (838–923) tried to demonstrate in his *Chronicle of Prophets and Kings* (*Ta'rikh al-rusul wa'l-muluk*) that historical reports could be deployed according to the rigorous criteria applied to *hadith* (the dicta ascribed to the Prophet). However, very few followed him down that path, even though his reconstruction both of pre-Islamic history and of Islam's first two centuries provided an authoritative framework for almost all later chroniclers. Far from adopting al-Tabari's strict method, most of his successors simply condensed and paraphrased his work. Five centuries later, the North African Ibn

Khaldun (1332–1406)—a gifted legal scholar and logician as well as a historian—argued in his famous *Muqaddimah* that history was the science of society and as such ought to be treated as a branch of philosophy. Again, his argument did not suffice to move historiography up to that exalted status, in spite of the substantial circulation and influence that his work enjoyed.

2. The Producers of Historical Knowledge

As we have seen in Sect. 1, history was written (and read) by highly literate men with some time on their hands. Given the social and educational structures of medieval Muslim societies, that means it was written chiefly by chancery secretaries and religious scholars (Arabic, '*ulama*'). It is tempting to describe these two groups in terms of well-defined ideal types and thus to posit opposing historical perspectives of palace and mosque. In reality these perspectives overlapped in complex and constantly changing ways.

Throughout Islamic times the '*ulama*' were given state appointments as judges or mosque officials, though it is true that many of the most prestigious scholars prized their independence and refused to accept such positions. But after the mid-eleventh century, even the independent-minded found it difficult to refuse professorships in the burgeoning religious colleges (*madrasahs*) that were endowed and to some degree controlled by the rulers and their senior military commanders. However scrupulous they might be, scholars could hardly avoid some entanglement with the princely courts. Quite apart from questions of material interest, most also had a profound fear of social disorder—as a popular dictum had it, one night of anarchy was worse than sixty years of tyranny. Thus they seldom challenged or even questioned the underlying political order of their societies, however severely they might criticize particular rulers.

On the other hand, their relations with the princely courts were always ambivalent. They knew that rulers came and went like the clouds. They might collaborate with the regime in power but rarely identified themselves closely with it. Their job was to embody Islamic beliefs and values in their own lives and to propagate these within society at large. Religious scholars came from a wide range of social backgrounds and interacted constantly with the urban population. They indeed derived much of their prestige (the only solid basis of their social influence) from their accessibility to all.

As historians, then, the religious scholars were the spokesmen for the community and its concerns. As such, they stood not so much against the princely courts as apart from them. They naturally placed the rulers and their doings at center stage in their historical writing, since the rulers were the heads of the Muslim

community within the territories they ruled—responsible not only for defense, public order, and justice, but also for ensuring that Islamic religious practices and moral norms were maintained. Without a Muslim ruler, however incompetent or corrupt, an authentic Islamic way of life was impossible. However, they did not regard rulers as the sole guardians of these matters. On the contrary, scholar-historians asserted the indispensable role of their own class by recording the lives and words of eminent men of religion in countless biographical dictionaries.

In contrast to the religious scholars, chancery secretaries were men of the regime, who shared and indeed articulated its perspectives on rulership and society. The middle and upper ranks especially tended to live in a world apart from the people they governed. Their lives and fortunes were held at the whim of an autocrat, whose purposes and moods were of far greater importance to them than the life of the larger society outside the palace walls. In spite of their narrow focus, the secretaries often took a surprisingly critical view of the regimes they served. They typically judged them in accordance with the political traditions of pre-Islamic Iran, though these were often blended with specifically Islamic values. Iranian political tradition emphasized the crucial obligation of the ruler to uphold justice within his realms. In practice that meant respect for established social hierarchies, equitable taxation, enforcing right religion, and mastery of the army. The chancery secretaries might well be sincere Muslims, but for them the ruler and his senior officials were the sole actors who really counted. Hence their historical writing was strongly focused on the court and on the struggle for power. The best historians provide searching analyses of this struggle, to be sure, but with a few notable exceptions they portray it as the affair of a few score individuals, not of broad social groups and forces.

3. *The Problem of Origins, 622–825 CE*

Islamic historiography was born in obscurity. The earliest historical texts that we now have only acquired their present form in the late eighth century. For earlier stages of development we are reduced to inference or, more accurately, speculation. We can at least be sure that Islamic historical writing emerged within Arabophone Muslim circles, for the most primitive elements in this literature are couched in archaic Arabic and reflect almost exclusively the internal concerns of the Muslim community. Islamic historiography emerged out of the triumph and trauma of the seventh century: the sudden rise of a new religion among the peoples of Arabia, the vast conquests which they made under the banner of that religion, the profound changes which religion and empire wreaked on their traditional culture and socio-political organization, and the bitter civil strife which

rent a unified community of believers into hostile sects and factions. To make sense of these tremendous events, early Muslims had to devise some way of recalling, ordering, and interpreting them.

By the early eighth century they had come to believe that the Qur'an held the key to the underlying meaning of these events. The Qur'an (the collection of the revelations received by Muhammad between his call to prophethood ca. 610 and his death in 632) is far from a historical record of the earliest Muslim community. It alludes to a few crucial events which formed that community but does not narrate them. Rather, it places Islam within the long sequence of prophetic messages through which God has repeatedly called his creatures back to obedience. In the Qur'an, Muslims could see how a prophet had been sent to every people, bearing the promise of salvation for those who believed and the threat of damnation for those who refused. Most peoples rejected their prophets and suffered destruction for their obduracy. In contrast, a very few (most notably the Jews and Christians) did accept their prophets and prospered so long as they adhered to the messages these prophets had brought them. Over time, however, even these believing peoples tended to fall away, and God called them to repentance by periodic chastisement and humiliation. Through this repeated pattern the Qur'an laid down a paradigm which could explain the course of historical events. This paradigm contained three elements: the covenant offered by God to his people, the ever-present threat that this covenant might be betrayed by its recipients, and the desperate hope for redemption even after betrayal. In the circumstances of the seventh and eighth centuries, the Qur'anic paradigm of Covenant–Betrayal–Redemption was very compelling indeed.

Paradigms do not create literary genres. To recount the things that had befallen them, the early Muslims developed short prose narratives (Arabic, *akhbar*, singular, *khabar*) focused on discrete events. The earliest historical narratives about the seventh-century conquests and the internal struggles within the community of believers sprang up in a spontaneous, uncontrolled manner. They represented the contradictory memories and claims of innumerable tribal spokesmen and sectarian advocates, not a master narrative propounded by official spokesmen of the caliphal court. They were originally (and perhaps remained for some decades) oral accounts, and changed constantly as they were told and retold. Moreover, highly partisan narratives inevitably provoked counter-narratives. With a very few exceptions, it is no longer possible to recover the primitive form of any of these early narratives. That was no doubt the case even by the early 700s.

The earliest methodical effort to identify and collect the most reliable (i.e., religiously and politically acceptable) of these floating, infinitely plastic narratives took place around 700 under the aegis of the Umayyad caliphs. But the Umayyads could impose no

monopoly on the process of collection and redaction; indeed their version of Islam's first century was nearly obliterated when they were overthrown by the Abbasids in 749–50 and subjected to a systematic *damnatio memoriae*. Between 750 and 830 a group of scholars with informal ties to the Abbasid court did generate a stable corpus of historical narratives. This corpus represents what we now regard as the 'mainstream' version of early Islamic history. But competing versions produced by rivals and opponents of the Abbasids survive, albeit in fragmentary or masked form. However obscure the realities of early Islamic history may be, the historical debate of the late eighth and early ninth centuries is very well known.

The narratives collected, redacted, or (sometimes) created in the early Abbasid period were usually gathered in 'monographs' devoted to a single event (typically a battle) or category of information (e.g., lists of provincial governors). However, this period also witnessed the first major syntheses. Ibn Ishaq (d. 767) produced his famous biography of the Prophet Muhammad, the oldest complete work of Islamic historiography to reach us, and the following decades saw a host of works on the early conquests, the first caliphs, etc. Structurally, these syntheses are compilations of discrete narratives (*akhbar*), not continuous integrated narratives. On the other hand, the editors shaped their compilations in accordance with a conscious (though usually unstated) design and agenda. Mid-range syntheses of this kind dominated Islamic historiography down to the late ninth century, and they never ceased to be produced altogether.

The early Abbasid period also witnessed a plethora of efforts to develop a reliable absolute chronology, though the first essays in this direction seem to have been made under the later Umayyads. Chronographers naturally used the *hijri* calendar, traditionally instituted in 642. But of course events that had occurred before 642 had to be backdated, while the precise dates for a host of later events could only be guessed at. Nevertheless, Muslim historians quickly came to feel that precise dating was an essential element of their craft; the effort to pin things down and put them in the correct chronological order became a permanent characteristic of Islamic historical practice. By the 850s at latest, we begin to see chronicles of the kind long familiar in Byzantine historiography—viz., a year-by-year calendar of significant events (usually recounted in very summary fashion), beginning at a fixed point and coming down to the present. In the Byzantine tradition, that fixed starting point was typically the Creation, while early Muslim historians favored the Hijra (622 CE). The annalistic structure of the chronicle was a rigid but useful way of ordering events, and by the eleventh century it had become by far the most widely used format in Arabic-language historiography.

The early ninth century witnessed the appearance of one last genre of synthetic historiography, the bio-

graphical dictionary. The first major work of this class was the *Great Book of Generations* (*Kitab al-tabaqat al-kabir*) of Ibn Sa'd (d. 845). His goal was to compile concise biographies of all those persons who had transmitted Islamic religious knowledge from generation to generation, beginning with the Prophet himself. These biographies seldom provided a detailed account of the events of a person's life, let alone any effort to probe his personality. Rather, they aimed at evaluating his religious standing and his reliability as a scholar. Ibn Sa'd had clearly addressed a deeply felt need, for succeeding centuries (including the twentieth) saw a profusion of such works. In addition to countless dictionaries dedicated exclusively to religious scholars, there were dictionaries dealing with poets, men of letters, mystics, physicians, bureaucrats, and ultimately all those (including numerous women) who had earned a place within the cultural memory of Islam.

By the last quarter of the ninth century Muslim historians produced a series of grand syntheses, which claimed as their subject matter all of Islamic history, or even the history of the world since the Creation. Among many remarkable works, pride of place undoubtedly belongs to al-Tabari's massive *Chronicle of Prophets and Kings*, a work still in progress at his death in 923 (see Sect. 1). Apart from its methodological rigor, al-Tabari's compilation is distinguished by its vast scale, its meticulous attention to biblical and Iranian (though not Roman) history, and its inclusion of conflicting perspectives on the critical events of Islam's first 250 years. Al-Tabari was a somewhat controversial figure in his own time, but his authority among later generations was immense. Most of his successors were content simply to abridge and condense his account of early Islamic history, but from him they did learn to place Islam within God's plan of creation and salvation by linking it to the religions and empires which had preceded it.

4. The Mature Tradition, 950–1800

By the time al-Tabari died, the Covenant–Betrayal–Redemption paradigm that had shaped historical discourse throughout the eighth and ninth centuries seemed only remotely connected to the new crises confronting the Islamic world. Historians now had to grapple with the collapse of the unitary caliphate and the rise of a host of warlord-dominated regimes throughout the Islamic lands—in short, the political disintegration of the community of believers. Moreover, the religious conflicts which had arisen in Islam's first two centuries could no longer be healed; indeed, by the mid-tenth century what had been fluid if bitterly contested allegiances were transmuted into rigidly defined sectarian divisions. Political and religious fragmentation were unwelcome facts, but facts they incontrovertibly were, and they deeply shaped the historical writing of the following centuries.

First of all, the new environment provoked a change in subject matter. In striking contrast to al-Tabari and his predecessors, historians began to focus on recent and contemporary events rather than the community's early history. Likewise, they increasingly framed their work in local and regional rather than universal terms; their subject was more often the history of Isfahan or Damascus than of the Islamic world as a whole. This emphasis on recent and local events encouraged historians to adopt dynastic history as a framework for their narratives, for dynasties (and the warlords who created them) were the most visible political reality in their world. The turn to dynastic history was linked to new interpretive concerns. For historians observing the rise and fall of the warlord states, the central problem was not a ruler's role in the grand drama of salvation, but rather the reasons for the success or failure of his regime.

Within this new historiographic culture, universal history—i.e., history which claimed to include the whole Islamic world since the time of the Prophet, or even the Creation—continued to be written. But even the finest work of this genre, the *Complete History* (*al-Kamil fi al-ta'rikh*) of Ibn al-Athir (1160–1234), is strongly inflected by the distinctive concerns of its era. Within its strict annalistic framework is imbedded a quite visible sequence of dynastic histories. It covers all the lands of Islam with remarkable balance and detail, but it does so by examining each region *seriatim* as it surveys the events of each year. Finally, its primary focus is the mundane political process and the pragmatic lessons that can be inferred from that process.

Two other universal histories of great originality are equally products of their time and place. The *Comprehensive Chronicles* (*Jami' al-tawarikh*) composed under the supervision of the vizier Rashid al-Din (d. 1310) is a work which symbolizes the aspirations to universal rule of the Mongol rulers for whom it was written, for it deals with the history of every people about whom Rashid al-Din could obtain any information. Even the usually despised and neglected Franks are included within his vast tableau. It is an irreplaceable encyclopedia of the historical knowledge available to Iranian scholars in the early fourteenth century. Finally, the *Book of Examples* (*Kitab al-'Ibar*) of Ibn Khaldun (1332–1406) is distinguished not only by its famous introductory volume (*al-Muqaddimah*) but by its analytic survey of peoples and cultures and its author's grasp of the socioeconomic structures which shaped events.

Apart from these shifts in subject matter and emphasis, the mid-tenth century witnessed the bifurcation of historical writing into two linguistic branches. Arabic (the traditional language of Islamic scholarship) was now joined by Persian, the language spoken at the emerging princely courts of Iran and Muslim Central Asia. Arabic continued to be used throughout the Iranian cultural zone (which came to

include Central Asia, Anatolia, and much of South Asia) for the religious sciences, but Persian rapidly supplanted it in other arenas: secular and religious poetry, government administration, didactic prose, and history. The use of Persian was more than a linguistic phenomenon, for Persian historiography quickly developed distinctive characteristics of its own—e.g., a preference for lengthy continuous narratives, an explicit appeal to the political precepts and institutions of ancient Iran, a certain vagueness or carelessness about chronology. By the eleventh century Arabic and Persian historiography clearly embodied two distinct traditions, each developing in its own way. This picture was further complicated by the emergence of Turkish as a vehicle of literary expression in the late fifteenth century. From that point down to the mid-nineteenth century Turkish-language works were composed throughout the Middle East—no surprise, since the political and military elites from the Nile to northwest India were overwhelmingly of Turkish origin during that period. However, the main centers of Turkish literature were located in predominantly Turkish-speaking areas—i.e., Uzbek-dominated Central Asia and the core Ottoman lands in Anatolia and the eastern Balkans. In spite of some specific characteristics of its own and many works of outstanding quality, Turkish historiography tended to model itself upon the Persian tradition, and in some regions the two languages coexisted on almost an equal basis as vehicles of literary expression.

5. The Transition from Islamic to European Models, 1800–2000

By 1800 the Muslim peoples of Asia and North Africa faced severe military and economic challenges from Europe. However, most Muslim polities (with the important exceptions of the Mogul Empire in India and the sultanates of Java) still retained their political independence, and the deep-rooted societal and cultural structures of the Islamic lands seemed intact. By 1900, however, almost every Muslim state had been swallowed up or made a client regime, and Muslim societies were enmeshed in massive, uncontrollable transformation on every level.

Change on such a scale inevitably compelled the creation of a radically new historiography, though the process was slow and complex. Works of great merit continued to be produced within the existing traditions throughout the first half of the nineteenth century. Indeed, excellent historical scholarship of a decidedly traditional cast was being written well into the twentieth century. Nineteenth-century historians were fully cognizant of the new world they inhabited, but for many decades traditional historiographic forms and approaches seemed adequate to the task. By the 1870s that was no longer the case. First, existing models of historical writing simply did not possess the analytic

tools that were needed to address the profound cultural and ideological problems created by the European challenge. For this purpose a new historiography was essential. No less important, by this time many Muslim intellectuals had studied in Europe or with European teachers, and they felt the powerful allure of Enlightenment and post-Enlightenment thought. To them it was self-evident that European military and economic power was grounded in the power of European ideas. Just as Muslim armies needed to adopt European weapons and tactics, so Muslim intellectuals had to appropriate European concepts and methods of inquiry.

European ideas and literary models were easier to perceive than to internalize, and the later nineteenth century witnessed a difficult struggle by Muslim intellectuals and scholars to develop new ways of writing history. Most late nineteenth-century historiography is thus a hybrid product. The point is illustrated by two major works, both marked by broad scope, critical use of documents, and awareness of European scholarship. Ahmet Cevdet Pasa's history of the Ottoman Empire (12 Vols., 1855–84) adheres to the traditional chronicle form, while 'Ali Mubarak's historical geography of Egypt (20 Vols., 1886–88) is modeled on a classic work by the fifteenth-century Cairene scholar al-Maqrizi (d. 1442). A model for historiography in a modern (i.e., European) style was provided by the *History of Islamic Civilization* (*Ta'rikh al-tamaddun al-islami*, 5 Vols., 1902–6) of Jurji Zaydan, a Syrian Christian novelist and man of letters. For both political and cultural reasons, however, Zaydan had no immediate imitators, and only after World War I did Middle Eastern historiography begin to internalize contemporary European modes of research and presentation. This shift was first visible in works dealing with modern history, but by the end of the 1930s the same trend had taken hold for studies of early and medieval Islamic times. The 'Europeanization' of historiography affected not only research methods and literary structure, but also the kinds of questions that Muslim scholars asked about their past. Even where (for a wide range of scholarly, political, and religious reasons) they rejected the specific analyses and interpretations of Middle Eastern and Islamic history asserted by European historians, they could not ignore the problems to which these scholars had called attention.

This transformation coincided with the professionalization of history after World War I—a sharp departure from traditional practice in the Islamic world. This process began with the establishment of European-style universities in Cairo and Istanbul just before the war, along with the creation of officially sponsored societies for historical research. Simultaneously, history was enlisted in the struggles for national independence and cultural autonomy of the interwar and postwar periods. Finally, by World War I a few Muslim scholars were taking advanced degrees from

the leading universities in Europe, and their numbers increased greatly in the following decades. The prestige and authority that these foreign-trained scholars enjoyed when they returned home in effect imposed a new model of historical thought and writing on their students.

Cross-cultural influence in historiography has not flowed in a single direction, however. Even in the 1920s and 1930s, a few Muslim scholars published their work in Western languages, and since the 1950s this trend has grown enormously. A growing number of scholars from Islamic Asia and Africa have not only taken their advanced degrees in Europe and North America, but have also been appointed to many of the most prestigious teaching posts there. This situation has begun to create a real dialogue among a plurality of theoretical and cultural perspectives. The history of the Islamic world is increasingly in the hands of its own scholars—the natural state of things, perhaps, but one which was hardly the case in the nineteenth and twentieth centuries.

See also: Islam: Asia; Islam: Middle East; Islam: Sub-Saharan Africa; Near Middle East/North African Studies: Culture; Near Middle East/North African Studies: Religion; Near Middle East/North African Studies: Society and History

Bibliography

- Crabbs J A 1984 *The Writing of History in Nineteenth-century Egypt: A Study in National Transformation*. Wayne State University Press, Detroit, MI
- Delanoue G 1982 *Moralistes et politiques musulmans dans l'Egypte du XIXe siècle (1798–1882)*. Institut Français d'Archéologie Orientale, Cairo, Egypt
- Donner F M 1998 *Narratives of Islamic Origins: The Beginnings of Islamic Historical Writing*. Darwin Press, Princeton, NJ
- Ende W 1977 *Arabische Nation und islamische Geschichte: die Umayyaden im Urteil arabischer Autoren des 20. Jahrhundert*. Orient-Institut der Deutschen Morgenländischen Gesellschaft, Beirut, Lebanon
- Encyclopaedia of Islam*, New Edition 1960 – in progress Entries: *ta'rikh* and individual historians. E J Brill, Leiden, Netherlands
- Fleischer C 1986 *Bureaucrat and Intellectual in the Ottoman Empire: A Biographical Study of the Historian Mustafa Ali, 1541–1600*. Princeton University Press, Princeton, NJ
- Hourani A H 1962 *Arabic Thought in the Liberal Age, 1798–1939*. Oxford University Press, London
- Hoyland R G 1997 *Seeing Islam as Others Saw It: A Survey and Evaluation of Christian, Jewish and Zoroastrian Writings on Early Islam*. Darwin Press, Princeton, NJ
- Humphreys R S 1991 *Islamic History: A Framework For Inquiry*, revised edn. Princeton University Press, Princeton, NJ
- Khalidi T 1994 *Arabic Historical Thought in the Classical Period*. Cambridge University Press, Cambridge, UK
- Lewis B, Holt P M 1962 *Historians of the Middle East*. Oxford University Press, London
- Meisami J S 1999 *Persian Historiography to the End of the Twelfth Century*. Edinburgh University Press, Edinburgh, UK

- Mustafa, Shakir 1983–93 *Al-Ta'rikh al-'arabi wa'l-mu'arrikhun*, 3rd edn., 4 Vols. [Arabic Historiography and Historians] Dar al-'Ilm li'l-Malayin, Beirut, Lebanon
- Noth A 1994 *The Early Arabic Historical Tradition: a Source-Critical Study*, 2nd edn. Darwin Press, Princeton, NJ
- Radtke B 1992 *Weltgeschichte und Weltbeschreibung im mittelalterlichen Islam*. [World History and the Description of the World in Medieval Islam] Orient-Institut der Deutschen Morgenländischen Gesellschaft, Beirut, Lebanon
- Rosenthal F 1968 *A History of Muslim Historiography*, 2nd edn. E J Brill, Leiden, Netherlands
- Wellhausen J 1899 *Prolegomena zur ältesten Geschichte des Islams*. [Prolegomena to the History of Early Islam] In: Wellhausen J, *Skizzen und Vorarbeiten* Georg Reimer, Berlin Vol. 6

R. S. Humphreys

Historiography and Historical Thought: Jewish Tradition

1. Biblical Traditions of History Writing

Together with the historical traditions of ancient Greece the Hebrew Bible contains the oldest corpus of historical writings. While not all of the various books which entered the Jewish biblical canon reveal a strong sense of history, quite a few—such as Samuel, Kings, and Chronicles—are primarily historical accounts. In contrast to the Greek historians, the authors and editors of the Bible were interested not in recording history but in giving a theological sense to history. The numerous biblical passages which command the Jews to remember (*zakhor*) do not imply a special interest in history itself but express a definition of the Jews as a historical people held together first and foremost by its will to observe its historically revealed laws and to remember its common history. God is described in terms of His historical deeds. He is the 'God of the fathers' and 'Your God who brought you out of the Land of Egypt.'

With the canonization of the Hebrew Bible in the second century CE (Christian Era) Jewish historical writing undertakes a clear shift. The first century CE produces the last major Jewish historical writings for about 15 centuries. In a historical work, of which only two out of five parts survived, Philo of Alexandria describes the persecution of Egyptian Jews by the Roman governor Flaccus and the delegation of Alexandrian Jews (including himself) to the emperor Caligula. The most important historian in this period is Flavius Josephus, whose account of the Jewish War against Rome in 66–73 CE was written a few years later in Roman exile and—together with his *Antiquities*—stands in the tradition of Roman rather than Jewish historiography.

2. Historical Thought in Rabbinic Literature and Medieval Sources

There is hardly a more telling symptom for the decline of Jewish historical writing than the fact that Josephus' histories fell into oblivion among the Jews during the Talmudic period. The same holds true for many historical accounts which did not enter the Jewish Biblical canon, such as the Books of the Maccabees, which describe the rise of the Hasmoneans against Hellenism. Preserved by non-Jews, they were re-discovered by Jews only in the sixteenth century. During the Middle Ages the history of the Second Temple period—beyond the Biblical accounts themselves—was known among Jews in rather distorted versions of the original accounts, such as the *Book of Antiochus* and the tenth-century *Book of Yossipon* (believed to be Josephus' original account), both produced in southern Italy.

The Talmud, which was to become the single most important source for Rabbinic Judaism over the centuries, is unhistorical in its nature. Although this vast corpus contains quite a few passages of historical value, its editors reveal no special interest in the chronology of events. Thus, rabbis of different centuries are made to engage in conversations, and biblical events are often presented out of their chronological order.

Medieval Jewish historical thought is framed by the memory of the biblical past on the one hand and by the longing for a messianic future on the other hand. The time span between a golden past and an ideal future is usually interpreted with clear reference to the biblical course of events. Biblical place names like Ashkenaz and Sefarad designate European countries, Christianity is Edom, and Islam is Yishmael. Vicious rulers are called Haman, like the archetypal biblical villain from the Book of Esther, and saviors are termed Mordechai after the hero of the same story. Messianic expectations led to concrete calculations about the End of the Days, and contemporary tragedies were often seen as the birthpangs of the Messianic Age.

Altogether, traditional Jewish historical thought as expressed in Rabbinic literature and medieval sources reveals at least a partly cyclical understanding of time. Every year when celebrating Passover observant Jews remember the exodus from Egypt, and during three annual days of fasting they mourn the destruction of Jerusalem. Each rescue from persecution is regarded a new Purim (in commemoration of the Book of Esther). Every Sabbath morning a weekly portion of the Pentateuch is read in the synagogue, thus reviving sequentially each year the historical accounts of the Five Books of Moses.

The most prominent genre of historical writing during the Middle Ages was the chronologies, which often tried to establish a reliable 'chain of tradition' (*shalshet hakabbala*) from Moses through the leading Rabbis of the Mishna and Talmud and up to the

authors' own time and place. They constituted a refutation of the anti-Talmudic challenge made by the Karaite sect and at the same time underlined the importance of the local Jewish community in which they were produced. Examples of this genre are the *Letter of R. Sherira b. Hanina* from the academy at Pumbedita in Babylonia, who in the late tenth century responded to scholars from Kairouan, and the twelfth-century *Sefer Hakabbala* (Book of Tradition) of the Spanish scholar Abraham ibn Daud. A stronger emphasis on contemporary events can be seen in the detailed accounts of the Crusade massacres concerning the cruelties which befell the Jewish communities of the Rhineland in the eleventh and twelfth centuries.

Among the most popular sources reflecting medieval and early modern Jewish historical thought in Central Europe were Memor Books, which contain collections of prayers together with a necrology of distinguished persons, and recorded local martyrs and persecutions. The oldest known Memor Book is from Mainz and dates back to 1296. A few hundred Memor Books are extant, mainly from seventeenth-century rural communities in southern Germany.

3. Sixteenth-century Jewish Historiography as a Response to Catastrophe

The sixteenth century saw a sudden upsurge of Jewish historical writing with the publication of 10 important historical works, more than all of those known to us from the Middle Ages. Quantitatively as well as qualitatively the development was significant, with a new interest in history itself, as Yosef Hayim Yerushalmi demonstrated in his seminal study *Zakhor* (1989). Writers like Yosef ha-Kohen, Abraham Zacuto, and Samuel Usque wrote in the wake of the expulsion of Jews from Spain and Portugal (1492 and 1497), which affected not only their own lives but also completed the wave of expulsions of Jews from Western Europe. The new search for explanation is apparent in their writings, perhaps most pointedly in Solomon ibn Verga's *Shevet Yehuda* (The Staff of Judah, 1554), but also in the historical oeuvre of the only Ashkenazic author of this group, David Gans from Prague. His *Tsemah David* (Offspring of David, 1592) contains the most precise account of Jewish and world history undertaken by a Jewish author until that time.

The Italian Jew, Azaria de' Rossi, who completed his *Me'or Enayim* (A Light to the Eyes, 1573–5) at about the same time, has the merit of recovering ancient sources that had been lost to Jewish tradition because they did not enter the biblical canon. Among them are works by Philo of Alexandria and the apocryphal *Letter of Aristeas*. For the first time the historical interest of Jewish writers did not concentrate on the biblical past or speculations about the messianic future but dealt with the concrete history of the Jews

during recent centuries and in their own time. These authors also showed a genuine interest in the histories of other nations. The impact of their work on later Jewish historiography was, however, minimal. It was only as a result of the encounter with academic scholarship in the early nineteenth century that modern Jewish history writing developed.

4. The Beginnings of Modern Jewish Scholarship and Critical Historiography in the Nineteenth Century

The beginnings of modern Jewish scholarship can be dated to the small circle of Jewish university students from Berlin who founded the *Verein für Cultur und Wissenschaft der Juden* in 1819. Leopold Zunz, the leading spirit of this group and the dominant figure in the field which would become known as *Wissenschaft des Judentums*, published in 1818 a systematic agenda for modern Jewish research under the title '*Etwas ueber die rabbinische Literatur*' (On Rabbinic Literature). The study of history, which had been marginal in the works of eighteenth-century Jewish enlighteners (*maskilim*), now became a central pillar of Jewish scholarly enterprise. From the beginning, a certain apologetic character was inherent in many attempts at reclaiming the Jewish past. As Immanuel Wolf, a founding member of the *Verein* expressed it: 'Scientific knowledge of Judaism must decide on the merits and demerits of the Jews, their fitness or unfitness to be given the same status and respect as other citizens.'

In a more direct way, the scholarly studies of Leopold Zunz served as instruments to achieve emancipation for German Jews. Like Wolf, Zunz believed that 'the equality of the Jews in customs and life will follow from the equality of *Wissenschaft des Judentums*'. In his *Gottesdienstliche Vorträge der Juden, historisch entwickelt* (1832), he documented the long tradition of Jewish sermonic literature, thus rebutting the Prussian government's dismissal of German synagogue sermons as innovations in Jewish liturgy. Five years later, in his study on Jewish surnames, *Die Namen der Juden*, Zunz proved the ancient origin of many 'non-Jewish' names among Jews and thus reverted to a decree which forbade the use of Christian first names among Jews.

It would, however, not do justice to the enormous enterprise of nineteenth-century *Wissenschaft des Judentums* to view it predominantly under the aspect of apologetics, as did later critics, especially Zionists, most prominently Gershom Scholem. Another important motif for the pioneers of Jewish historiography was to retrieve the lost treasures of the past for a rapidly assimilating Jewish community as a means of strengthening their particular identity as modern Jews. As Yosef Hayim Yerushalmi writes, 'History becomes what it had never been before—the faith of fallen Jews' (1989, p. 86). Finally, the purely scholarly

interest in the Jewish past at a time in which history writing and reading became one of the most popular intellectual endeavors should not be disregarded.

There was no bridge between those first scholarly accounts of Jewish history and earlier Jewish writing. It was through the non-Jewish world that historiography entered Jewish writing. Consequently, it was not a Jew, but a Protestant French author in Dutch exile, Jacques Basnage, who wrote the first multi-volume postbiblical Jewish history. He characterized his *Histoire du peuple Juif depuis Jésus Christ jusqu'à présent*... (1706–11, rev. edn., 1716–21) in his subtitle as a continuation of Josephus, and there was indeed some truth to this statement. To be sure, his account would in no way measure up to modern scholarly standards, and was still full of missionary intentions, but it was nevertheless the first serious attempt to write a comprehensive postbiblical Jewish history.

5. Political Agendas and the Writing of Jewish History

The first Jewish writer to embark on such an endeavor was Isaak Markus Jost, a member of the *Verein für Cultur und Wissenschaft der Juden* and one of the pioneers of *Wissenschaft des Judentums*. The modern secular education he received as a student at Berlin and Goettingen is visible in his attempt to present a sober and objective view of Jewish history in his nine-volume *Geschichte der Israeliten seit der Zeit der Maccabäer bis auf unsere Tage* (1820–28). Jost's work is clearly centered on the German-Jewish experience and draws a picture of Judaism as a religion which has lost all its original national characteristics. Depicting modern Jews as a purely religious community was part of the larger German-Jewish struggle for Emancipation. The title of his later work, the three-volume *Geschichte des Judentums und seiner Sekten* reveals the urge to present Judaism as a religion analogous to Christianity.

When Heinrich Graetz wrote his 11-volume *Geschichte der Juden von den ältesten Zeiten bis auf die Gegenwart*—the most popular such history ever written—one generation later (1853–76), he deviated from the early works of *Wissenschaft des Judentums* representatives in his stronger emphasis on the existence of a Jewish people and by a more passionate presentation. Although he was later claimed as a proto-Zionist writer, it is important to note that his definition of the Jewish nation is that of a community held together by an essentially religious idea. Like Jost, for whom modern Jewish history begins with the enlightened Prussian monarch Frederick the Great, Graetz located the beginnings of Jewish modernity in Germany, with the appearance of the philosopher Moses Mendelssohn and the Berlin Jewish Enlightenment.

It was not until the late nineteenth century that a clear break from the earlier Germanocentric view of Jewish history and from the strong emphasis upon a 'history of suffering and scholarship' (Graetz) appeared in the work of the Russian-Jewish historian Simon Dubnow. His 10-volume *World History of the Jewish People* was written in Russian but published first in German translation (1925–29). Motivated partly by his own political agenda—he was the founder of the autonomist Jewish movement which represented Jewish diaspora nationalism in eastern Europe—his version of Jewish history centered on institutions of Jewish life in the diaspora, most notably the *kehilla*, the semi-autonomous Jewish community.

Dubnow was also among the founders of the YIVO Institute for Jewish Research—an institution established in Berlin and Vilnius in 1925 to systematically research the Jewish past and present in Eastern Europe and other Ashkenazic communities. In contrast to the traditional German dominance of nineteenth-century *Wissenschaft des Judentums* and to Hebrew-language Zionist scholarship, the YIVO deliberately presented its research in Yiddish, the language spoken by the Jewish masses of Eastern Europe.

From the end of the nineteenth century, Jewish historiography was increasingly influenced by new subdisciplines of *Wissenschaft des Judentums*, such as ethnography, sociology, and demography. The *Gesellschaft fuer Juedische Volkskunde*, a society of Jewish folkloristics founded in 1898 and led by the Hamburg rabbi Max Grunwald, played a pioneering role in the scholarly research of Jewish folk traditions around the globe. Its journal, the *Mitteilungen fuer Juedische Volkskunde*, published the most important findings in the field for almost three decades (1898–1929). In Russia, an expedition to the traditional communities of the Ukraine led by the playwright and folklorist An-Ski from 1912 to 1915 was the culmination of a long search for the remnants of rural Jewish life.

Jewish demographers created their own institutional framework when they established an Office for Statistics among the Jews in Berlin, which from 1904 published its own journal and was closely related to the burgeoning interest in Jewish sociology best expressed in the pioneering works of Arthur Ruppin. Most of these endeavors were clearly related to the 'Jewish Renaissance' propounded by the emerging Zionist movement. Zionist scholars reacted against what they alleged to be the orientation of nineteenth-century Jewish scholars exclusively to the Jewish past and broadened their interest to include contemporary issues within their research.

After the establishment of the Hebrew University in 1925, scholars with a Zionist outlook formed a vaguely connected group of Jewish historians, sometimes referred to as the 'Jerusalem School,' emphasizing the centrality of Palestine in the course of Jewish history. Their most outspoken representative was the later

Israeli minister of education Benzion Dinur (Dünaburg), who in 1935 helped to establish the most important Hebrew-language historical journal, *Zion* (an earlier series of the journal had been aborted), together with the first professor of Jewish history at the Hebrew University in Jerusalem, German-born Yitshak (Fritz) Baer.

From the very early works of Zionist historiography, usually written by Zionist activists who were not professional historians (such as Nahum Sokolov and Adolf Böhm), a teleological view of history leading to the return to the Jewish homeland is clearly visible. At the same time, the Jewish past in the Land of Israel is placed in the center of interest at the expense of diaspora history. In contrast to most historians, Benzion Dinur thus commences the period of exile not with the destruction of the Second Temple in the first century CE but much later, with the Muslim conquest of Jerusalem in the seventh century CE. In a similar vein, he concludes the period of diaspora domination as early as 1700, when a group of pious Jews from Eastern Europe made their way to Palestine. In his *History of Zionism, 1600–1918* (1919) Nahum Sokolov anachronistically dates the Zionist movement back to the seventeenth century.

While Dubnow's conception of Jewish history represented a corporate view of Jewish diaspora life and the Jerusalem School emphasized the centrality of Palestine, the last historian who attempted a single-handed multi-volume universal Jewish history, Salo Wittmayer Baron, stood for a more positive approach to individuals' diaspora experience. His 16-volume *A Social and Religious History of the Jews* (1952–76) reaches only the year 1650. His more affirmative view of the diaspora was certainly influenced by his own experience of spending most of his life in the United States, where he was appointed in 1930 as the first professor of Jewish history at a Western university, Columbia University in New York.

Jewish historiography has thus to a large extent been determined by the political context, the geographical situation, and the ideological and religious outlook of the respective historians. Since all these factors differed considerably in the Jewish world, the outcome was a highly diverse array of Jewish historical writing – even though, at least up until the second half of the twentieth century, most of those who undertook to write Jewish history were themselves Jews. The destruction of European Jewry, the establishment of the State of Israel, and the flourishing of Jewish studies in the United States led to an entirely new situation for Jewish historiography after World War II.

6. Twentieth-century Trends

With the exception of Baron no single historian embarked on a monumental Jewish history project in the tradition of Jost, Graetz, and Dubnow in the course of the twentieth century. There were a few

attempts to publish Jewish world histories as collective enterprises, as early as the ambitious plan developed by Ismar Elbogen of the Berlin *Hochschule für die Wissenschaft des Judentums* in the late 1920s; these, however, fell victim to the economic and political crisis of the early 1930s. More successful, but for financial reasons also never completed, was a multi-volume project initiated by scholars of the Hebrew University.

After World War II Jewish history became more and more a field for specialists who would concentrate on limited areas and periods. This tendency toward specialization was also the result of an increasing professionalization and acceptance of the subject as a distinct field to be taught at university level, with new centers emerging in Israel and the United States. Until the 1960s, the teaching of Jewish history in the US had been restricted to a handful of universities and a few institutions with a strong emphasis on Jewish education, such as rabbinical seminaries and Jewish teachers' colleges or the Jewish-sponsored Brandeis University. With the establishment of new Israeli universities the singular role of the Hebrew University in Israel diminished, while a true explosion in the field of Jewish studies enabled a similar diversification at American universities beginning in the 1970s. European institutes of higher learning followed to a more modest degree in the 1980s and 1990s. By 2000, Jewish history had become a highly specialized and diversified field represented at almost every major institution of higher learning in the United States and many European countries.

This development led to a more complex picture of Jewish historiography. The personalities and political programs of individual historians and the formations of more or less clearly defined schools no longer dominate the professional scene. Zionist historians who had fought an ideological battle over Jewish history, mostly against their non-Zionist Jewish colleagues, up to the 1930s were essentially left without opponents as a result of the Holocaust and the destruction of alternative Jewish options, such as diaspora nationalism and Bundist socialism. In the State of Israel, apologetical historiography (now mainly directed against Arab and British versions of the history of Palestine) did not entirely disappear but was more and more balanced by the emergence of a younger generation of historians, by whom the existence of the Jewish state was already taken for granted. This resulted in a gradual break with long-accepted notions rather than an intellectual revolution as more and more historians and sociologists of a new postwar generation attacked the older schools of Zionist historiography in the 1980s and 1990s.

While mainly questioning issues of Israeli history or prehistory, such as the role of the Zionist leadership in Palestine with respect to attempts at the rescue of European Jewry in World War II, and the accepted notions surrounding the first Arab–Israeli War in 1948–49, the views of the so-called 'New Historians'

also challenge more general Zionist notions of Jewish history. In a highly politicized context the two sides of the 'Israeli Historians' Debate' stand not only for different views on Jewish or Israeli history but mark also different paths for Israel's political future. In contrast to Israeli historiography, which has been dominated by a classical left-liberal Zionist intellectual establishment, the 'new' historians and sociologists often represent a post-Zionist stance and as such constitute a major challenge to the master narrative of Jewish and Zionist history.

In the last decade of the twentieth century, Jewish historiography moved from its early Germanocentric and later Eurocentric emphases to a stronger integration of widely neglected communities. In particular, the histories of Jews in the Muslim world have slowly been integrated into the general picture. This is partly the result of the emergence, mainly in Israel, of historians of oriental background, but also of the establishment of a few academic positions concentrating on Sephardic Jewry at American universities. It was in the United States, where the history of Jewish women emerged as a subfield of gender studies, with significant publications in the 1990s about Jewish women in Imperial Germany and in the United States, and a historical encyclopedia of Jewish women. New trends among general historians, including post-modern ideologies, have also left their imprint on Jewish historiography.

Europe, too, witnessed a revival of Jewish historical scholarship in the 1990s. The larger Jewish communities of Western Europe, France and Britain, emerged as new centers of Jewish scholarship, although most of the historical studies there deal with their respective national Jewish histories. Central and East European institutions relating to the Jewish past were destroyed by the Nazi onslaught, together with those who participated in them. Ironically, the Nazis entered into much more systematic Jewish historical research than non-Jewish scholars had done in the Imperial and Weimar periods; they established research institutes for Jewish history in Munich, Frankfurt, and Berlin. Studies undertaken there, however, were preoccupied not with Jewish history *per se*, but with the so-called 'Jewish Question' and sought to justify Nazi discrimination and persecution against the Jews from a pseudo-scholarly perspective.

It is interesting to note, though, that some of the research begun there—especially on the Court Jews and on Sephardic Jews in Hamburg—was published with only slight alterations in the 1950s. More modern Jewish historical research in Germany developed as a result of the establishment of the first Jewish Studies institutes and of the emergence, starting in the 1960s, of a new interest in German-Jewish history among modern German historians. While Jewish history is mainly written by Jewish historians and for a Jewish public both in prewar Europe and in today's Israel and America, the postwar European situation is different.

With only small Jewish communities, research is mainly undertaken by non-Jewish scholars, and mainly read by non-Jews.

7. Major Issues in Modern Jewish Historiography

Perhaps even more than scholars of other particular histories Jewish historians have been divided by a set of fundamental questions. First of all, there is the question of whether Jewish history exists at all or whether the history of Jews merely constitutes a chapter within the histories of the nations among whom they lived. If Jewish history exists in a broader sense, the question arises whether it is the history of a people or a purely religious history—or whether at some point the national history turned into a religious history. Finally, an additional problem of Jewish history is its theological interpretation by some modern Jewish and non-Jewish thinkers, which makes Jewish history appear as though it escapes the regular laws of history. This last argument is strengthened by the conspicuous role of antisemitism throughout the centuries and the claim of the uniqueness of the Holocaust which brought about the near destruction of European Jewry.

The doubt whether Jewish history exists *per se* seems rather superfluous in an age of expanding programs in Jewish history and within the framework of an article discussing the rich tradition of Jewish history writing. Nevertheless, one can still find this opinion among some modern historians. It is based partly on the Christian theological notion that active Jewish history came to an end with the early stages of Christianity, and partly on the secular claim that Jews lack a clearly defined political history based on a common territory. In the nineteenth century such a belief was inherent in the theological characterization of ancient Judaism as 'late Judaism,' while on the other side of the spectrum of historical outlook Marxist historians denied a separate course of Jewish history on the basis that Jews not only lacked a territory of their own but also a division into classes of society in the modern period. Nowadays, the questioning of the existence of a Jewish history throughout the ages is less explicit. More indirectly, however, such a view can be detected in the almost exclusive interest of many historians in the histories of the Jews in their respective societies. Thus, German Jewry or French Jewry are often seen as isolated entities, as part of German or French history, but without any connection to the developments of other Jewish communities. On the other hand, there has been a contrary tendency among some Zionist historians to neglect the respective contexts of Jewish histories and view them only through a national Jewish perspective.

If we accept the notion that Jewish history exists as such and is stretched over numerous centuries and

over five different continents, the question remains whether it is primarily a national or a religious history. The pioneers of Jewish historiography in nineteenth-century Germany often preferred to depict Jewish history in terms of the development of a religious idea (the purest form of monotheism) and its bearers. This concept was in line with the demands of Emancipation that Jews integrate into European societies as citizens who only differed in their religious creed. Thus, the notion of a Jewish people was characterized as an ancient idea long left behind. With diaspora nationalists like Simon Dubnow and the emerging Zionist school the national element was restored to Jewish historiography. Both schools emphasized that Jews never ceased to constitute a nation even if they lacked their own territory. Dubnow went so far to stress that a nation which had overcome the need for its own territory constituted the highest and most sophisticated form of nationalism. Zionists, on the other hand, emphasized the unbroken longing for return to the Land of Israel and the use of the Hebrew language over the centuries.

A final peculiarity of Jewish history is the theological weight it enjoyed especially in a Christian context. In medieval and early modern Europe Jews often constituted the only non-Christian minority. The traditional position Jews enjoyed in Christian Europe was a mixture of protection and discrimination. Jews served as the classical 'other,' being protected because they bore testimony to Jesus Christ while at the same time discriminated against as Christ-killers and as unbelievers. On the basis of this peculiar role, many Christian authors of postbiblical Jewish history regarded this particular history as outside the course of 'regular' history. They claimed accordingly that it should be judged by theological rather than historical criteria. On the Jewish side, the most prominent proponent of a Jewish '*Sonderweg*' theory, albeit with a different background, was the philosopher Franz Rosenzweig, a student of Friedrich Meinecke, who stressed in his main work *The Star of Redemption* (1921) that general rules of history do not apply to Jewish history. More recently, the discussion of the uniqueness of the Holocaust has brought the question of a peculiar course of Jewish history to a different level. In this view, it is not Jewish history that is incomparable to other histories, but rather the persecution of the Jews which culminated in the attempt of their total annihilation.

In his programmatic and provocative essay, 'Ghetto and emancipation' (1928), the historian Salo Baron fiercely objected to what he called a 'lachrymose version of Jewish history.' At the beginning of the twenty-first century, and half a century after the greatest tragedy of the Jewish people, numerous voices once again express their reservations against a one-sided picture of Jewish history reduced to a history of persecution. While they stress the importance of research on antisemitism and the Holocaust, they also

warn not to read Jewish history backwards starting from its most tragic chapter.

See also: Historiography and Historical Thought: Christian Tradition; Historiography and Historical Thought: Current Trends; Historiography and Historical Thought: Islamic Tradition; Jewish Diaspora; Jewish Law; Judaism; Judaism and Gender; Religion, History of

Bibliography

- Funkenstein A 1993 *Perceptions of Jewish History*. Berkeley University Press of California, Berkeley, CA
Liberles R 1995 *Salo Wittmayer Baron: Architect of Jewish History*. New York University Press, New York
Meyer M A 1974 *Ideas of Jewish History*. Behrman House, New York
Myers D 1995 *Reinventing the Jewish Past: European Jewish Intellectuals and the Zionist Return to History*. Oxford University Press, New York
Myers D N, Ruderman D B (eds.) 1998 *The Jewish Past Revisited, Reflections on Modern Jewish Historians*. Yale University Press, New Haven, CT
Schorsch I 1994 *From Text to Context: The Turn to History in Modern Judaism*. University Press of New England, Hanover, NH
Yerushalmi Y H 1989 *Zakhor: Jewish History and Jewish Memory*, 2nd edn. Schocken, New York

M. Brenner

Historiography and Historical Thought: Modern History (Since the Eighteenth Century)

1. Reorientation of Historical Thought in the Eighteenth Century: The Emergence of a Secular Outlook

There is good reason for beginning this article on Western historical thought and historical writing in the modern age with the eighteenth century. The Western tradition is much older, going back at least to the Hellenic Age. There is much continuity between the older traditions, particularly those of the humanistic historiography of the Renaissance and historical writing in the eighteenth century. Yet there are also profound changes in the conceptions of history and of historical knowledge and in the institutional context in which history is written that mark a sharp break with previous periods.

Henri Berr and Lucien Febvre in the article 'History' in the first edition of the *Encyclopedia of the Social*

Sciences in 1932 identified two characteristics of the modern conception of history as basic to the modern historical outlook and practice: the commitment to establishing the 'truth' and to placing historical accounts within 'universal history.' A third is the transformation of historical studies into a professional discipline. The older ancient and humanistic historiography, they argued, saw history as a form of rhetoric in the service of politics and morality and as an art form. 'History as it is conceived today,' they noted, 'may blossom into art, may be crowned with philosophy; but it is primarily and necessarily the solid establishment of facts' (Berr, p. 358), however difficult this may be.

This formulation presupposes a sharp distinction between history as literature and as a scientific discipline, even if its approach to science takes into account the elements of contingency and particularity that distinguish history from the generalizing natural sciences; but it assumes that historical accounts based on a solid factual basis narrate a coherent development that reflects historical reality.

There is a good deal to be said for this description of the transformation which historical studies underwent in the eighteenth and particularly in the nineteenth century; but especially from the perspective of the late 1990s it seems one-sided and incomplete.

On two levels, historical thought and historiographical practice in the eighteenth century moved toward greater historical realism. On the one hand, an increasingly secular outlook insisted on the exclusion of the supernatural and held that historical explanation must not violate a common understanding of natural reality. This secular approach was relatively new in the eighteenth century. On the other hand, it coexisted often with a residual religiosity, as it still did even in the nineteenth century with a conception of a hidden divine plan, as in Leopold von Ranke's trust in the 'hand of God' (1973, p. 138), or in a less hidden form in G. W. F. Hegel's dialectic. Thus Enlightenment historians such as J. C. Gatterer and A. L. Schloetzer did not challenge the biblical chronology or the biblical account, although for the post biblical period, or for the non-Hebraic ancient peoples, they insisted on an approach in accord with natural reality.

Apart from the great works of synthesis in the eighteenth century, David Hume, Montesquieu, Voltaire, Edward Gibbon, G. T. Raynal, and Adam Ferguson, there exists a tradition of erudition that introduced standards of source criticism, which went back to the humanist scholarship of the Renaissance and Reformation and to the erudition of seventeenth-century historical scholarship. The belief that great philosophic works of synthesis are not enough and that scholarly history cannot be written until the authenticity and veracity of the sources on which it rests have been established expressed itself in the great methodical collection of sources begun in the seventeenth century with the Jesuit *Acta Sanctorum*,

published after 1643, to provide a firm basis for ecclesiastical history, and with Ludovico Muratori's efforts in the *Scriptores rerum italicarum* (1723–51) to lay the foundation for an Italian national history. The German scholars at the universities, first J. L. Mosheim and J. S. Semler at Halle and then J. A. Ernesti and J. O. Michaelis at Göttingen, established a tradition of historical exegesis that laid the foundations for historical hermeneutics and the new discipline of historical philology which emerged in the latter part of the eighteenth century and became the key elements of the critical method of B. G. Niebuhr and Ranke in the nineteenth century.

A key concept of eighteenth-century thought was the linear character of history. History was no longer understood as a cumulation of events but as a continuous story possessing inner coherence. Reinhart Koselleck has suggested that between 1750 and 1850 the concept 'history' (*Geschichte*) replaced that of 'histories' (*Geschichten*). Each of the great historical works of the eighteenth century followed a continuous, coherent story whether in Gibbon's *Decline and Fall of the Roman Empire*, Hume's *History of England*, or Voltaire's *Essai sur les mœurs*.

Yet the conception of world history that arose with the Enlightenment constituted not only an extension of the historical outlook but at the same time also a narrowing. And this conception dominated almost all of Western historical thought and historical writing until well into the second half of the twentieth century. The history of the world was identified with the history of civilization and the latter rather narrowly with that of the West. Two variations emerged in the conception of what constituted world history, one identified with Condorcet, centering around an idea of progress on a world scale, the other identified with Herder, which stressed the diversity of peoples and cultures. Yet on closer examination, these two conceptions shared many characteristics, specifically in their stress on development.

The Enlightenment has too many aspects to permit one to reduce it to a simple formula. Two ideas were central to it, but neither was free of contradictions. The first assumed the essential equality of all human beings in all cultures; the second maintained that only human beings who had attained a high degree of culture and education deserved to be free and equal citizens. The latter view distinguished between societies that were in the vanguard of civilization, understood as Western civilization, and less-developed peoples. If G. W. Leibniz and Voltaire had still regarded China as a civilization equal and possibly superior to that of the West, Christoph Meiners considered the yellow peoples as inferior to Europeans, and the Blacks as barely different from animals. The notion emerged of a distinctive Western civilization with roots in classical Greece. J. J. Winckelmann's rediscovery of Greek art in 1764 contributed to an idealization of Greek standards of beauty and personal

culture which were adopted as models for a modern culture distinguishing the West from other cultures and giving it its superior standing. Meiners' scientific racism represented an extreme position, but the idea of the unique position of the West and the relative inferiority of all non-Western cultures and peoples became a central assumption in Western historical consciousness.

2. Nineteenth-century Conceptions of History and Historical Science: Between Romantic Historism and Scientistic Positivism

This outlook was dominant well into the twentieth century. It reflected the changes that had taken place in the world in political and military strength, economic power, and scientific and technological development. The two divergent historical outlooks in the first half of the nineteenth century, one represented by Hegel's dialectical philosophy of history, the other by Ranke's historism, shared this outlook despite their philosophic differences. Hegel's perception of the world as an ongoing dialectical process by which, through progressive contradictions, the absolute mind, reason, fulfills itself and takes on concrete forms in the institutional life of the modern European world, and specifically in the Prussian political and social order after the French Revolution and Napoleon, reflects its origins in specifically German political and philosophical traditions. But it has parallels in less involuted intellectual currents of thought outside of Germany, from Turgot, Condorcet, and Thomas Paine in the late eighteenth century to Henri de Saint Simon, Auguste Comte, J. S. Mill, Thomas H. Buckle, Herbert Spencer, and Benedetto Croce in the nineteenth century. For all of them world history is the history of the European world, a world which for some may include the United States but for all excludes the Slavic peoples. For Hegel only the West has a history. China, India, and Persia may have been the earliest civilizations but they have remained stagnant.

According to Hegel, Black Africans are incapable of cultural creativity. Karl Marx with his conception of static Asiatic modes of production took over much of Hegel's conception of the inferiority of non-Western societies and he and Friedrich Engels used this to justify Western imperialism as the only way to bring the non-Western world into the realm of modern civilization.

Theoretically Ranke's conception of the historical world represented the antithesis of Hegel's. Ranke rejected the idea of progress as well as the schematization of historical development as it occurred in Hegel's philosophy of history. History for him deals with the particular and the contingent and avoids generalizations and determinism. In fact, however, Ranke saw history as a linear process at the culmination of which stood the post-Napoleonic Western and Central Euro-

pean world. Ranke proclaimed that 'every age is equally immediate to God' (1973, p. 52), but at the same time denied that China or India had a true history. Except for his early work on the Serbs, his writings deal with the countries he defined early as the Latin and Germanic. His incomplete *World History* is a history of the West.

Whatever the philosophical and political differences, nineteenth-century thinkers as well as the educated public generally turned to history as the key to understanding things human that replaced traditional philosophy. The charges, often repeated, that the eighteenth century, particularly the Enlightenment, ignored history, generalized, and moralized, and did not recognize individuality and uniqueness of historical contexts is disproved when we look at the great historical works of the eighteenth century which we have already mentioned. Despite the impact of romanticism in the first part of the nineteenth century, with its stress on imagination and poetry, a cult of science dominated historical study and writing, which was not necessarily in conflict with a romantic outlook.

This search for a scientific approach to history took two different directions: one that has frequently been called positivism, the other historism, neither of which can easily be reduced to a common denominator.

The first of these orientations took the natural sciences as its model. Building on Turgot and Condorcet, Auguste Comte sought to formulate laws of development by which societies moved from simpler to more complex forms of organization. The driving force was the transformation of knowledge from theological and then metaphysical stages to positive science that would provide the means for the scientific organization of society. But positivism remained largely grand theory with little empirical application to historical inquiry. More significant was Karl Marx's approach to history. Marx saw the driving force of history in the lawful development of economic forces, which constituted the basis of historical transformation, and the revolt of this basis against the social, political, and intellectual superstructure. In the course of this conflict the contradictions within the social order, which involved inequality, would be overcome in a classless society. Marx thus intended to replace Hegel's idealist dialectic with a materialistic one.

If positivism both in its Comtean and in its Marxist form in its search for a scientific history envisaged a utopia at the end of history, the main currents of historical scholarship sought to avoid a speculative approach to history but nevertheless did not free themselves from metaphysical and even theological presuppositions which they refused to acknowledge. Ranke may perhaps serve as the most influential representative of this orientation. He too was committed to a strictly scientific approach to history. For him, however, this approach must recognize that general laws have no application to history that deals

with the particular and the contingent. For him the task of the historian was to recapture the past as 'it actually was' (*wie es eigentlich gewesen*), free of any bias, requiring the application of philological methods to the critical analysis of the sources as F. A. Wolf and Niebuhr had applied them. But for him history was more than the reconstruction of the past. It involved interpretation and synthesis, as it did for J. G. Droysen. Ranke was convinced that the particular event contained the general, and that the general became apparent to the researcher who immersed himself in the sources, because—and here we return from the realm of strict science to that of metaphysics—the historian can observe the general tendencies that manifest themselves in history, tendencies which do not represent God's direct intervention in the world, but nevertheless reflected divine purpose.

3. Professionalization of Historical Studies

The stress on the scientific character of historical studies led to the formation of a professional discipline of history. Professionalization was not restricted to history but was part of a process in which scholarly studies in all fields, in the liberal arts as well as the natural sciences, were institutionalized at centers of higher learning, primarily the universities and research institutes. The German model was imitated elsewhere, relatively early in France with the introduction of the German seminar method at the *Ecole Pratique des Hautes Etudes* in Paris in 1868 and in the United States with the introduction of a research oriented graduate program at The Johns Hopkins University in Baltimore in the 1870s leading to a Ph.D. With some delay the German model was adopted in Japan, Great Britain, the Netherlands, Italy, Spain, Russia, Scandinavia, Greece, Romania, Latin America, and elsewhere, each time, it must be stressed, with major modifications that reflected different scholarly traditions. Thus, the sharp distinction between the cultural and the natural sciences which dominated German thought played less of a role elsewhere.

A basic tenet of professional historians was the total objectivity or what Ranke called the 'impartiality' of the historian. In fact, however, the new discipline was highly politicized everywhere. This intensified interest in archival research went hand in hand with the wave of nationalism that gripped all Western countries and also modernizing non-Western countries such as Japan. The broad cosmopolitan view of many historians in the eighteenth century gave way to a new focus on national history. Ranke still wrote the histories of a variety of European states, Serbia, England, France, Prussia, and also the Reformation in Germany, and the Papacy. Later historians, who took him as a model for their own scholarship, in most cases concentrated on their own nation.

However, the development of modern historical scholarship was not a German monopoly. The opening of archives and the collection of documents, primarily those of the Middle Ages, became a prerequisite for scholarly national history. In Paris the *Ecole des Chartes* was founded in 1821 to train scholars in the critical use of documents. Muratori's collection of Italian documents in the eighteenth century has already been mentioned; this was followed by the *Monumenta Germaniae historica* launched in Germany in 1819 and by similar undertakings in France, Great Britain, Spain, Poland, Bohemia, Greece, and elsewhere.

The Rankean belief that immersion into the sources guarantees an impartial view of the interrelation among historical events turned out to be an illusion, something which was openly recognized by the scholars of the Prussian school such as Heinrich von Sybel, J. G. Droysen, and Heinrich von Treitschke who frankly saw historical scholarship as a tool for the legitimation and propagation of the German national cause under Hohenzollern rule. Moreover, the process by which scholars were recruited guaranteed a high degree of homogeneity in social composition and political outlook.

Professionalization was to guarantee objectivity, but also to provide a scientific or scholarly discourse distinct from that of literature. Ranke had stressed the scientific character of history, but at the same time he also recognized that history combines science with art. The great prestige which history enjoyed everywhere in the nineteenth century—a number of the leading French statesmen from François Guizot to Alexis de Tocqueville, Alphonse de Lamartine, Adolphe Thiers, and Jean Jaurès were important historians—rested on the literary quality of their histories. History could not escape being literature, and imaginative literature at that. In fact there existed a close parallel between the great histories written by professional historians and the great realistic novels of the nineteenth century—Stendhal, Honoré de Balzac, Charles Dickens, Gustave Flaubert, Gottfried Keller, Gustav Freytag, Ivan S. Turgenev, Fyodor Dostoyevski, and even Thomas Mann's *Buddenbrooks* (1902). They all constructed a plot that followed a temporal line and involved individual subjects as agents. And the great historians of the century, Jules Michelet, Denis Fustel de Coulanges, Thomas B. Macaulay, and Theodor Mommsen, who in 1902 received the Nobel Prize for literature, were read as great literature, as was Ranke himself. Michelet is a striking example of the crossing of borders between critical scholarship and imaginative literature, the exhaustive mining of the archives for the purpose of creating a national memory and identity.

But the role professionalization played in nineteenth-century historical writing should not be overestimated as it often has been. A great deal of history did not follow the dominant paradigm of narrative

national history such as three of the most important and lasting works of the mid-century, Fustel de Coulanges's *La Cité antique* (The Ancient City), de Tocqueville's *L'Ancien régime et la Révolution Française* (The Old Regime and the French Revolution), and Jacob Burckhardt's *Die Kultur der Renaissance in Italien* (The Civilization of the Renaissance in Italy). Moreover, many leading writers of history such as Thomas Carlyle, Macaulay, John Green, George Bancroft, Lamartine, Louis Blanc, and, of course Tocqueville, were not professional historians.

The great literary histories of the eighteenth century, for example, those of Hume, Gibbon, and Voltaire, were intended to entertain, educate, and edify a broad educated public. Professional historians wrote often for specialists. But in fact the division between literary and professional history and their audiences was not as deep as it may appear. Certainly Hume, Gibbon, and also William Robertson were serious scholars and the great professional historians of the nineteenth century, such as Ranke, Fustel de Coulanges, Theodor Mommsen, and Lord Acton were read avidly by a broad public for whom history had become the key science replacing philosophy as a guide to understanding the human world.

4. Twentieth-century Reorientations of Historical Thought and Scholarship: The Turn to the Social Sciences

Outside the discipline proper, critical methods were applied to many areas: the history of religion (F. D. Strauss, Ernest Renan), the history of law, the history of literature and art, and other fields. While the historians proper neglected the study of society and economics, an important group of economic historians, the so-called 'Historical School of National Economy' in Germany (Bruno Hildebrand, Karl Knies, Wilhelm Roscher, and Gustav Schmoller) pursued a middle course between an empirical approach, seeking generalizations of economic behavior and development, reminiscent of positivism, and an intense regard for the particular historical setting of culture and society in which economic behavior occurred, which reflected historicist principles.

The late nineteenth and the turn to the twentieth century witnessed profound reexaminations of the historical outlook that had been dominant until then. On the one hand, there was a continued, even a heightened faith in science, technology, and the progress of Western civilization which largely survived the trauma of World War I. On the other hand, there was an increasing sense of discomfort with the quality of modern civilization. Darwinism, or rather Social Darwinism, played a fundamental role in both of these outlooks. The world was seen as a struggle for the survival of the fittest that could forecast the pro-

gressive victory of the white races over the rest of the world, which for Oswald Spengler was a desperate struggle for survival. At the turn to the twentieth century, some Europeans and Americans foresaw progress in terms of growing tolerance and democracy, while others preached racism and ultra nationalism.

As far as historical thought was concerned, both Burckhardt and Friedrich Nietzsche voiced their discomfort with a modern industrial society with its concomitant social and cultural aspects, a mass civilization moving to commercialism and democracy, and extended this dismay to a critique of the modern faith in science. Nietzsche, Wilhelm Dilthey, Henri Bergson, and Georges Sorel in differing ways called for a restoration of the vital against the mechanical aspects of life and culture. Established patterns of historical scholarship, as practiced by the historical profession, were called into question simultaneously at the end of the nineteenth century from Russia and Poland to North America. Generally, however, the critics of the historical profession did not criticize the latter because it was too scientific but rather because it was insufficiently so and had not taken into consideration the tremendous changes in modern society and civilization in an industrial and democratic age.

The most spectacular confrontation with the established profession occurred in Germany with the so-called *Methodenstreit*, occasioned by the publication of the first volume of Karl Lamprecht's *Deutsche Geschichte* in 1891. Lamprecht criticized his colleagues for their narrow focus on the state and their neglect of broader cultural and social aspects of history. He also accused them of working with an outdated concept of science and historical science that restricted itself to events and personalities rather than taking into consideration the structural context of historical change. The German professoriate, reflecting the peculiar semi autocratic character of the Prussian German state, vociferously rejected Lamprecht's initiative and saw in it a threat to the established social and political order. Lamprecht's attempt to formulate laws of historical development was rejected internationally but his call for a broadening of historical perspective to encompass society, economics, and culture was generally welcomed outside of Germany. Historians in Russia (Paul Vinogradoff, Vasiliev Kliuchevskii, Paul Miljukov), Romania (Nikolae Iorga), Poland (Jan Bujak, Frantczek Rutkowski), Norway (Halvdan Koht), Belgium (Henri Pirenne), as well as the circle around the journal *Revue de synthèse historique* founded by the philosopher Henri Berr in 1900 as a forum for an international discussion, and the American 'New Historians' (F. J. Turner, J. H. Robinson, Charles Beard, Carl Becker, H. E. Barnes) called for an expansion of the scope of history and a closer cooperation with the empirical social sciences.

Two variations of a social science approach to history need to be mentioned. The first is Max Weber's historical sociology that moves between the historicist

emphasis on the recognition of the uniqueness of cultures and a historical social science working with clearly defined concepts and a rigorous logic of inquiry. Weber's conception of social causation and historical development was in some ways indebted to Marx, but unlike Marx emphasized the interrelation between culture, including religion, and economics, and placed much greater weight on contingency. Weber recognized that societies constituted meaningful relationships that required an interpretive sociology but avoided the sharp distinction that idealistic philosophers such as Wilhelm Windelband, Heinrich Rickert, and Wilhelm Dilthey made between the generalizing natural sciences and the cultural sciences (*Geisteswissenschaften*) that favored empathetic understanding (*Verstehen*).

A second approach to a history taking into account the social sciences was offered by the so-called 'Annales school.' It is difficult to reduce the *Annales* to a common denominator. The journal by this name, launched in France in 1929 by Lucien Febvre and Marc Bloch with the support of Henri Pirenne, continued initiatives begun early in the century in Henri Berr's already mentioned *Revue de synthèse historique*. A similar project was initiated in Poland in the journal *Roczniki Dziejow Społecznych i Gospodarczych*, the editors of which, Rutkowski and Bujak, were in close contact with the *Annales*. In the place of the concentration of the established historiography on events and great personalities, the *Annales* explored impersonal forces and structures.

The 'human geography' of Vidal de la Blache played an important role in the early works of Lucien Febvre and Marc Bloch. Human history had to be understood in the concrete context of space and time. But unlike the German Friedrich Ratzel, who saw soil and race (*Volk*) as determining factors in history, Paul Vidal de la Blache, as did Febvre, Bloch, and Fernand Braudel later, recognized the role which human agency and culture played in the formation of the geographic setting. The historians of the *Annales* attempted to pursue a 'total history,' generally limited to a restricted region, in which all aspects, geographic, economic, political, religious, and demographic, were to be included. Quite consciously the dividing line between the established disciplines were overcome and in the *Maison des Sciences de l'Homme*, opened in Paris in 1968, economists, sociologists, anthropologists, linguists, psychoanalysts, and literary critics were housed with history as the integrating science.

In other ways too, the *Annales* represented a radical break with older historical conceptions. The nation no longer represented the basic unit of history; it was replaced by the region or by a broad comparative approach that extended beyond Europe (Braudel, Pierre Vilar). The concept of a linear time, so basic to all forms of nineteenth century historical thought, and with it the idea of progress, were abandoned and a historical epoch viewed in its own terms, not as part of

a historical process. Burckhardt had already done something similar in 1860 in *Die Kultur der Renaissance in Italien*, and had worked with a very broad concept of culture, as had also Jan Huizinga in *Herfsttij der Middeleeuwen* (1919), but unlike the *Annales* historians they focused heavily on elites. In contrast the *Annales* dealt with the population at large, but ran the danger that at times their history became impersonal and void of concrete human beings.

In retrospect, the racist historiography of German National Socialism of the period 1933–1945, although it built on widespread Social Darwinist ideas and ultra nationalism, remained an interlude, which even in Germany had limited influence. The Marxism-Leninism of the Soviet Union and its dependent states in its dogmatic forms, in many ways a perversion of Marxism, similarly represents an interlude. As early as 1956, historians in Poland and Hungary, and until the freeze after the Soviet occupation in Czechoslovakia, and even in the Soviet Union (Aaron Gurevich), participated in the broader historical discussions of the international community.

5. New Orientations after 1945: The Turn to a Cultural Dimension

A subsequent article will deal with trends in historical thought and study in the second half of the twentieth century. The victory of the Allies in World War II and the dominance of the United States in the postwar world revived older beliefs in progress and science. The 1950s and 1960s witnessed a turn to the quantitative social sciences in historical studies.

The years since the late 1960s have seen a re-orientation in historical outlook in the Western world generally, including to a lesser extent also Eastern Europe and Russia. The conflicts of the 1960s, the racial confrontation in the United States, the Vietnam War, the student revolts throughout Europe and North America, and the suppression of the Prague spring all contributed to a reorientation as did the fundamental technological changes which resulted in the transformation of an industrial society concentrated on production into an economy based on information technology, in which consumption and services play a key role and the old socioeconomic classes gave way to new ones. The confidence in science and economic growth was severely tested as a new generation became increasingly conscious of ecological concerns and called for the emancipation of subordinated segments of the population such as women and ethnic minorities at home and the peoples of the former colonies abroad. The very quality of the modern Western world, including its science, came under attack. All of this had consequences for historical studies after the 1960s.

See also: Historicism; Historiography and Historical Thought; Christian Tradition; History and the Social Sciences; History: Forms of Presentation, Discourses, and Functions; History, Professionalization of; History: Theories and Methods; Modernity; Modernization and Modernity in History

White H 1973 *Metahistory. The Historical Imagination in Nineteenth-century Europe*. The Johns Hopkins University Press, Baltimore, MD

Woolf D R (ed.) 1998 *A Global Encyclopedia of Historical Writing*. 2 vols. Garland Publishing, New York

G. G. Iggers

Bibliography

Bann S 1984 *The Clothing of Clio: A Study of the Representation of History in Nineteenth-century Britain and France*. Cambridge University Press, New York

Barnes H 1962 *A History of Historical Writing*. Dover, New York

Barracough G 1978 *Trends in History*. Holmes & Meier, New York

Bentley M 1999 *Modern Historiography*. Routledge, London

Berr H, Febvre, L 1932 History In: *Encyclopedia of the Social Sciences*, Macmillan, New York, Vol. 7, pp. 357–68

Blanke H W 1991 *Historiographieggeschichte als Historik*. Frommann-Holzbock, Stuttgart-Bad Canstatt, Germany

Breisach E 1983 *Historiography. Ancient, Medieval, Modern*. University of Chicago Press, Chicago

Butterfield E 1955 *Man on His Past. The Study of the History of Historical Scholarship*. Cambridge University Press, Cambridge, UK

Cassirer E *Philosophie der Aufklärung (The Philosophy of the Enlightenment)*. Mohr, Tübingen, Germany

Chickering R 1993 *Karl Lamprecht: A German Academic Life (1856–1915)*. Humanities Press, Atlantic Highlands, NJ

Fueter E 1911 *Geschichte der neueren Historiographie (History of Modern Historiography)*. Oldenbourg, Munich, Germany

Gazi E 2000 'Scientific' History. *The Greek Case in Comparative Perspective (1850–1920)*. Peter Lang, New York

Gooch G 1913 *History and Historians in the Nineteenth Century*. Macmillan, London

Iggers G 1968 *The German Conception of History. The National Tradition of Historical Thought from Herder to the Present*. Wesleyan University Press, Middletown, CO

Kelley D 1998 *Faces of History. From Herodotus to Herder*. Yale University Press, New Haven, CT

Keylor W 1975 *Academy and Community. The Foundation of the French Historical Profession*. Harvard University Press, Cambridge, MA

Koselleck R 1975 Geschichte, Historie. In: Brunner O, Conze W, Koselleck R (eds.) *Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland* Vol. 2. Klett, Stuttgart, Germany, pp. 593–717

Novick P 1988 *That Noble Dream. The 'Objectivity Question' and the American Historical Profession*. Cambridge University Press, Cambridge, UK

Ranke L von 1973 The theory and practice of history. In: Iggers G, von Moltke K (eds.) Bobbs Merrill, Indianapolis, IN

Reill P H 1975 *The German Enlightenment and the Rise of Historicism*. University of California Press, Berkeley, CA

Ringer F 1969 *The Decline of the German Mandarins. The German Academic Community 1890–1933*. Harvard University Press, Cambridge, MA

Thompson J W 1942 *A History of Historical Writing*, 2 vols. Macmillan, New York

Historiography and Historical Thought: South Asia

This article addresses two interlinked subject matters, the history of South Asia to which scholars from all over the world have contributed, and historiographical developments in the region of South Asia itself. It briefly covers the historical tradition in pre-modern times and primarily focuses on developments since World War II.

1. Historiography of Ancient India

1.1 Historical Traditions in Ancient India

Compared to classical Europe, ancient India was rather deficient in historical traditions. The Hindu mythological texts called the *Puranas* (circa first to fourteenth century) contained genealogical lists of Indian dynasties. These lists, by way of narrating the achievements of the kings, also recorded rudimentary information about wars, conquests, and the rise and fall of empires. A very small number of works chronicle the history of individual kings and dynasties. Of these, the most famous is Kalhana's *Rajatarangini* (Stein 1900), a history of the Kashmir princes. Incidentally, the author recognized one historical principle, namely, that the historian must try to be unbiased in his/her judgment, 'like a judge.'

The Buddhists and Jains developed a tradition of sacred history, especially hagiographies centered on the life-stories of their founders. Some of the stories concerning the Buddha, like the account of his death (*Mahaparinivvana-sutta*, Bongard-Levin 1986) bear the stamp of authenticity.

1.2 European Scholarship on Ancient India

In the nineteenth century, European scholars like Pargiter tried to reconstruct the dynastic history of ancient India using the Puranic lists, inscriptions, coins, archaeological remains, and foreign accounts of India like those of Alexander's historians and the Chinese Buddhist pilgrims. European Orientalists also tried to reconstruct the religious and social history of India, using religious texts and the *smritis*, normative

writings laying down the rules by which the Brahminical Hindu was expected to live in matters both ritualistic and secular (like inheritance and punishment for crimes). They also undertook the massive task of publishing critical editions of religious texts which laid the foundation for serious work on religious, cultural, and social history. A large number of German, French, and East European scholars have contributed to this enterprise.

1.3 *Historiography of Ancient India: The Indian Contribution*

A large body of Indian scholars started research on ancient Indian history following the lead of the European scholars. Their work was heavily focused on archaeological evidence, especially the study of inscriptions. The archaeological work, by Europeans and Indians, helped move back the known beginning of civilization in South Asia to about 3000 BC through the discovery of the Indus valley civilization. The work of the Indian Indologists also reflected the new-found nationalistic pride in the glories of ancient India. Some of it focused on traditions of democratic/oligarchic politics and developed statecraft to challenge the notion that these were European imports into India. However, the central concern of the earlier work was with the meticulous reconstruction of dynastic history, especially the correct chronology. Some scholars like R. K. Mukherji and R. C. Majumdar did research on the history of Indian cultural influence in south-east Asia and invented the now discarded notion of Greater India.

1.4 'Orientalism': A Controversy

Edward Said's *Orientalism* (1995), which projects the view that western scholarship on Asia projected the notion of an inferior 'other' of Europe and thus reflected the new relationships of power generated by imperialism, has had a seminal influence on the historical work on South Asia (since 1980) even though Said's primary point of reference was the Middle East. The controversy generated by this work has drawn attention to the many strands of Orientalist studies in the West and to the fact that French and German studies of South Asia do not project any unquestioned superiority. For ancient India, the controversy has led to more research on western historiography than to any reorientation in exploration of substantive themes.

1.5 *Indology after World War II*

The post-World War II historiography of ancient India has moved away from these earlier preoccupa-

tions and, under the influence of Marxism as well as anthropology, concerned itself with discovering structures like feudalism and patterns of large-scale social change over time. The writings of D. D. Kasambi (1965), R. S. Sharma (1990), Romila Thapar et al. (1990) are prime examples of this initiative. The multi-volume *History and Culture of the Indian People* edited by R. C. Majumdar (1951–1969) reflects a more traditional approach, while Louis Dumont's *Homo Hierarchicus* (1972) applies Levi-Strauss' structural analysis to the study of the caste system.

2. *Historiography of the Indo-Islamic Period*

The thirteenth century saw Turkish and Afghan dynasties and tribes, Muslims by faith, firmly established as rulers in many parts of India. The sultanate based in Delhi, where there was a succession of dynasties, became the seat of a pan-Indian empire with fluctuating boundaries. These dynasties were succeeded by the Mughals or Timurids in the sixteenth century and their effective rule lasted until the late eighteenth century. This long period, marked by the encounter between the older Indian civilization and the religion and culture of Central Asia's Islamic peoples, has been the subject matter of a rich historical literature.

2.1 *The Medieval Persian Chronicles*

India's central Asian invaders brought with them a rich tradition of historiography. This literature, written mostly in Persian, is known as *tarikhs*, a term, translated as annals. The *tarikhs* are narratives of dynastic successions, wars and conquests. Some of the writers had leanings towards Islamic fundamentalism and hence were inclined to present the wars of the Muslim invaders and Hindu dynasts as *jihad* or 'holy' wars. These accounts were later used by some British writers to project the Indo-Islamic era as an age of Muslim tyranny. This perception informed the ideology of India's proto-nationalism, which saw the struggle of the Rajput clans and other Hindu chiefs to retain their territories as the Hindu struggle for independence. However, chronicles provide ample evidence of a pragmatic bent in the policies of the Muslim kings, anxious not to alienate their Hindu subjects. The historians at the court of the third Mughal emperor Akbar, project his policy of peace with all, which was not to the liking of the more orthodox elements at his court. One component of the historical literature of the period is the memoirs written by the emperors, Babar and Jahangir and by Babar's daughter, Gulbadan. These sensitively written accounts offer a portrait of the age and the family life of the Mughals.

2.2 Regional Histories

The Indo-Islamic era also produced a rich and varied body of historical writings written in the regional languages. Best known among these are the Marathi *bakhars*, modeled on the Persian *tarikhs*, the chronicles of the Ahom kings of Assam, known as *buranjis*, and the *Madla Panji* retained in Orissan temples which recorded the annals of the local dynasties. Besides these, the Vaishnavas in Bengal produced a body of hagiographical literature celebrating the lives of their founder, Chaitanya and other pioneers of the movement. All these products of the period have been later used to reconstruct the history of the period.

2.3 Colonial and Post-colonial Historiography of the Indo-Islamic Era

The earliest 'modern' historical writings on the Indo-Islamic era go back to the nineteenth century. Works like Orme's *History of Indostan* are amateur efforts focused on the political history of the period. The famous collection of selected source material, translations from the Persian chronicles, Eliot and Dawson's *History of India as told by her own Historians* (1981), contain introductions which are essentially historical statements. One of the compiler's stated objects was to prove how very inferior 'Muslim rule' was compared to its British counterpart. In general histories of India, written by administrator-scholars like Vincent Smith (1981) the same concern to prove the superiority of British rule is less obviously present. It only appears as a view of Indian history which projects a long-term proneness towards political unification that finally succeeds only under the British. Such works contain an implicit evaluation of moral and material welfare of the people in pre-British times. The twentieth-century economic historian W. H. Moreland (1920) undertook a comparison between India at the death of emperor Akbar with conditions in the nineteenth century. Serious academic research on the Indo-Islamic era, based on the study of the chronicles as well as documentary evidence like the *akhbars* or newsletters of the Mughal period, belong to the twentieth century. It produced magnificent works of scholarship like Sir Jadunath Sarkar's multi-volume histories of Aurangzeb and the later Mughals focused on political narratives. There were other studies dealing with the religious policies of the Mughals and art and architecture of the period. The burgeoning nationalist consciousness sought to emphasize the syncretist traditions of the Indo-Islamic era, implying that the conflict between communities was a product of British rule.

Since World War II, a major development in this field of study resulted from the work of a group of scholars based at Aligarh. Irfan Habib's *Agrarian System of Mughal India* (2000), drawing upon a wide

range of source material, was hailed as a path-breaking contribution to the study of agrarian history. The research of the Aligarh 'school' uses the methods of modern historiography and has drawn upon a vast quantity of new material. While some of the scholars subscribe to Marxist theories of history, the impact of Marxism is not very obvious in their work.

In recent years some Western historians have projected the view that the Mughal empire, like all its predecessors in India, was a 'segmentary state' under a thin veneer of centralization. This has led to a controversy with some of the Aligarh historians.

3. Historiography of the Colonial Era

3.1 Colonial Historiography under Colonial Rule

British historians writing in the colonial era produced a large body of works which recorded the history of the conquest, the wars with indigenous powers, and histories of the local powers like the Marathas, Sikhs, and Mysore with whom they had to contend for power. These histories are in the purely narrative mode, supplemented by ethnographic accounts of various elements in Indian society. They are informed by a simple assumption regarding the legitimacy and naturalness of the British conquest. One of the pioneers in the field, James Mill (Mill and Wilson 1997), writing in the early decades of the nineteenth century, first projected a clear notion of cultural and racial superiority, significantly absent from earlier British writings on India.

British writings in the latter half of the nineteenth and early decades of the twentieth century often assumed a defensive posture in response to the growing nationalist criticism of their rule, especially the charge that the colonial nexus had induced a drainage of resources from India, thus steadily impoverishing the country. British historians challenged this view and instead projected a picture of economic transformation and development as a result of colonial rule. This controversy has not yet died out. The British historiography of colonial rule before World War II almost entirely focused on the activities of the government. A number of studies dealt with individual viceroys. Some twentieth-century works briefly refer to the burgeoning nationalist movement (one historian described it under the heading 'Political Sentiment') but with one exception they dismissed it as a phenomenon not to be taken seriously until the very eve of the transfer of power. However, the British colonial historiography increasingly adopted the methods of modern research with its emphasis on 'objective' analysis of records. But since the records in question were for the most part products of the colonial government, the resulting research inevitably reflected the built-in bias of the documents, a bias which the

historian generally shared. Indian historians, excepting those writing on economic history, rarely ventured into comments on the history of British rule. Exceptions to this pattern were the publicists writing on the rising of 1857 (one of whom described it as India's First War of Independence) and authors who considered British rule to be beneficial for India. The strong ideological and political undertone manifest in the historiography of the colonial era has assumed more subtle forms in the post-colonial period.

3.2 *Historiography of Modern India after Decolonisation*

Since the transfer of power in 1947, attention has shifted from the working of the British administration to the structure and dynamics of indigenous politics, though a number of monographs on the political history of the Raj, as well as its administrative and constitutional initiatives, were written during this period. Indian historians now came to be deeply concerned with the history of the Indian nationalist movement and its social-cultural origins. A group of historians approached the subject from the overt perspective of nationalism; others were influenced by Marxist class analysis and saw Indian nationalism as an expression of colonial bourgeois aspirations. A group of British historians centered at Cambridge explored the close relationship between British administrative and colonial initiatives on the one hand, and the structure and dynamics of indigenous politics on the other. The competition between those who were admitted to a share in power by the colonial ruler and those who were not was seen to be the central feature of indigenous politics. Nationalism itself was perceived to be a strategy reflecting the aspirations of the 'have-nots.' This view was questioned by Indian historians who saw the thesis as a continuation of the old colonialist perception dismissing nationalism as a conspiracy of the self-seeking few. However, the thesis was directly, or indirectly, questioned outside India as well. A group of historians under the leadership of Professor Low of the Australian National University recognized the reality of Indian nationalism and proceeded to unravel its local and class roots. American scholars brought to the subject the insights of social sciences, especially sociology and anthropology. There were considerable developments in economic history, using economic theory as well as statistical methods. The new techniques were used to tackle the old questions regarding the negative or positive impact of colonialism. The debate, as noted, remains inconclusive. The inadequacy of statistical evidence is one reason for this state. It should be noted here that Indian historical research has now become an international enterprise to which the Japanese, Russian, and continental scholars have made significant contributions.

One major new development in the subject is centered on the series of publications entitled *Subaltern Studies* (Guha et al. 1988). Borrowing the concept of the Subaltern from the Italian Marxist, Gramsci, the contributors to this series have focused on the history of the underprivileged elements in South Asian society and used the post-modernist techniques of discourse analysis to unravel the autonomous consciousness of the dispossessed and the record of their resistance to exploitation and domination by the elite. This movement, too, has become an international enterprise with historians and social scientists from all parts of the anglophone world contributing to the *Subaltern Studies*.

Pakistani historians and their pre-independence predecessors who focused on the development of Muslim consciousness and aspirations in the sub-continent have been deeply influenced by the theory of two nations, the ideological basis of the demand for Pakistan. Their perception is in sharp contrast to the nationalist perception of co-existence and mutual accommodation between the major ethnic cultures of the region. The history of Sri Lanka, despite its many links with the civilization and society of the mainland, has been unjustifiably peripheral to the concerns of the historians of India. But there was a strong tradition of historiography in the region drawing, *inter alia*, on the Buddhist religious texts for the ancient period. Some of the products of this historiography are influenced by Marxist theories of history.

One marked feature of developments in the historiography of modern India is the diversification of its subject matter. History of art and its links with nationalist sentiment, women's studies, and ecological history are some of the areas explored in recent years. This has helped bring the subject within the sphere of contemporary intellectual interest and brings to an end its intellectual isolation.

See also: Colonization and Colonialism, History of; Historical Sociology; Historiography and Historical Thought: Current Trends; History and the Social Sciences; History: Overview; Imperialism, History of; Orientalism; South Asian Studies: Culture; South Asian Studies: Economics; South Asian Studies: Politics; Subaltern History; Time, Chronology, and Periodization in History

Bibliography

- Bongard-Levin G M (ed.) 1986 *Tripitaka, Sutraṭitaka, Mahāparinirvāṇasūtra*. International Institute of Buddhist Studies, Tokyo
- Dumont L 1972 *Homo Hierarchicus: The Caste System and its Implication*. Paladin, London
- Eliot H M, Dowson J 1981 In: Nizami K A (ed.) *The History of India as Told by its own Historians: The Muhammadan Period*. Delhi, India

- Frykenberg R E *India to 1858*; Raychaudhuri T. *India, 1858 to the 1930s*; Robin J. Moore, *India in the 1940s* Winks, R W., 1999 *The Oxford History of the British Empire, Historiography*, Oxford University Press, Oxford Vol. 5
- Gallagher J, Seal A, Johnson G 1973 *Locality, Province and Nation Essays on Indian Politics 1870 to 1940*. Cambridge University Press, Cambridge, UK
- Guha R, Chakravarti S, Gayatri 1988 *Selected Subaltern Studies*. Oxford University Press, New York
- Habib I 2000 *Agrarian System of Mughal India 1556–1707*, 2nd edn. Oxford University Press, New Delhi, India
- Kasambi D D 1965 *The Culture and Civilisation of Ancient India in Historical Outline*. Routledge, London
- Majumdar R C 1951–1969 *History and Culture of the Indian People*, 10 Vol. Bharatiya Vidya Bhavan, Bombay, India
- Mill J, Wilson J H 1997 *The History of British India*. Routledge, London
- Moreland W H 1920 *India at the Death of Akbar*. Macmillan, London
- Philips C H (ed.) 1961 *Historians of India, Pakistan and Ceylon*. Oxford University Press, London
- Said E W 1995 *Orientalism*. Penguin, Harmondsworth, UK
- Sarkar S 1985 *Modern India: 1885–1947*, 2nd ed. Macmillan, New Delhi; 1997 *Writing Social History*, Oxford University Press, Delhi
- Smith V 1981 In: Spear T G (ed.) *The Oxford History of India*. Oxford University Press, Delhi, India
- Stein M A (trans.) 1900 *Kalhana's Rajatarangini: A Chronicle of the Kings of Kashmir*. Westminster, UK
- Sharma R S 1965 *Indian Feudalism, 300–1200*. Calcutta University Press, Calcutta
- Thapar R 1990 *A History of India*. Penguin, Harmondsworth, UK

T. Raychaudhuri

Historiography and Historical Thought: Southeast Asia

Embracing the southeastern peninsula of Eurasia and the Indonesian and Philippine archipelagoes, Southeast Asia has recorded its history in a dozen written languages and countless oral traditions. Indic models of writing were influential in the millennium before 1200 CE, and Chinese and Islamic ones in the subsequent six centuries, while European enlightenment interpretations of history became increasingly dominant in the nineteenth and twentieth centuries. Diversity has always been a hallmark of this region. The Vietnamese emulated the Chinese literary tradition with their dynastic chronicles; cyclical Indic ideas had a longer career in the Theravada Buddhist societies of the mainland, while linear Muslim and Christian ideas became influential in the islands. Nevertheless, synchronicities are apparent in the developments within seemingly autonomous literary and historiographic traditions.

Oral poetry and traditions were important for most peoples, but particularly for those in the highlands and eastern islands (including the Philippines) which did not absorb Indic or Chinese literary forms. As late as the 1960s, at least one people in Roti (near Timor) retained oral historical narratives which accurately recorded the names of chiefs for three or four centuries. Most of the written chronicles which took shape in the sixteenth and seventeenth centuries incorporated older traditions with the repetitive features of oral narrative.

1. Chronicle Traditions

The dynastic chronicles of Vietnam were expressed in Chinese and followed the model of didactic Chinese chronicles heavily informed by the moral perspective of Tang and Sung Confucianism. Some such work survives from the fourteenth century, but the most important is the strictly Confucian chronicle composed by Ngo Si Lien in 1479. This in turn incorporates the *Dai Viet su ky* of Le Van Huu (1272), which included a historical justification for the independent imperial status of Dai Viet—then occupying little more than the Red River delta.

The remaining written traditions grew out of Indic models in Sanskrit and Pali, or in the Malay case from Islamic writing in Arabic and Persian. Sacred histories, in Mon (the then language of what is today southern Burma), Burmese, and Khmer, as well as in Pali, traced the establishment of Buddhist law and practice from the time of the Buddha to the founding of local temples in Southeast Asia. Among the earliest examples are the northern Thai *Mulasasana* (around 1420) and the Pali *Jinakalamalipakaranam* (1516), both written by monks in Chiang Mai.

Royal and dynastic chronicles are almost as old. The *Pararaton*, a unique chronicle in old Javanese of the Javanese kingdoms of Singosari and Majapahit down to 1481, was written after the fall of Majapahit to Muslim forces around 1527. The earliest extant *yazawin* (royal genealogy or chronicle) in Burmese was written by the monk Shin Thiawuntha in 1520. Dynastic chronicles (*phongsawadan*) were being compiled in Thai from at least the sixteenth century, although the Luang Prasoet text conventionally dated 1680 is the oldest of the extant versions which have come down to posterity. Written by an astrologer at the cosmopolitan court of King Narai (1657–1688), this chronicle marks a trend to more secular dynastic history well aware of Siam's position among its neighbors as well as European, Indian, and Chinese visitors.

The advent of Islam created more discontinuities in the south than Theravada Buddhism did in the north. As Southeast Asian Islam shifted from an Arabic to an indigenous idiom in the fifteenth and sixteenth centuries, the Malay language written in Arabic script

became its favored vehicle. Only in Java and South Sulawesi, and of course in non-Muslim Bali, did the older literary idiom and alphabet continue alongside the Arabic/Malay one. Probably the oldest extant Malay chronicle is that of the Kings of Pasai, in north Sumatra, the first substantial Islamic state in Southeast Asia, from the 1290s. This text, *Hikayat Raja-raja Pasai*, is known in only one copy, made in 1814 from an older version held in Demak, the fifteenth-century capital of Muslim Java. Since it ends with the story of the fourteenth-century conquest of Pasai by Hindu-Buddhist Majapahit (Java), and the bringing of many Pasai Muslims as captives to Java, its kernel may have originated with this group of Malay-speaking Muslim exiles. It represents a collection of semi-mythical stories about the Pasai kings, much more reflective of local oral traditions than of mainstream Islamic historiography.

Several Malay texts celebrating other royal lineages, first written down in the sixteenth or the early seventeenth century, take this story-telling mode to a high art form extremely illuminating for the historian, though usually devoid of chronology. The *Hikayat Banjar*, for example, tells the story of the establishment of the leading Islamic state in southern Borneo through a conservative ethic extolling the virtues of a Javanese cultural model. Much the best-known text of this genre is the 'Genealogy of (Melaka) Kings' (*Sulalat'us-Salatin*), a version of which Thomas Raffles printed and popularized in 1821 under the name 'The Malay Annals' (*Sejarah Melayu*), in a conscious attempt to provide a national history for Malays. It weaves through a host of lively anecdotes its political message that the great metropolis of Melaka fell to the Portuguese in 1511 because the fundamental contract between ruler and ruled was violated by the last Melaka ruler. Originating in some form a generation after this Malay disaster, the text appears to have been essentially created in the early seventeenth century by the Johor writer Tun Seri Lanang, though subsequently much modified. Not much later in time, but a great deal more Islamic in spirit, was the encyclopedic 'Garden of Kings' (*Bustan as-Salatin*), written by the Gujarat (India)-born theologian Nuruddin Ar-Raniri around 1640. Written in Malay at the behest of the Sultan of Aceh, in northern Sumatra, this includes a substantial chapter on the history of Aceh, and another on Melaka, and represents the most skilful use in Malay of Islamic traditions of historiography.

Java's conversion to Islam was a turbulent and drawn-out process centered in the fifteenth and sixteenth centuries. The syncretic Islamo-Javanese courts which emerged from it produced a chronicle tradition in Javanese verse known as *babad*, part mythical, part prophetic and instructional, and part genealogy. The central narratives designed to show the legitimacy of kings descending through the complex upheavals of the wars of Islamization are collectively known as the *Babad Tanah Jawi*. All modern historians, beginning

with Raffles, have drawn on this set of traditions, although the task of editing a canonical text has so far defeated modern scholarship. The most chronologically accurate of the early Javanese histories, the Kartasura verse chronicle *Babad Ing Sangkala*, was written in 1738 but thought to incorporate a particularly accurate text originally compiled before 1670.

A particularly interesting school of historical writing developed in South Sulawesi (Celebes) in the Makasar and Bugis languages. The written syllabary was only introduced (probably from southern Sumatra) in the fifteenth century, and appears to have been especially valued for the recording of aristocratic genealogies. Marriage alliances among the numerous Bugis and Makasar states required a careful chronicle of previous marriages and begettings going back to the supernatural origins of the leading families. From around 1600, genealogies were developed into state chronicles at Makasar, then a major maritime power much influenced by Islamic and European literary models. From the 1620s, this chronicle tradition could draw on court diaries, which recorded royal births, deaths, and marriages, as well as notable foreign visitors, wars, and innovations. Keenly aware of chronology, the state histories adopt a critical tone towards their sources, a distinctly linear notion of the progress of the kingdom through various innovations, and a desire to learn from failures as well as successes. The Goa (Makasar) chronicle began with an explanation of why it was important that future generations remember the past, 'to avoid the twin dangers, either of our thinking that we were all great kings, or of others alleging that we were worthless.'

2. An Autonomous Modernity?

While the seventeenth century witnessed remarkable new departures in historical writing, there was a further development in the late eighteenth century of more abundant historical writing in a self-conscious, secular, and historically careful pattern in several Southeast Asian traditions. While the better survival of manuscripts written after the wars of the eighteenth century may be one explanation, there also appears to have been a creative sense of crisis among Tai and Vietnamese writers in particular which gave rise to reflection on the past. An early sign of this trend was U Kala's *Great Chronicle of Kings* (*Mahazawingyi*), the first comprehensive and chronologically reliable Burmese chronicle, written in the 1720s. In the late eighteenth century this was incorporated into the first history which might be called national, the 'New Chronicle of Myanma' (*Myanma Yazawin*), also the first to use its sources in a critical sense.

Among Vietnamese, the new trend was embodied by the scholar-official Le Quy Don (around 1726–1784), who wrote a Le Dynasty history in biographical form,

often also referred to as a national history—*The Complete History of Dai Viet (Dai Viet Thong Su)*. His encyclopedic *Miscellaneous Records of Pacification in the Border Area (Phu Bien Tap Luc)* is a careful accumulation of primary sources relating to technological, economic, cultural, and political developments in the southern Vietnamese kingdom (known to foreigners as Cochinchina) which was a rival of his own. The Tai kingdoms endured a mauling from Burma in the late eighteenth century, and virtually all responded in the period 1790 to 1830 by carefully reconstructing their histories with a view to avoiding such disasters in the future. In Java for a similar reason there was a spate of chronicle-writing after a long period of warfare ended with the peace of Gianti in 1755.

3. European Influences

These elements of modernity all preceded any direct European influence on the writing of history. The nineteenth century, however, gave rise to numerous writers who were in contact with European scholar-officials with their own agendas for gathering histories of Southeast Asian societies. Among the earliest of these were writers in Malay who cooperated in the quest for manuscripts of English and Dutch officials. Munshi Abdullah bin Abdul-Kadir (1797–1854) was a Malay writer for Stamford Raffles, and used his understanding of British mores to launch a critique of the Malay society of the sultanates. In his vivid autobiographical works, *Hikayat Abdullah* (1849) and *Hikayat Pelayaran Abdullah* (Chronicle of Abdullah's Voyages), he introduced such concepts as progress, race, freedom and the rule of law into Malay historical writing. The Riau court writer Raja Ali Haji (around 1809–1870) was in regular contact with the Dutch Malay scholars of his day and learned something of their perspective on history. The spirit of his two outstanding histories, *Genealogy of Malay and Bugis Princes (Silsilah Melayu dan Bugis dan sekalian raja-raja-nya, 1865)* and *The Precious Gift (Tuhfat al-Nafis, around 1866)* is nevertheless heavily didactic in the best Islamic tradition.

European writer-officials brought to the study of Southeast Asia an approach often extremely skeptical of chronicles, and more concerned with the written records of European travelers and the epigraphic and archaeological evidence from the early states. Beginning with Stamford Raffles' *History of Java* (1817) there was a fascination with the ruins of vanished kingdoms, and a tendency to find their successors in a state of decline which could only be reversed by European rule. The Edinburgh-trained doctor John Crawfurd (1783–1868) deserves particular attention as an historian and encyclopedist of Southeast Asia as a whole. His *History of the Indian Archipelago* (3 Vols., 1820), followed by the *Descriptive Dictionary of the Indian*

Islands and Adjacent Countries (1856) covered almost the whole region, based on first-hand knowledge of Java and Singapore as an administrator, and the official missions he led to Burma, Siam, and Vietnam in the 1820s, each of which yielded published reports which began the rendering into English of the chronicle traditions of these countries.

The Philippines was the area best endowed with early European writings because of the prominence in the colonization of the islands of Spanish religious orders, each of which produced chronicles. Among the best of these, containing much material about the pre-Hispanic Philippines, were those of the Franciscan Marcelo de Ribadneyra (1601), the Jesuit Pedro Quirino (1604), and the Dominican Diego Aduarte (1640). An outstanding later example of the genre was the 14-volume chronicle of Spanish and missionary activity in the islands by the Recollect Juan de la Concepcion (1788–1792). As nineteenth-century Spanish accounts became ever more negative about the achievements of Filipinos, the first generation of nationalists turned for inspiration to the oldest of the Spanish secular writers, Antonio de Morga, whose *Sucesos de las Islas Filipinas* (1609) received its first modern edition in English in 1868. The intellectual pioneer of Philippine nationalism, Jose Rizal, published in Paris a new Spanish edition of the work in 1890, with the professed intention 'to awaken your [Filipino] consciousness of our past, already effaced from your memory, and to rectify what has been falsified and slandered.'

As Southeast Asia became partitioned among the colonial powers in the second half of the nineteenth century, the more talented colonial officials pioneered modern-style 'national' histories in English, French, and Dutch. These usually combined some of the chronicle record of ancient kingdoms, the interventions of foreigners, and the rise of colonial power within the borders later marked out by the colonial powers. Sir Arthur Phayre, an army officer who spent the years 1835 to 1867 in key posts in Burma, was the first of these with his *History of Burma* (1883), followed in 1926 by a book of the same name by G. E. Harvey. Adhémar Leclère wrote a *Histoire du Cambodge* in 1914, some time before the more demanding equivalent for Vietnam, *Histoire Moderne du Pays d'Annam (1592–1820)* (1920), by Charles Maybon. A history of Siam was first published in French on the basis of missionary accounts by Turpin in 1771, but the first of the modern European histories was that of the British Consul in Chiang Mai, W. A. R. Wood (1926), relying much on the chronicles.

Sir Richard Winstedt consciously shaped the Malay-language school syllabus with his revealingly-named *Tawarikh Melayu* ('Malay,' not 'Malayan' History, 1918), in collaboration with Daing Abdul Hamid. Winstedt's English-language *Malaya and its History* began its many editions in only 1948, long after the more blatantly imperial *British Malaya* of Sir

Frank Swettenham (1906). The best of the early Dutch writing on the archipelago was devoted to particular kingdoms and culture areas, particularly Java, though the encyclopedic *Oud en Nieuw Oost-Indien* (8 Vols., 1724–1726) of the protestant pastor Francois Valentijn ranged widely. Only with the need for text-books in Netherlands-Indian schools was there a coherent set of ‘national’ textbooks written, of which the most influential work was the Dutch-triumphalist *Geschiedenis van Nederlandsch Indie* (1930) of F. W. Stapel.

4. Nationalist History

All of Southeast Asia except Siam/Thailand was under Western colonial control from 1900 to 1941, and was ruled by the Japanese for nearly four turbulent years (1942–1945). The sudden Japanese surrender in August 1945 inaugurated a period of revolutionary struggles in the successor states, with relative peace imposed on Indonesia only around 1966, Vietnam after 1975, and Cambodia in the 1990s. The first half of the twentieth century was a period of extraordinary debate and rethinking of the meaning of history, as a Western-educated generation increasingly detached from cultural connections to the past developed new ways of using it for their modernizing purposes. First Social Darwinism and then Marxism were seized upon as means to explain what had to be done to save the nation from colonial humiliation. By the 1950s, a nationalist, anticolonial historiography was dominant everywhere except Malaya/Malaysia and Thailand, turning militant opponents of the colonial order into the most exemplary heroes. Marxist categories were selectively adopted, often to devalue the precolonial past as generically ‘feudal.’ Debates remained intense between communist, religious, and ethno-nationalist readings, and often between local and national perspectives.

In Indonesia, nationalist writers suddenly occupied center stage when the Japanese occupation removed all Dutch textbooks. Muhammad Yamin (1903–1962), a Sumatran convert to a Java-centric view of nationalist history, became the most influential writer and patron of a nationalist history which glorified Sriwijaya and Majapahit as precursors of modern Indonesia. His work on the fourteenth-century Javanese kingdom of Majapahit, *Gajah Mada* (1945) and *Tatanegara Majapahit* (3 Vols., 1962), has long-term significance.

In Thailand a similarly influential figure through the most nationalist phase was Wichit Wathakan (1898–1962), a commoner who became an immensely popular writer of historical novels and plays, as well as a 12-volume world history (1931). After the anti-monarchy revolution of 1932, he replaced Prince Damrong Rajanubhab (1862–1943) as the government’s arbiter of culture and history. Where Damrong had chron-

icled a Thai identity developed from the monarchy’s gradual adoption of liberal modern values, Wichit glorified the purity and heroism of the Thai race.

In Vietnam the debate between traditionalists, nationalists, and Marxists centered around the evaluation of the Nguyen dynasty which came to power with some French assistance in 1802, and of the Tay Son rebellion which preceded it in the late eighteenth century. Tran Trong Kim (1883–1945), author of the first modern Vietnamese history (1928, constantly reprinted until the 1960s) and briefly Prime Minister under the Japanese, broke with the dynastic chroniclers in accepting the Tay Son leader Quang Trung as a national hero. Nevertheless he accepted the first Nguyen emperor Gia Long, who defeated Quang Trung’s forces, as a dynastic founder and unifier. Dao Duy Anh’s *Short History of Vietnamese Culture* (1938) was the first and perhaps best of the Marxist-influenced studies which saw Vietnamese culture as constantly evolving through social forces. For later Marxists, the Tay Son were heroic peasant rebels, and the last dynasty a disaster which failed to adapt to historical conditions.

5. Historiography of Southeast Asia as a Whole

These nationalist preoccupations allowed little incentive for indigenous historians to enquire into larger Southeast Asian historical patterns except through the Marxist mold. One of the few exceptions is Nguyen Van Huyen (1908–1975), later Education Minister in the Hanoi government. As a student and Vietnamese teacher in Paris he discovered Dutch scholarship on Indonesia, and used it to good effect in his study of common patterns of housebuilding in the region—*Introduction à l’Etude de l’Habitation sur Pilotis dans l’Asie du Sud-est* (1934).

For the most part it was European scholars interested in the precolonial history of the region who found a broader regional perspective useful. German-speaking prehistorians like Heger and Heine-Geldern were the first to use the term in describing the presence of bronze-age Dongson drums throughout the region, or similarities in temple architecture. French scholarship of the 1920s and 1930s began to uncover the relations between the ancient kingdoms of Angkor, Champa, Pagan, and Java. George Coedès (1886–1969) first brought this material together in a path-breaking book he entitled *Histoire Ancienne des États Hindouisés d’Extrême-Orient* (Hanoi, 1944), though better known in its English translation (1968). Coedès broke with the ‘Greater India’ school which regarded the Southeast Asian kingdoms as Indian colonies, but continued to see Indian agency through visiting traders and scholars. Later historians such as Paul Mus, J. C. van Leur, Bernard Groslier, Ian Mabbett, Oliver

Wolters, and Claude Jacques, and still more the nationalists, were inclined to see Southeast Asians borrowing selectively and creatively what they found useful in Indian civilization, and even sharing in the creation of that civilization.

Histories of the whole region began to appear in English in the 1950s, the most influential being that of D. G. E. Hall (1955), who came to the first British chair of Southeast Asian History, that of the school of African and Oriental Studies (SOAS) in 1949, after an early career (1921–1934) at the University of Rangoon. The other colonial seedbed of historians of the region was the University of Malaya, opened in Singapore in 1949 and eventually spawning a range of institutions in independent Malaysia and Singapore. Such regional historians as Northcote Parkinson, C. D. Cowan, John Bastin, Wang Gungwu, Alistair Lamb, David Bassett, Jan Pluvier, Anthony Reid, Leonard Andaya, and Barbara Andaya learned their craft in these institutions.

The US borrowed from the British tradition initially, with first D. G. E. Hall and then Oliver Wolters pioneering the teaching of the subject at Cornell University. The first generation of American-trained historians, generally studying the dynamic of diverse nationalisms, were less inclined to perceive the region as a whole. The most important US history of the region to date, *In Search of Southeast Asia* (Sleinberg 1971), required the collaboration of six country specialists and is basically divided by country. In contrast, the *Cambridge History of Southeast Asia*, (Tarling 1992) took the heroic line that each chapter should be chronologically and thematically demarcated, but cover the whole region. This ambitious experiment was dominated by historians based in Australia, New Zealand, and Southeast Asia, reflecting a shift which had occurred in the previous two decades.

The approach of Fernand Braudel and the Annales School was brought to Southeast Asia in the 1980s in relation to the Early Modern, a period which had until then been largely abandoned to the 'European expansion' school. Denys Lombard (1936–1998), a colleague of Braudel and founder of the interdisciplinary journal *Archipel*, developed in *Le Carrefour Javanais: Essai d'Histoire Globale* (3 Vols., 1990) the kind of 'total history' pioneered by Braudel. Anthony Reid (b. 1939) found both the broad interdisciplinary and the ocean-centered focus of Braudel useful in giving coherence to his *Southeast Asia in the Age of Commerce, c.1450–1680* (2 Vols., 1988–1993).

In global terms, the practice of history in Southeast Asia since 1950 has been relatively weak in economic and institutional history, and has produced few outstanding biographies or intellectual histories in the strict sense. Its strengths have been its interdisciplinarity and its deep concern to explicate cultural diversity and autonomy. The two Southeast Asianists who have most influenced historians were trained in other disciplines. The cultural anthropologist Clifford

Geertz (b. 1926) defined his task as 'to uncover the conceptual structures that inform our subjects' acts,' and likened it to 'trying to read (in the sense of 'construct a reading of') a manuscript—foreign, faded, full of ellipses, incoherencies, suspicious emendations, and tendentious commentaries.' Though history is never absent from his numerous studies of Java and Bali, it is most salient in *Negara: The Theatre State in Nineteenth Century Bali* (1980), which influentially argued that elaborate royal rituals were the defining acts of Bali states, not simply their decoration or arcane cosmology.

Anglo-Irish Benedict Anderson (b. 1936) made his career at Cornell University, where his brand of culturally-informed political history had more influence on historians than on his own discipline of political science. His pathbreaking *Imagined Communities: Reflections on the Origins and Spread of Nationalism* (1983) was preceded by a series of more specific studies on Javanese and Indonesian political culture, and the influence of the Japanese occupation on them. In the 1980s, deeply alienated from Suharto's Indonesia, he produced revisionist essays which placed the histories of Thailand and the Philippines in a broader political economy perspective. Among the able historians trained at Cornell, Michigan and elsewhere were a number of Southeast Asians who had an impact in their own countries as well as internationally. Reynaldo Ileto (b. 1946) pioneered a new kind of history from below by beautifully demonstrating how popular Christianity provided the idiom of resistance in his *Pasyon and Revolution: Popular Movements in the Philippines* (1979). Nidhi Aevriongse wrote almost entirely in Thai after returning to teach in Chiang Mai, but his revisionism has shifted Thai historiography away from the agency of kings and Westerners to the social basis of cultural change.

See also: Area and International Studies: Development in Southeast Asia; Historiography and Historical Thought: Current Trends; Historiography and Historical Thought: East Asia; Historiography and Historical Thought: Modern History (Since the Eighteenth Century); Islam: Asia; South Asian Studies: Culture; South Asian Studies: Politics; Southeast Asia, Archaeology of; Southeast Asia: Sociocultural Aspects

Bibliography

- Coedès G 1968 *The Indianized States of Southeast Asia*. University of Hawaii Press, Honolulu, HI
Cowan C D, Wolters O W (eds.) 1976 *Southeast Asian History and Historiography: Essays Presented to D. G. E. Hall*. Cornell University Press, Ithaca, NY
Hall D G E 1955 *A History of South-East Asia*. Macmillan, London

- Hall D G E (ed.) 1961 *Historians of South East Asia*. Oxford University Press, London
- Reid A, Marr D (eds.) 1979 *Perceptions of the Past in Southeast Asia*. Asian Studies Association of Australia/Heinemann, Singapore
- Steinberg D J (ed.) 1971 *In Search of Southeast Asia: A Modern History*. Praeger, New York
- Tarling N (ed.) 1992 *The Cambridge History of Southeast Asia*, 2 vols. Cambridge University Press, Cambridge, UK
- Woolf D R (ed.) 1998 *A Global Encyclopedia of Historical Writing*. Garland, New York/London, 2 Vols.
- Wyatt D K, Woodside A (eds.) 1982 *Moral Order and the Question of Change: Essays on Southeast Asian Thought*. Yale University Southeast Asia Studies, CT

A. Reid

Historiography and Historical Thought: Sub-Saharan Africa

The discovery of Africa as an object of scholarly research is part of a wider expansion of academic pursuit. It coincided with the emergence of women's studies. The focus of investigation shifted to the underdogs, including slaves, colonized nations, and other silent majorities of our day as well as the Middle Ages, in Africa, and elsewhere (Wachtel 1971, Clancy 1979, Ginzburg 1980, Feierman 1993).

This article concentrates on sub-Saharan Africa. It covers recent origins of written history, the 'discovery' and valuation of its oral traditions, the interaction of anthropology and history, the Anglo-American debate, and the role of the diaspora.

1. Origins of History

The first Sahel states emerged on the northern fringe of the savanna: Ghana (eighth to the eleventh century), Mali (thirteenth to fourteenth century), and Songhay (twelfth to sixteenth century).

Arab authors were the most knowledgeable on Africa's past. Those who made the largest contributions to reconstructing the history of Africa, and western and central Africa in particular, include al-Bakri (1029–1094), al-Idrisi (1154), al'Umari (1301–1349), and Hassan Ibn Mohammed al-Wuzza'n, known as Leo the African (1494–1552). Ibn Khaldun (1332–1406) is unquestionably the 'Father of History,' even though Ibn Battuta (1304–1369) remains the best thanks to his personal accounts and the detail in his descriptions (Levtzion and Hopkins 1981).

Written European sources followed, beginning in the fourteenth century. Their authors were Portuguese navigators. These documents preceded the substantial

addition made by codification or oral records in Arabic by Mahmoûd Kâti and Es-Sa'di, both Timbuktu scholars.

In the course of the nineteenth and twentieth centuries, accounts written by Africans in indigenous languages have abounded. A large share of these histories was in Kiswahili, Hausa, Fulfulde, Kanembu, Diula, and Malagasy.

The first professional historians, caught in the Hegelian trap, discounted the history of Africa (Neale 1986, Jewsiewicki and Mudimbe 1993, p. 1). However, the continent and its 'people with no history' were deemed worthy of a present. Ethnography was called upon (Jewsiewicki and Mudimbe 1993, p. 8) to justify the backwardness of the region. Underdeveloped Africa was expected to make do with a 'prehistory' (Moniot 1991, p. 197).

The International African Institute was established in London in 1928. The School of Oriental and African Studies began in 1948 to teach a new subject called history of Africa (Vansina 1992, p. 77). An overdue offspring of academic research was born. Curiously, African universities in French colonies, the local intelligentsia, and their backers in France did not see any need for the discipline (Vansina 1992, Moniot 1995, p. 647).

2. History and Anthropology

Almost eight years after publication of *African Historiographies* (Jewsiewicki and Newbury 1986), Jewsiewicki and Mudimbe published *History Making in Africa* (1993). The majority of the authors who contributed to the former work were historians. Anthropologists dominated the latter. This shift in the weight accorded to the two disciplines is a sign of the times. When focused on Africa, anthropology and history do not fit neatly into the university compartmentalization scheme. Rebellious grassroots reality has gradually overcome the rigidity that had governed the allocation of professorships.

As time passed most of the convenient cubbyholes have been banned to the realm of memory. It was once the fashion to contrast societies with no history, acephalous, precapitalist societies, with societies blessed with a state and a history culminating in capitalism. The first group was viewed as inherently handicapped by communalism, polytheism, oralism, and tribalism. The second group of nations was endowed with the virtues of individualism, monotheism, literacy, and polity, to name just a few. Criticism of the concepts of tribe and ethnicity took two different approaches, a French school and an Anglo-Saxon school (Coquerel 1983, Young 1986, Chrétien and Prunier 1989, Barber and Moraes Farias 1990, Coplan 1991, 1993, van Jaarsveld 1992, Harries 1993, Jenkins 1997).

The adoption of precolonial Africa as a concept for a catchall analysis of the continent marked a milestone. This approach created the illusion of a different, idealized world. Postcolonial Africa was viewed as the child of a past to which the umbilical cord had been severed. This is the country cherished by the nationalists during their struggle for independence and the image invoked by politicians. Cheikh Anta Diop was fond of climbing the continent's family tree back to its glorious Egyptian ancestors, while Joseph Ki-Zerbo (1972) and J. F. Ade Ajayi convey us to a not so distant past, labeled the precolonial era (Mudimbe and Jewsiewicki 1993, pp. 1–2). Historians responsible for this breakdown into periods have changed their minds and now support the anthropologists' view of the existence of a continuum between the colonial societies and their worthy successor, sovereign Africa (Bayart 1989, Paillet 1987).

The focus of historians on the past and the concentration of ethnologists on the present explain the farsightedness of the former and the shortsightedness of the latter. The outlook of these two isolated disciplines changes radically as soon as historians discover humanity as the center of their field of investigation and anthropologists take account of time, the framework for human behavior, and activity.

3. *The Silent Make Themselves Heard*

Africa is known for its chanted, sung, or simply recounted oral history. The *griots*, professional chroniclers of West Africa, the *biiru*, specialists in Rwandan royal ritual, ordinary storytellers in the cities and the country have provided a wealth of material to researchers accustomed to studying written material.

In 1962, Jan Vansina published his authoritative work, *De la tradition orale*. His name is invariably associated with the acceptance of oral accounts as a source of African history and with the elaboration of a rigorous scientific method. It took historians, or at least part of their fraternity, a long time to learn to appreciate the 'auriture' of oral genres (Coplan 1991, 1993, p. 8).

Zealous pursuit of sources engenders such excesses as equating the death of an old timer with the gutting of a library. This attitude refers us to a kind of precolonial library, to paraphrase Mudimbe. It is also called the memory of the griots and 'oral archives' (Bazin 1979, p. 450). Writing history Western style entails the fault of reducing documents recorded by interviewers to simple sources. This is denounced by Jewsiewicki (1986, p. 16) and Schoenbrun (1993, p. 55).

An increasing hue and cry (Farias 1992) has been raised in opposition to this tendency. The critics are anxious to upgrade the role of reputedly passive traditionists (Ba 1980). While David W. Cohen (1989,

p. 12) insists on the contrast between material provided by the Africanist community and that produced by Africans who collect data on a daily basis, Jan Vansina was an early advocate of another reading of these documents, as revealed in a text of Jewsiewicki (1979, p. 81).

As soon as we escape the pull of the two fetishes constituted by the roving reporter and the Western-style historiographer, we are attracted by other types of history, and vice versa. We can then pay heed to nonformalized oral texts, unofficial samples taken from lower rungs of the social ladder. Their particular form and their non-verbal character require another method of analysis, yielding a more fertile approach to Africanist research (Olivier de Sardan 1976, Diawara 1990, Barber 1991). The field is further enriched by the involvement of anthropologists who tap the upstream end of oral production and characterize its interface with society (*surface sociale*). Their tools bridge a major gap neglected by historical studies which concentrate entirely on the oral text and lose sight of the life, environment, and experiences of participants in society (Moniot 1986, p. 58).

The history of the continent has become a professional endeavor. It involves 'interrogating and revealing the historian's operation' (Moniot 1995, p. 647). From now on, the Africanists can forget the 'automatic legitimization of being specialists in a small exotic field' (Cooper 1995, p. 236).

4. *The Anglo-American Debate*

Vansina (1992, p. 91) takes advantage of the lessons of historiography consisting in recognizing the past in the present and vice versa, and aims his harshest criticism at the postmodern drift of David W. Cohen. The latter (1994, p. XV) stresses the fact that historical knowledge is not the monopoly of poetry or formalized narratives but that it is constantly being proclaimed and cited. Henige (1995, p. 314) vehemently attacks what he calls Cohen's 'predilection [...] for subordinating the substance of sources to their interpretation'.

5. *The Diaspora Includes Africa; in Fact, it Is Africa*

What position do African specialists take in this controversy, which almost boils down to an 'American-American' debate? Jewsiewicki (1986, p. 9) makes a statement that has lost none of its pertinence for historians and other Africanists. 'The Africanist historian and the African historian share neither the same responsibilities nor the same existential constraints' (cf. Vansina 1994, p. 221). This is certainly true, except that another schism has split the African scholars' camp, divided between a diaspora and a home com-

munity, and occasionally bridged by a coming and going between the two shores. Africans abroad still focus more on debates that are often kindled overseas, in America or Europe.

Nevertheless, practically everything written is in the language of academics, which is not the idiom of indigenous populations. A UNESCO policy to remedy this situation, adopted in the 1970s, remains a pious vow even though issues of *l'Histoire Générale de l'Afrique* have appeared in Kiswahili and Hausa. Anyway, who is able to read them? Where can you get a copy and what would it cost? Does the body of literature constitute history for the average person? The question posed by the *African Historiographies* subtitle, 'What history for which Africa?' is as relevant as ever. The role of Africans in this international forum appears to be rather marginal. The empire of Western historiography is bigger than ever (Vansina 1994, p. 220). The hopes raised by studies in the English-speaking world, which benefits from the North American alternative to the intellectual heirs of the British Empire and from a more substantial organization (Jewsiewicki 1986, p. 15), continue to fade. In the meantime, the New World option has become an attractive outlet for the brain drain from tropical dictatorships. Zaire has forfeited its chances to develop a school of thought liberated from the domination of Brussels. The ex-colonies reproduce the models elaborated in the former centers of colonial power (Chrétien 1997) despite the fact that English-speaking and French-speaking African authors have begun to read one another's works (Coquery-Vidrovitch 1997). Their marginality becomes even more pronounced as the tendency to consume (È Nzièm 1986, p. 25) concepts developed outside Africa grows. It makes little difference whether they are the products of non-Africans or of Africans of the diaspora, which never stops expanding (Jewsiewicki 1979, p. 81).

Intellectual dependence resembling the economic impasse of the continent is worsening. While additions to the circle of recognized Africanists are becoming rare, a heavy storm of productivity is brewing in Africa, as far as reading about and interpreting the past are concerned. Student theses, cassettes recorded by individuals in different villages, radio stations, and other relevant institutions are mushrooming. The harvest has never been so bountiful (Cohen 1994, p. XV, Diawara 1997).

This analysis is not intended to deny the essential, unanticipated impact of African studies on social sciences in general. The 'dissolution of world history' by the history of Africa is a reality (Feierman 1993). In evaluating African studies, one is inclined to dismiss the crucial role played by the diaspora, which actively propagates its share of the African experience (Macamo 1999). Expatriates are deprived of their birthright by attributing any success they achieve to their severance of ties to their dismal native lands in the tropics. Mudimbe and Appiah (1993, p. 113),

referring to the impact of African studies on American disciplines, cite the necessity 'to locate ourselves as much in the current history of those disciplines as in the current state of African Studies [...].'

In fact, a wider influence of African studies on intellectual life is detected—in Western Europe and elsewhere as well as in America (Feierman 1993, p. 200, note 13). The impact of Africanist historiography on that of France, the former métropole, is undeniable (Coquery-Vidrovitch and Jewsiewicki 1986, pp. 143–44).

A glimpse at what is on the horizon suggests the tide is turning, the groundswell stemming from the unprecedented creativity of local bards, both griots and other producers of historical material. This resurgence is amplified by a foreign legion, which gives a boost to those who have remained on their native soil. Hope is raised by the fact that disciplines applied to Africa are inevitably linked to academic trends in Europe and the US. No longer cramped by a narrow view of a particular field of study, Africanists have demonstrated the benefits of multidisciplinary collaboration. In addition, Africanists have been instrumental in fostering a dialog and creating an interface between traditional academic approaches in Anglo-Saxon and French-speaking countries.

See also: African Studies: History; Anthropology, History of; Historiography and Historical Thought: Christian Tradition; Near Middle East/North African Studies: Culture; Near Middle East/North African Studies: Society and History; Oral History; Urban Ethnography

Bibliography

- Ba A H 1980 La tradition vivante. In: *Histoire Générale de l'Afrique*. UNESCO, Vol. 1, pp. 191–230, Paris
- Barber K 1991 *I Could Speak Until Tomorrow. Oriki Women and the Past in a Yoruba Town*. Edinburgh University Press, Edinburgh, UK
- Barber K, Farias M, Fernando de P (eds.) 1990 *Self-Assertion and Brokerage. Early cultural Nationalism in West Africa*. Birmingham University African Studies, Series 2, Birmingham
- Bayart J-F 1989 *L'État en Afrique. La Politique du Ventre*. Fayard, Paris
- Bazin J 1979 La production d'un récit historique. *Cahiers d'Études Africaines* 73–76. 19(1–4): 435–83
- Chrétien J-P 1997 Une crise de l'histoire de l'Afrique en langue française. *Politique Africaine* 68: 141–49
- Chrétien J-P, Prunier G (eds.) 1989 *Les Ethnies ont une Histoire*. Karthala-ACCT, Paris
- Cissé Y T, Kamissoko W 1988 *La Grande Geste du Mali des Origines à la Fondation de l'Empire*. Karthala-Arsan, Paris
- Cissé Y T, Kamissoko W 1991 *Soundjata, la gloire du Mali*. Arsana Karthala, Paris
- Clancy M T 1979 *From Memory to Written Record. England 1066–1307*. Arnold, London
- Cohen D W 1989 The undefining of oral tradition. *Ethnohistory* 36: 9–17

- Cohen D W 1994 *The Combing of History*. University of Chicago Press, Chicago
- Cooper F 1995 Work, class and empire: An African historian's retrospective on E. P. Thompson. In: *Social History*, pp. 235–41
- Coplan D B 1991 Fictions that save: Migrants' performance and Basotho national culture. *Cultural Anthropology* 6: 164–91
- Coplan D B 1993 History in Eaten Whole: Consuming Trapes in Sesotho Auriture. In: Mudimbe V Y, Jewsiewicki B (eds.) *History and Theory*, Beiheft 32, Wesleyan University, Middletown, CT, pp. 80–104
- Coquerel P 1988 *Le verouillage ethnique en Afrique du Sud*. UNESCO-OUA
- Coquerel P 1989 Les Zulu dans l'Afrique du Sud contemporaine. In: Chrétien J-P, Prunier G (eds.) *Les Ethnies ont une Histoire*. Karthala-ACCT, Paris, pp. 417–25
- Coquery-Vidrovitch C 1997 Réflexion comparée sur l'historiographie africaniste de langue française et anglaise. *Politique Africaine* 66: 91–100
- Coquery-Vidrovitch C, Jewsiewicki B 1986 Africanist historiography in France and Belgium. Traditions and trends. In: Jewsiewicki B, Newbury D (eds.) *African Historiographies. What History for Which Africa?* Sage, London, pp. 139–50
- Cuoq J 1975 *Recueil des sources arabes concernant l'Afrique occidentale du VIIe au XVIe siècle (Bilād al-Sūdān)*. Editions du CNRS, Paris
- Diawara M 1989 Women, servitude and history: The oral historical traditions of women of servile condition in the kingdom of Jaara (Mali) from the fifteenth to the mid-nineteenth century. In: Barber K, Farias P F de Moraes (eds.) *Discourse and its Disguises. The Interpretation of African Oral Texts*, Centre of West African Studies, Birmingham, pp. 109–37
- Diawara M 1990 *La Graine de la Parole*. Franz Steiner Verlag, Stuttgart, Germany
- Diawara M 1994 Der Blick vom anderen Ufer. Oder: Die Entdeckung des Weissen. In: Schlumberger J A, Segl P (eds.) *Europa—aber was ist es? Aspekte seiner Identität in interdisziplinärer Sicht*. Böhlau, Cologne, Germany, pp. 255–83
- Diawara M 1997 Über die Bedeutung der Arbeit in der Welt der Mande. In: Spittler G, Beck K (eds.) *Arbeit in Afrika*, LIT, Munster, pp. 49–68
- Diop C A 1960 *L'Afrique Noire Précoloniale*. Présence Africaine, Paris
- Diop C A 1967 *Antériorité des Civilisations Nègres: Mythe ou Vérité Historique?* Présence Africaine, Paris
- Duby G, Lardreau G 1980 *Dialogues*. Flammarion, Paris
- È Nzièm N 1986 African historians and Africanist historians. In: Jewsiewicki B, Newbury D (eds.) *African Historiographies. What History for Which Africa?* Sage, London, pp. 20–7
- Farias P F M 1992 The oral traditionist as critic and intellectual producer: An example from contemporary Mali. In: Falola T (ed.) *African Historiography—Essays in Honour of Jacob Ade Ajayi*. Longman, London, pp. 20–44
- Feiermann S 1993 African histories and the dissolution of world history. In: Bates R H, M V Y, O'Barr J (eds.) *Africa and the Disciplines*. The University of Chicago Press, Chicago, pp. 167–212
- Ginzburg C 1980 *Cheese and the Worms. The Cosmos of a Sixteenth-century Miller*. Routledge & Kegan Paul, London
- Harries P 1993 Imagery, symbolism and tradition in a South African Bantustan: Mangusuthu Buthelezi, Inkatha, and Zulu history. In: Mudimbe V Y, Jewsiewicki B (eds.) *History Making in Africa, History and Theory*. Beiheft 32, Wesleyan University, Middletown, CT, pp. 105–25
- Henige D 1995 Homphaloskepsis and infantilizing of history. *Journal of African History* 36: 311–18
- Jaarsveld F A 1992 Recent Afrikaner historiography. *Itinerario* 16(1): 93–106
- Jenkins R 1997 *Rethinking Ethnicity: Arguments and Explorations*. Sage, London
- Jewsiewicki B 1979 L'histoire en Afrique ou le commerce des idées usagées. *Revue Canadienne des Études africaines*. 13(1–2): 69–87
- Jewsiewicki B 1986 Introduction. One historiography or several. A requiem for Africanism. In: Jewsiewicki B, Newbury D (eds.) *African Historiographies. What History for Which Africa?* Sage, London, pp. 9–17
- Jewsiewicki B 1991 Le primitivisme, le postcolonialisme, les antiquités 'nègres' et la question nationale. *Cahiers d'Études Africaines* 121–122, 31(1–2): 191–213
- Jewsiewicki B, Newbury D (eds.) 1986 *African Historiographies. What History for Which Africa?* Sage, London
- Ki-Zerbo J 1972 *Histoire de l'Afrique Noire: d'Hier à demain*. Hatier, Paris
- Levtzion N 1973 *Ancient Ghana and Mali*. Methuen, London
- Levtzion N, Hopkins J F P 1981 *Corpus of Early Arabic Sources for West African History*. Cambridge University Press, Cambridge, UK
- Macamo E S 1999 *Was ist Afrika? Zur Geschichte und Kultursoziologie eines modernen Konstrukts*. Duncker & Humblot, Berlin
- McIntosh R J, McIntosh S K 1979 Terra-cotta statuettes from Mali. *African Art* 12(2): pp. 52–3, 91
- Miller J 1980a Listening for the African past. In: Miller J (ed.) *The African Past Speaks*. Dawson-Archon, Folkstone, UK, pp. 1–59
- Miller J (ed.) 1980b *The African Past Speaks*. Dawson-Archon, Folkstone, UK
- Moniot H 1970 Les sources de l'histoire africaine. In: Deschamps H (ed.) *Histoire Générale de l'Afrique*. T I, PUF, Paris, pp. 123–47
- Moniot H 1986 Profile of a historiography. Oral tradition and historical research in Africa. In: Jewsiewicki B, Newbury D (eds.) *African Historiographies. What History for Which Africa?* Sage, London, pp. 50–8
- Moniot H 1991 L'histoire africaine, un observatoire de la periodisation *Périodes: la construction du temps historique. Actes du Ve colloque d'histoire du temps présent*. Éditions de l'École des Hautes Études en Sciences Sociales et Histoire au Présent, Paris
- Moniot H 1995 L'histoire à l'épreuve de l'Afrique. *Cahiers d'Études Africaines* 138–139, 35(2–3): 647–56
- Mudimbe V Y, Appiah K A 1993 The impact of African studies on philosophy. In: Bates R H, Mudimbe V Y, O'Barr J (eds.) *Africa and the Disciplines*. The University of Chicago Press, Chicago, pp. 113–35
- Mudimbe V Y, Jewsiewicki B 1993 Africans' memories and contemporary history of Africa. In: Mudimbe V Y, Jewsiewicki B (eds.) *History Making in Africa, History and Theory*. Beiheft 32, Wesleyan University, Middletown, CT, pp. 1–11
- Neale C 1986 The idea of progress in the revision of African History, 1960–1970. In: Jewsiewicki B, Newbury D (eds.) *African Historiographies. What History for Which Africa?* Sage, London, pp. 112–22
- Olivier de Sardan J-P 1976 *Quand nos Pères Etaient Captifs*, Paris

- Piault M (ed.) 1987 *La Colonisation, Rupture ou Parenthèse*. Paris
- Shoenbrun D L 1993 A past whose time has come: Historical context and history in Eastern Africa's Great Lakes. In: Mudimbe V Y, Jewswiecki B (eds.) *History Making in Africa, History and Theory*. Beiheft 32, Wesleyan University, Middletown, CT pp. 32–56
- Stock B 1983 *The Implication of Literacy*. Princeton University Press, Princeton, NJ
- Vansina J 1962 *De la Tradition Orale*. Musée Royal de l'Afrique Centrale, Annales, Sciences Humaines, No. 36, Tervuren
- Vansina J 1985 *Oral Tradition as History*. University of Wisconsin Press, Madison, WI
- Vansina J 1992 Some perceptions on the writing of African History: 1948–1992. *Itinerario* 16(1): 77–91
- Vansina J 1994 *Living with Africa*. University of Wisconsin Press, Madison, WI
- Wachtel N 1971 *La Vision des Vaincus*. Gallimard, Paris
- Young C M 1986 Nationalism, ethnicity and class in Africa: A retrospective. *Cahiers d'Études Africaines* 103: 421–95

M. Diawara

History and Literature

Literature and history have many things in common, and one of them is the aesthetic form of narration. The many contemporary experiments notwithstanding, modern historiography and the modern novel still share the common feature of narration. In narratology, some theoreticians have differentiated between the narrative modes in the two fields; others have shown how they conflate and that they can hardly be distinguished. This article wants to show what narration in literature and history has in common and in what respect the two modes differ from each other.

For Aristotle (see Chapter 9 of his *Poetics*) the matter of the division of labor between historiographer and poet was easily decided: the historiographer relates 'what has happened' and the poet 'what might happen'. In contrast to Plato's devastating criticism of literature, Aristotle considered poetry to be of high value. In his eyes, literature dealt with problems of a more general and thus more significant nature, while the historian had to report about specific, and thus, less important matters.

Until approximately 1750, historiography was considered a branch of literature. Since the eighteenth century (when Aristotle's distinctions were rediscovered), the differences between the two disciplines have been stressed, and it was only then that the term 'literature' became synonymous with 'poetry' and 'fiction.' Gottsched and Lessing still shared Aristotle's prejudice against history. For them the field of history was merely a quarry from which they mined some of the raw material they used to build their preconceived

literary plots. Of real importance to them was a moralistic story, and examples were taken from the repertoires in the areas of history, legend, and mythology. However, the late eighteenth century is an important turning point. As Koselleck has demonstrated, it is the time of the development of a genuinely new philosophical awareness about history (Koselleck 1976). The concept of a continuously progressing history as such replaces both the religious concept of a Christian teleology as well as the idea of history as a chaotic conglomerate of countless individual stories of an exemplary nature. Through this philosophical development, literature is seen as part of history, and thus it loses its Aristotelian superiority.

Today the relationship between fiction and historiography poses a number of hermeneutic questions, and there is no modern Aristotle who might offer a simple and acceptable solution fitting for our time. Narration can be viewed as one of the primary tools of knowledge, and it determines the structure and aesthetic form in the presentation of a real or a fictional event. Narration discloses the dimension of historical reality as well as the sphere of poetic fiction. The logic of narration provides access to the logic of historiography and fiction; it makes the structure of both historical consciousness and fantasy tangible. In every narrative, real or fictional events are supported by a structure through which happenings are accentuated, selected, or eliminated according to the law of relevance. Neither in historiographical nor in fictional narration are we dealing with a reproduction or duplication of events, but with a certain organization of actual or imaginary occurrences or experiences.

1. Narration in Historiography

Perhaps because it is so difficult to liberate it from its established forms, narration is constantly under attack. In the discussion of modern historiography from 1970–2000, it looked as if 'narration' would be replaced by 'description' and 'explanation.' 'Historiography,' Szondi argued, must 'divest itself increasingly of its narrative character'; narrative historiography must 'turn into description if a new concept of history is to do justice to our knowledge of history as an anonymous process, as a sequence of events and changing systems' (Szondi 1973). The arguments of the antinarrativists can be summarized in the following way: the narrative recording of events in historicism was based on the belief that history was made by individuals. However, this categorization did not adequately take into account the development of social processes. Processes and structures can only be described, not narrated. Thus the old '*histoire événementielle*' has to be replaced by the new '*histoire structurale*.' 'The nouvelle histoire' of French historians in particular, with their descriptions of structures

underlying long periods of time, and their quantifying methods, forced the old paradigm of writing history by narrating events into a defensive position (Furet 1982, LeGoff and Nora 1974).

However, according to some historiographers, history as a narration of events and history as a description of structures are not mutually exclusive. These historiographers concede that history as a narration of events is no longer the center of the historiographer's work and that it no longer provides the obvious support of historical writing. Nevertheless, they also concede that narration is allowed to stand as a possibility alongside the writing of history as a structural analysis of processes (Borst 1976). Other scholars will have nothing to do with a strict separation of the two methods and point out that narration and description are intertwined. The argument is that the fact that modern history is a series of processes, it can only be understood through the alternate explanation of events through structures and vice versa (Koselleck 1976 and Kocka 1984). Although the structures become more accessible through description and events through narration, nevertheless in practice it would be impossible to maintain a demarcation between the two methods. The complementary characters of description and narration are increasingly emphasized (Stone 1979).

The defenders of the descriptive method like to confront the narrators among the historiographers with their lack of theory. They assert that the historiographer must not only narrate but also explain, and that this is only possible by means of a theoretical concept of the realities, as it has long been practiced by social scientists. The proponents of the narrative method in writing history defend themselves vehemently against the charge of being anti-theoretical. From their initially defensive position, they have turned to attack. Danto has asserted that narration is vital for description and explanation; that the form of narration is at the same time a form of explanation. The narrative form always surpasses the given, and categorically uses all-encompassing concepts. Narrative structures influence our historical thinking in the same way theories do in the case of the sciences (Danto 1965). Ricoeur expresses the same idea in his phrase 'to tell a story means to explain it' (Ricoeur 1983). Rüsen agrees that narration is part of the logical structure of historiography and that consequently 'description and explanation have a function subordinate to the narrative scheme' (Baumgartner and Rüsen 1982).

2. Narration in Literature

In connection with Gottfried Benn's poetics, Grimm described a type of novel from which the narrator had disappeared. It was called 'the novel of the phenotype'

or 'the inward directed novel,' labels indicating a tendency toward lack of action and lack of characters (Grimm 1963). The world and its events are banned here. The reduction of the 'I,' the isolated subject, is significant for this novel: the 'I' speaks only of itself. Consequently, one reads in Rilke's *Aufzeichnungen des Malte Laurids Brigge* from 1910: 'It must have been before my time that stories were narrated, really narrated. I have never heard a person narrate a story.' Benjamin's discussion of the decline of storytelling in his essay *Der Erzähler* about the Russian writer Nikolai Lesskow echoes Rilke's statement (Benjamin 1972).

The aesthetic designation of the phenotype novel—exemplified in the novels of Virginia Woolf, Benn, and Rilke—is also appropriate for the writing of some contemporary authors; the authors of the French '*nouveau roman*' (like Alain Robbe-Grillet) as well as postmodern American writers (like William Gass) and contemporary German authors (like Peter Handke) come to mind. In the novel of the avant-garde (Joyce, Dos Passos, Gide, Döblin, Musil, Broch), narration was either complemented or largely replaced by montage techniques and cognitive means, by reflection or essayistic writing. Here it was not so much a matter of narration of events, action, and mimesis as one of the presentation of ideas and the provision of knowledge (Hinck 1995).

However, the best-known postwar novelists (like Garcia-Marquez, Vargas Llosa, Philip Roth, Bellow, or Grass) uphold traditional established narrative models. One can write purely narrative novels. However, nonnarrative epic fiction is nonexistent. Like historiography, the novel obviously cannot dispense entirely with narration, and thus it makes sense, now as ever, to adhere to the science of narrative technique (Lützel 1986). In the field of narrative technique, literary scholars are far ahead of the narratologists in the camp of the historiographers. Historians could learn from reader theories as developed by such literary theorists as Jonathan Culler ('the competent reader'), Umberto Eco ('the model reader'), Stanley Fish ('the informed reader'), Wolfgang Iser ('the implied reader'), Michel Riffaterre ('the superreader'), and Gerald Prince ('the narratee') (Culler 1975, Eco 1979, Fish 1970, Iser 1974, Riffaterre 1971, and Prince 1982).

3. The Conflation of Historiography and Fiction

In *Metahistory*, Hayden White draws interesting parallels between historiography and fiction. He strives for a poetics of historiography and attempts to describe the forms of historical presentation in order to compare them with literary forms (White 1973). White starts with the premise that no sequence of events—by itself—is a story; it is the task of the

historiographer to transfer the events, the facts, to the narrative framework of a story. The connections between happenings do not occur spontaneously; they are the result of the historiographer's reflection. In each historiographical text, White distinguishes between the 'what' (the narrated events) and the 'how' (the type of narration, the 'emplotment'). He was not the only scholar to see that historiographical and fictional narration are related, but he was the first to make a systematic attempt to study the structure of historical consciousness—to show the fictional elements and their formal manifestations in historiography.

The situation in nineteenth century historiography and philosophy of history serves as an example. White could have given his book the title *The Rhetoric of Historiography* as a parallel to Wayne C. Booth's 1961 study *The Rhetoric of Fiction*. In *Metahistory* he concludes that the writing of history has little to do with such qualities as true or false, correct or incorrect, and that ultimately historical writing cannot be distinguished from fiction. The historian's style, White argues, depends on the combination of the following elements: first, there are three different strategies; he calls them 'explanation by formal argument,' 'explanation by ideological implication,' and 'explanation by emplotment.' Specific modes of articulation correspond to each strategy. In this context we are only interested in the strategy of 'emplotment' (narration) with its four corresponding 'poetic' modes of 'romance,' 'comedy,' 'tragedy,' and 'satire.'

White's references to certain structural similarities between the unfolding of tragedies or comedies and that of historical representations are interesting, but these similarities should by no means be construed as an actual fictionalization of history. Demandt has shown how thought-structures were mirrored in the poetic metaphors used by historiographers, and how, in turn, the ideologies of the historiographers influenced their use of rhetorical figures (Demandt 1978). However, the use of metaphors does not signify a fictionalization of history. There exist basic narrative concepts that are equally valid for historical and for fictional narration. Without action, images, and metaphors, narration *per se* is unimaginable.

Returning to White's example of the different depictions and interpretations of the French Revolution, it is evident that this is not a case of fictionalization because the historical process is not depicted as an 'as if,' but as something that actually did occur. No one will deny that the depiction of history is a creative process. However, its qualitative difference from fictional narration remains, despite the use of literary interpretation patterns and flowery metaphors. While fiction has as its basis the so-called 'as-if' structure, and thus proves to be pretense, illusion, and unreality, historiography always relates to actual events. Hamburger clarified the difference through the example of a historical report about

Napoleon and a novel about the French emperor. The historiographer cannot see Napoleon 'in the subjectivity, the ego-originality of his inner thoughts, of his very "existence."' If he does so, however, we 'find ourselves in a novel about Napoleon, a piece of fiction' (Hamburger 1968).

The narrative techniques with which novelists depict the psyches of their characters have been analyzed by Dorrit Cohn. She differentiates between psycho-narration, quoted monologue, and narrated monologue (Cohn 1978). These are narrative techniques reserved for fictional narration, which cannot occur in historical writing. Just as the use of a literary (tragic or comic, for example) interpretive category cannot turn a historical narrative into a fictional one, a historical novel does not become a piece of historiography simply by including historical reality in the fictional field. Käte Hamburger validly defines the line between fiction and reality, 'where there is no crossing over from one category to the other,' when she discusses the different perceptions of mimesis in novel and historiography. According to her, there is a definite relationship between the narrating process and the narrated material in historiography, while characters created by novelists are simply narrated persons. In fiction, there is no relationship, only functional context, between the narrating process and the narrated material. In other words, the fictionalization of the persons described consists in the fact that the characters are not presented as objects but as subjects, as 'ego-origines' (Hamburger 1968).

This difference between historiography and fiction also becomes evident when we look at the results of more recent studies of the theory of functionality. Assmann has shown that scholars in the fields of analytic language philosophy and of pragmatic communication theory insist on this distinction (Assmann 1980). In the analytic philosophy of language the point is made that fiction has a very peculiar status as far as its logic of statements is concerned: the authors of fiction are not obliged to come up with any proof; their judgments are quasi, pseudo, or as-if statements. In communication theory, fictionality is studied with respect to its reception effects: it involves the elimination of the normal connection that exists between the participants of an act of communication. Dealing with fiction leads to a temporary liberation from thinking in terms of fixed institutions; it means the absence of social sanctions and a lack of verification.

4. The Concepts of Truth

Historia vero testis temporum, lux veritas, vita memoriae, magistra vitae, nuntia vetustatis: Cicero's classic definition of the function of historical writing is still valid today—with a few deviations and reservations. The modern formulations are somewhat less pre-

tentious, but given such terms as 'demand for truth,' 'imparting of meaning,' 'provision of identity,' and 'memory,' the function of historiography is not defined so differently from Cicero's wording. Only the postulate '*magistra vitae*' is viewed more skeptically. Authors of novels are equally concerned with the pursuit of truth, the imparting of meaning, the provision of identity, and memory. Friedrich Schlegel wrote in his introduction to the *Wissenschaft der europäischen Literatur*: 'As history aims at recognition and truth it approaches science, but as it is also representation (...) it is related to art.' Here the double-faced aspect of historiography has been precisely formulated.

No historiographer would be so arrogant as to assume that his or her report could encompass the whole truth of a historical event. Historiography can claim with justification that its statements are 'truthful' insofar as they have been gained 'scientifically,' that is in so far as they can be proven. And they can be proven because they can be proven wrong: the truth capacity of the narrative statements of historical writing is based on the fact that they can be falsified. In this respect, one can imagine no greater difference between historical writing and fiction.

Since the statements of a novel are fictional to begin with, they cannot be falsified. While writers of fiction may embellish an event from the past with fantasy, historiographers must check the sources, which dictate what they cannot say. The truth of a poetic statement is not achieved by fiction imitating the principles of historiography. The propensity for truth and truthfulness itself are presented differently in the two narrative modes. The intent to provide the truth does exist in both literature and historical writing, but it is aimed in different directions. Historiographers attempt to achieve the most exact comprehension and representation of actual historical events; novelists have no such ambition. On the contrary, their strength lies not in the realm of facts but in that of possibilities.

The aesthetics of the modern as well as of the postmodern novel are focused on the opening up of possibilities. Modernist literature aims to distance itself from the mimetic-realistic principle that dominated the novel of the second half of the nineteenth century. Musil, for example, in *Der Mann ohne Eigenschaften*, plays with several types of utopias and arrives at the category of 'possibility,' in opposition to confidence in the power of history. The same is true in the postmodern novels of Thomas Pynchon.

5. The Concept of Totality

In examining the truth postulate in literature and historiography, one must consider the premises on which historical and novelistic narration are based. In traditional narration the novelist, like the historiographer, is guided by ideas of totality as we know

them, on the one hand from the aesthetics of Hegel to Lukács, and on the other hand from Hegel's and Marx's substantialist philosophy of history. In both cases, the attempt is made to see events in such a way that they make sense within the context of a totality. Here the historical and artistic totalities are structurally identical. Danto has criticized as outdated the substantialist philosophy of history, with its concept of a teleological, continuous development. He states that this philosophy acts as if it knew of a 'divine plan' of history, as if one could determine past and present from the perspective of the future. Danto wants to replace the speculative and dogmatic substantialist philosophy of history with an analytical (descriptive and coordinating) theory of history. In this theory, the vantage point would not be from a postulated future goal back to the present and past, but rather from the past to the present, with no certainties about the future (Danto 1965). The ancient trust, reaffirmed by Nietzsche, in the continuity of history is questioned, as are the Christian and Marxist tenets of the eschatological or utopian direction of history.

Such beliefs once formed the foundation for the concept of continuity in history. Many contemporary historians shun these labels and are consequently faced with two choices: giving up the concept of continuity altogether, or adopting it from the tradition of historiography and perhaps developing it in their own fashion. Historians at present are more aware than ever of the heuristic character of the concept of continuity; they see it merely as a 'regulative idea' in historical narration (Rüsen 1982). Today the focus is more on discontinuity. In this respect the impact of Foucault as an outright anti-Hegelian thinker (especially with his 1969 study *L'archéologie du savoir*) is not to be underrated.

The dismissal of the concepts of continuity and totality has also had an impact on the writing of the novel. The idea of the 'self-contained' work of art that provides totality by way of typical characters is dead. Its place has been taken by the 'open' novel with its open-endedness, as first envisioned by the early German Romantics. In his *Theorie des Romans* (1916), Georg Lukács declared that the modern novelist could outdo the historiographer if, after the loss of totality of meaning and being, he or she would counter the partial insights into the course of history with concern for a new, homogeneous cosmology. In contrast, however, the modern novel plays an important role as an instrument of cognition because such concepts of systems are now distrusted. Not a new cosmology of being, but a view of the disparity of what is to be, of the diversity of possibilities, is what this medium can provide.

In twentieth-century philosophy, a similar discussion to the one between Plato and Aristotle about the truth-value of fiction took place. Adorno was convinced that, due to its autonomy, art 'as a creative, nonalienated work' was the last vestige of freedom.

According to him, art and literature alone can unveil what is hidden in reality and can destroy ideology and lies (Adorno 1970). What Adorno criticized as a web of ideological deceptions is close to what Roland Barthes calls the 'mythicizing' of everyday life. Like Adorno, Barthes affirms that poetic language maintains direct contact with reality and, unlike historiography, escapes mystification (Barthes 1957). In contrast to Adorno and Barthes, Eagleton and Jameson argue that the writing of literature is just another form of ideological production (Eagleton 1976). Jameson maintains the following position: 'Ideology is not something which informs or invests symbolic production; rather the aesthetic act is itself ideological, and the production of aesthetic or narrative form is to be seen as an ideological act in its own right' (Jameson 1981).

All four authors make extremely one-sided statements. The common ground of historiography and literature is narration. And when narration, through the eyes of Jameson, is seen as an ideological act per se, we cannot expect much truth in either historiography or fiction. But if, on the other hand, literature is able to destroy ideology and deception—to cite Adorno—this virtue should also be attributed to historiography. Narration can lead in both directions: it can help to approach truth, or it can produce ideological veils. Neither literature nor historiography is immune to the virus of error or the distortion of truth, but both are capable of making reality be seen more clearly.

6. Meaning, Memory, and Identity

Among the functions of historiography and the novel are, along with the search for truth, the imparting of meaning and the provision of identity. The imparting of meaning (*Sinnbildung*) is far from the creation of meaning (*Sinnstiftung*) (Rüsen 1982). Surely there were, and still are, historiographers and novelists who have declared themselves ideological creators of meaning, but neither modern novelists nor critical historiographers are interested in such ambitious goals. In historical narration, the imparting of meaning is achieved through memory alone. And although novelistic narration cannot be reduced to memory, it plays an important role here as well. This is especially evident when historians fail to bring to mind certain aspects of historical experience through memory; then it is often the novelist who must assume this task. The 1950s in West Germany provide a clear example of this phenomenon. With regard to the National Socialist past, repression rather than grief was prevalent. It was a novel that broke this taboo: *Die Blechtrommel*, by Günter Grass (1959). This literary breakthrough was followed by the works of the historiographers. George Steiner listed a number of examples of this function of the 'writer as remembrancer' (Steiner 1973). It is

particularly in cases of the conscious falsification of history by historians that the writer steps in.

Writers and historiographers can help define the identity of a generation, a nation, or a culture. Since the Enlightenment and secularization, historiography and fiction have been burdened increasingly with the function of providing an identity. By dealing with memories, the experiences of the present attain a coherent perspective that provides the reader with an orientation for the future. Repeatedly correlating experiences of the present with those of the past can bring about the dissolution of stagnant patterns of identity, with a critical view of ideological prejudices (Lützel 1998). The factual orientation of the historiographer as well as the novelist's interest in what is possible are part of the provision of identity. By means of memory, a potential of historical and literary meaning is activated that, along with the demonstration of actual limits and the allusion to yet untested possibilities, makes the provision of identity possible. Historical and literary consciousness is a complex construct in which memory is the prerequisite for interpretation and expectation. To structure an identity means, first, to identify with stories. During an identity crisis, society may show greater interest in historical remembering (as was the case in Europe in the nineteenth century), or demand a fictional provision of identity (as is presently the case in Latin America). The weaker the link with the historical past and the more suspect it appears, the stronger will be the desire for fictional patterns of identity.

See also: History and Memory; History, Theories and Methods; Anthropological Universals

Bibliography

- Adorno T W 1970 *Ästhetische Theorie*. Suhrkamp, Frankfurt, Germany
- Assmann A 1980 *Die Legitimität der Fiktion*. Fink, Munich, Germany
- Barthes R 1957 *Mythologies*. Editions du Seuil, Paris
- Baumgartner H M, Rüsen J 1982 Erträge der Diskussion. In: Lämmert E (ed.) *Erzählforschung. Ein Symposium*. Metzler, Stuttgart, Germany
- Benjamin W 1972 *Gesammelte Schriften* 2. Suhrkamp, Frankfurt, Germany
- Borst A 1976 Das historische Ereignis. In: Baumgartner H M, Rüsen J (eds.) *Seminar: Geschichte und Theorie: Umriss Historik*. Suhrkamp, Frankfurt, Germany
- Cohn D 1978 *Transparent Minds: Narrative Modes for Presenting Consciousness in Fiction*. Princeton University Press, Princeton, NJ
- Culler J 1975 *Structuralist Poetics. Structuralism, Linguistics and the Study of Literature*. Routledge and Kegan Paul, London
- Danto A C 1965 *Analytical Philosophy of History*. Cambridge University Press, Cambridge, UK
- Demandt A 1978 *Metaphern für Geschichte*. Beck, Munich, Germany

- Eagleton T 1976 *Criticism and Ideology. A Study in Marxist Literary Theory*. Humanities Press, Atlantic Highlands, NJ
- Eco U 1979 *The Role of the Reader*. Indiana University Press, Bloomington, IN
- Fish S 1970 Literature in the reader: Affective stylistics. *New Literary History* 2: 123–62
- Furet F 1982 *L'atelier de l'histoire*. Flammarion, Paris
- Grimm R 1963 *Strukturen; Essays zur Deutschen Literatur*. Sachse und Pohl, Göttingen, Germany
- Hamburger K 1968 *Die Logik der Dichtung*. E Klett, Stuttgart, Germany
- Hinck W 1995 *Geschichtsdichtung*. Vandenhoeck & Ruprecht, Göttingen, Germany
- Iser W 1974 *The Implied Reader; Patterns of Communication in Prose Fiction from Bunyan to Beckett*. Johns Hopkins University Press, Baltimore, MD
- Jameson F 1981 *The Political Unconscious: Narrative as a Socially Symbolic Act*. Cornell University Press, Ithaca, NY
- Kocka J 1984 Zurück zur Erzählung? Plädoyer für historische Argumentation. *Geschichte und Gesellschaft* 10: 395–408
- Koselleck R 1976 Wozu noch Historie? In: Baumgartner H M, Rüsen J (eds.) *Seminar: Geschichte und Theorie*. Suhrkamp, Frankfurt, Germany
- LeGoff J, Nora P 1974 *Faire de l'histoire*. Gallimard, Paris
- Lützeler P M 1986 *Zeitgeschichte in Geschichten der Zeit*. Bouvier, Bonn, Germany
- Lützeler P M 1998 Nomadentum und Arbeitslosigkeit. Identität in der Postmoderne. *Merkur* 52: 908–18
- Prince G 1982 *Narratology: the Form and Functioning of Narrative*. Mouton, Berlin, Germany
- Ricoeur P 1983 *Temps et récit*. Seuil, Paris
- Riffaterre M 1971 *Essais de stylistique structurale*. Spanish, Seix Barral, Barcelona
- Rüsen J 1982 Die vier Typen des historischen Erzählens. In: Koselleck R (ed.) *Formen der Geschichtsschreibung*. Deutscher Taschenbuch Verlag, Munich, Germany
- Steiner G 1973 The writer as remembrancer: A note on poetics 9. *Yearbook of Comparative and General Literature* 22: 51–7
- Stone L 1979 The revival of narrative reflections on a new old history. *Past and Present* 85: 3–24
- Szondi P 1973 Für eine nicht mehr narrative Historie. In: Koselleck R, Stempel W D (eds.) *Geschichte—Ereignis und Erzählung*. Fink, Munich, Germany
- White H 1973 *Metahistory: The Historical Imagination in Nineteenth-century Europe*. The Johns Hopkins University Press, Baltimore, MD

P. M. Lützeler

History and Memory

1. The Alliance Between History and Memory

1.1 Remembered History

Before history is registered, chronicled, and interpreted by professional historians, it is witnessed and remembered by those that are contemporary to it. Such a memory of experienced history is necessarily

partial, biased, subjective, and, therefore, also highly variegated. While the present of experience is witnessed as a rather formless, dissonant, and even chaotic mass of perceptions, actions, and events, it assumes a much more unified shape and meaning when it is transformed into what is called the 'past.' To some extent, the multifarious, disrupted, and inchoate experiences of the former present are preserved in personal memories, if they manage to resist the tendency to adapt to the retrospective interpretation and normalization of the past. For many historians, there is very little overlapping between the present as experienced and witnessed and the past as reconstructed by historians (Koselleck 1997). While individual experience is particular, partial, and extremely limited in perspective, the reconstructed past provides a coherent and inclusive view of what was never in this way to be witnessed. In contrast to the past, the present does not produce a common reality that can be experienced by the witnesses in a similar way.

The plurality of lived experience and the unified reconstruction of history, however, are not fixed in a radical polarity. These poles are mediated in many ways, bringing the realms of present experience and retrospective history closer together. The personal experience of history is not only radically subjective, it is also shaped by the generalizing and collectivizing factors of social norms and cultural knowledge. Individual behavior and memory is not wholly erratic because it is stamped by generational patterns of experience, which means that it is already to some extent embedded in and framed by collective and cultural constructions of history.

It is true that personal memories of history differ as one individual life differs from another life. This, however, does not exclude certain similarities arising from biographical patterns. These life patterns with their respective perspectives, challenges, and responsibilities have their impact on the constitution of memories. The second world war, for instance, is remembered differently by the different nations as it is obviously remembered differently by the victims, the perpetrators, and the bystanders. But even among the Germans, it was remembered differently by different generations, by men and women, by those who remained in place and those who were forced to migrate from east to west. It is the task of the oral historian to acknowledge these individual and collectively typical memories as historical sources and to investigate history from below, collecting the various points of view of those who did not plan and control history but who were affected and afflicted by it.

There are other reasons why the singularity and interiority of memories should not be overestimated. Memories are generated in the individual very much like the linguistic capacity and in accord with it. They are elaborated and stabilized to the extent that they are transformed into narratives and exchanged in social

interaction. The memories of an individual are always supported and warranted by the memories of other individuals. Subjective memories are embedded in familial interaction and social communication. As individuals share the memories of their parents and grandparents, the range of remembered history is extended over three generations. Thus, an embodied and participatory historical memory of approximately a 100 years, is built up by oral transmission.

The three-generational oral memory is transient and perishes continually with every new cycle of oral memory that is built up. This continuous fading out of remembered history is a normal, natural, and necessary process of social forgetting which wipes the slate clean for following generations and their experiences. An exception to this rule occurs when the remembered history has an heroic or traumatic character. In cases of historical triumph or trauma, arts of memory are devised to stabilize the otherwise transient memory and anchor it in the consciousness of following generations (Giesen 2000). The survivors of the holocaust are bearers of such a memory which by a general cultural consensus is to be made permanent. The transformation of personal memories into cultural memory is achieved by media of memory such as writings, films, museums, monuments, and video testimonies.

In societies without writing, historical consciousness is divided into a recent and a distant past. The recent past is the three-generational oral memory which is transformed continually over time while the distant mythical past is a heroic memory which is preserved by epic poetry and other ritual and ceremonial arts of memory. This cultural heritage is re-enacted periodically and celebrated in public festivals by the community and carefully preserved and transmitted by bards and other memory specialists.

1.2 Commemorated History

Experiential subjective memory is embedded in collective memories of a widening scope: that of social interaction and that of cultural identity. The boundary lines between individual, collective, and cultural memory are not water-tight, but, as Ricoeur has put it: *'mémoire privée et publique se constituent simultanément selon le schéma d'une constitution mutuelle et croisée'* (Ricoeur 1998, p. 32). While private memory is informal and erratic, public memory shows a much higher degree of organization. It is mediated, ritualized, and based on social institutions. Other than remembering, the term 'commemorating' refers to an intentional, formalized, and collective action. Public memory is built on a collective will and communal actions. In the transition from private to public and cultural memory, the media play an important role. It is particularly writing that has revolutionized cultural memory. Following the anthropologist André Leroi-

Gourhan, Jacques Le Goff has described the history of cultural memory in five stages as the history of changing media from orality and writing to ever more mechanical technologies of storing information (Le Goff 1986). Not only did it vastly extend the range of remembered history, it also liberated the poets from their traditional office of 'remembrancers,' setting them free for other, more individual tasks. There are two kinds of media with which the historical memory of a group is stabilized and transformed into an extended and continuous past: material media and performative media. Among the material media are books, photographs, films, monuments, relics, and other material carriers of signs that retain traces of the past and can serve as documentary sources in the reconstruction of history. The performative media are symbolic forms of action, such as rites, festivals, and ceremonies linked to symbolic dates. Traditions and celebrations periodically renew and revitalize the memory of the past through acts of repetition and commemoration. While the material media store information for later use, they tend to generate specialized knowledge and to externalize the memory of a group. It is, therefore, the function of institutions of education and the performative media to check this tendency and to restore the mnemonic quality of historical knowledge, to bring it back to popular consciousness, and to reconnect it with the realm of lived experience. Usually, material as well as performative media are involved in acts of commemoration. Halfway in between material and performative media are the modern mass media such as newspapers, cinema, radio, and television, not to forget the large historical expositions that are becoming more and more important in shaping an image of the past and providing occasions that trigger collective memory through public debates.

In addition to their personal, local, and familial recollections, individuals also hold a share in that 'corporate memory' (Lewis 1975) which is constituted by public symbols and rites of commemoration. 'Commemorative ceremonies are distinguishable from all other rituals by the fact that they explicitly refer to prototypical persons and events, whether these are understood to have a historical or a mythological existence' (Connerton 1989, p. 61). The creation of a corporate memory was the aim of ancient societies and was revived by the modern nation-state and other politicized groups. A corporate memory is extremely limited in scope. Unlike books and libraries, it cannot contain many and diverse historical events, nor does it do justice to finer distinctions. It selects only very few items and reshapes these historical events for its purposes. The long chronicle of history has to be condensed into a few poignant moments that acquire a high symbolic value for the identity of a nation and state. Examples for history telescoped in national memory are the 4 July in the US or the 14 July in France. Such regularly recurring dates provide occa-

sions for speeches and celebrations which re-invigorate both the corporate memory and the collective identity. In Israel, the historical event of the fall of Massada became a national cult in the new state. For many centuries, Massada had been forgotten in the Jewish tradition and was restored by historical and archaeological scholarship only in the 1960s. For an historical event to become an item of public memory, the event has to be taken out of its context and to be transformed into a symbol. Among the first to emphasize this was the sociologist Maurice Halbwachs who wrote in his book on the social frames of remembering: 'In order for a person or historical event to enter social memory, it has to be transformed into an instruction, a concept, a symbol. It is invested with meaning, it becomes an element in the system of ideas of a society' (Halbwachs 1985a, pp. 389–390). In the case of Massada, the political message of the historical event was coined in a poem by a Palestinian Jew who in the 1920s had written the line: 'Massada shall not fall again.' The memory of Massada was restored from Hellenic sources that had not been part of the heritage of remembered Jewish tradition (Vidal-Naquet 1981). This memory was 40 years later corroborated by archaeological findings which in themselves, however, were rather scarce. Before these long forgotten external sources could become a cornerstone in the construction of a national myth, they had to be reanimated and kindled by the spark of historical imagination. The story of Massada became a central point of reference in Israeli corporate memory, because in the new state, a new future required a different past (Lewis 1975).

In antiquity, the commemoration of historical events took the form of religious feasts which were celebrated annually like the feast of Passover in which the Jews commemorated their liberation and exodus from Egypt. Such events were integrated into the religious calendar with the effect of turning history into myth. The English still celebrate annually the memory of the attempted catholic assault on their parliament of 1605, with the folkloristic rites of the Guy Fawkes-day starting as early as in 1606. Another way to commemorate historical events is to celebrate them according to the 'round' decimal dates of chronology. In contrast to annual feasts, these cyclical patterns seldom recur in a life-time and, therefore, lack the repetitive structure which is necessary for a continuous memory. This form of long term commemoration already goes back to antiquity. There is a stela of Ramses II commemorating the establishment of the cult of Seth in Avaris after 400 years. As a practice of communal commemoration, however, the decimal pattern was introduced only in the early modern era. One of the earliest examples is the commemoration of Luther's reformation in 1615, 100 years after Luther's actions of protest at the Augsburg Reichstag. This mode of recalling past events via anniversaries in the decimal system has become the

dominant mode of remembering history in the nineteenth and twentieth century. Examples are the 500th anniversary of the Turkish conquest of Constantinople which was celebrated in Turkey in 1953 or the 500th anniversary of Columbus' discovery of the new world celebrated in 1992. Such events do not recur within the lifetime of individuals. Therefore, it is rather a mode of spotlighting historical dates than a form of creating a continuous, communal memory.

1.3 Myth and History

Myth has been established by theoreticians of the nineteenth and twentieth centuries as the inveterate antagonist of history. Myth and history were considered as two incompatible modes of presenting the past, the one in a cyclical, the other in a linear time frame. In this polemical constellation, the term myth was also associated with fabrication and lie, while truth and objectivity were reserved for history as presented by critical historians. Beyond these well-known connotations of the cyclical and the untrustworthy, the term myth has in the meantime also acquired the connotation of 'meaningful history.' 'Myths create meaning through which the past is linked to the present in such a way that it can orient the future' (Münkler 1994, p. 21). This description is not far from Assmann's definition of myth as 'imagined and remembered history' (Assmann 1992, see also Kirsch 1998). Myths are conceived by him as a complex of icons, symbols, and narratives about the past that define the cultural identity of a community. Such a communally imagined and transmitted past provides for the group the basic values that it needs for an explanation of the present and an orientation for the future. It is exactly this symbolic use of the past for the construction of an identity that was suspended in the process of professional historical scholarship as developed in academic institutions in the nineteenth century. In this specific sense, the term 'myth' is used as an antonym to (modern) histor(iograph)y; while the former refers to the presence of the past, the latter insists on the pastness of the past, enforcing 'a liberation of the present from the past' (Arnaldi 1974). In the era of historiographic scholarship, however, remembered history is not the opposite of factual history, on which it has to rely as its basis and source of correction. If myth is defined as remembered history, it is severed from other connotations such as 'lie,' 'imposture,' or 'sacred history.' It refers to 'a foundational narrative with the aim to illuminate the present by an account of origins. It is through remembering that history turns into myth. This does not mean that it becomes unreal. On the contrary, it is by turning history into myth that history becomes reality and assumes the quality of a normative and formative power' (Assmann 1992, p. 52).

Myth in this new sense of remembered history stresses mnemonic qualities such as symbolic power, narrative form, iconic impact, and a highly focussed and sensual form of relevance. Invested with a memorable shape, symbolic meaning, and pragmatic relevance, myth provides qualities that professional history had explicitly discarded. With a developed sense of the shortcomings of both myth and history, they are no longer confronted as antagonistic and mutually exclusive but rather seen as complementary modes of referring to the past. The mnemonic and symbolic use of the past is no longer perceived as an abuse but as an important social function which cannot be provided by scholarly historiography.

2. Memory and Historiography

2.1 The Identity of Memory and History

The relation between historiography and memory has itself a history which has evolved over various stages. Before the foundations of critical historical scholarship were laid in the eighteenth and nineteenth centuries, history, memory, and myth were not yet clearly distinguished. On the contrary, it was considered the central function of the writing of history to preserve the memory of a dynasty, the church, or a state in order to legitimize such institutions and to ensure their continuity by providing for them an honorable past. In this first stage, historiography was the foundation of the community and the 'handmaid of authority' (Plumb 1969). The *res gestae* of heroes and kings were sung by bards and written down by chroniclers in order to rescue them from oblivion and to establish an honorable memory for the noble dead.

Although a consciousness of a specific function of historiography was established in the writings of Herodotus, he emphasized in the proemium to his work the traditional link between history and memory. While giving up a strictly ethnocentric stance, history, for him, is still 'an art of memory' where 'memory' is closely related to 'fame.' His project claims to ensure 'that the actions of people shall not fade with time' and that 'great and admirable monuments produced by both Greeks and barbarians shall not go unrenowned' (Herodotus 1992, pp. 3–4, see also Hutton 1993). His historiographical project is aimed at fixing the events of the past in the memory of those to come. For Herodotus, the function of historiography was already twofold: it was divided into a memorial function and a cognitive function ('to set forth the reasons why they, i.e., the Greeks and the barbarians, waged war on each other'). The memorial function of historiography was never given up throughout Greek and Roman antiquity. Cicero still defined historiography as a weapon against oblivion, a formula that became a topos and was elaborated in the Renaissance. Official historiography of the Tudor dynasty in sixteenth century

England still counters oblivion which is personified as 'the ancard enemy,' 'the suckyng serpent,' 'the deadly darte,' and 'the defacer' (Campbell 1964).

The intimate link between memory and history shaped the historiographic genres in a specific way. It narrowed the criteria for the selection of persons and events to be memorized; only those of highest rank were singled out for a continuation in memory and only those feats and achievements were selected that contributed to the honor and fame of those that were remembered. In addition to this, only such events were selected that supported the opinions and interests of the ruling class.

The non distinction between memory and history is not only a feature of early stages of civilization but also a characteristic of authoritarian and totalitarian rule. But, whereas in the early stages neither media nor institutions had been established to back up an authority-independent account of the past, in later stages censorship had the function to destroy the rivaling media and carriers of counter histories threatening the security of the uniform view and authoritarian voice of history.

2.2 The Polarization of Memory and History

2.2.1 *The discrediting of memory by historians.* The second stage is marked by a separation between memory and histor(iograph)y. This distinction was established when professional historiography arose as an independent and specialized institution. It was not before the nineteenth century that historiography turned into an autonomous discursive discipline which set up its own standards of truth and authority, developing specific rules of verification and inter-subjective argument. The development of modern historiography as a cognitive enterprise has its roots in classical antiquity and was revived in early modernity. Histor(iograph)y started from memory and it reached its new status in emancipation from memory. Before the rise of critical historiography, history had been adapted to the demands of the present; it served specific functions for the state or community such as justifying the present, controlling the future and legitimizing authority. A critique of memory, therefore, was also a critique of power and authority. Critical historians insist on their distance to the institutions of corporate memory. Lukian called them 'apolis,' thereby indicating that they are not committed to the norms and ideals of a particular community. In spite of their ideal of objectivity, of course, even professional historians are not completely free from loyalties and prejudices. But the professional historian differs from others in that 'he is aware of this fact, and instead of indulging his prejudices seeks to identify and correct them' (Lewis 1975, p. 52). The criticism of historical memory not

only aims at filling its gaps and correcting its errors but also to expose the intentional fallacies and forgeries of memory, the constructions that secure the establishment and maintenance of power. Critical historiography sharpened its tools in the period of Renaissance humanism in the struggles for power and truth, but it developed its ideal in disinterested objectivity only in the nineteenth century. It was in this period that the links between history and identity were severed and historical knowledge turned into a rapidly growing mass of abstract, disconnected, and disembodied information.

2.2.2 The rehabilitation of memory by philosophers and historians. It was Nietzsche who saw a connection between the massive production of abstract historical knowledge and a growing apathy of cultural memory, which led him to express an urging demand for new and invigorated identities (Nietzsche 1988). According to Nietzsche, the historical disciplines had become a dangerous burden for the wellbeing of society. Scholarship produced a pathological form of memory without the possibility of forgetting, an ever growing archive of irrelevant data. Every group and collective unit, however, was in need of a clear cut and limited knowledge of the past, i.e., of a usable past. Nietzsche diagnosed for his age a severe crisis and pathology of cultural memory which had lost its capacity of limitation. For him, the important function of memory was not the storing of knowledge but the limitation of knowledge, the creation of profile and relevance in an indifferent mass of information. Historical knowledge was no longer a tool for the creation of a meaningful life but an oppressive burden that crushed identity and extinguished the sparks of creative energy.

Nietzsche's polarization of history and memory became a topos among historians of the twentieth century. The specificity of histor(iography) was defined in contrast to the demands and mechanisms of memory. When Maurice Halbwachs defined the mechanisms of social memory in the 1920s, he introduced history as a negative foil for the process. Halbwachs investigated the function of shared memories for the cohesion of groups. He spoke of a collective memory and showed that the group was sustained and nurtured by common memories as these memories were in turn sustained by the group. This means that the stability of memories is linked to the stability of groups; when its members are disseminated, the memories are also dissolved. Erratic and dissonant memories had no place in Halbwachs's functionalist and constructivist theory of memory. Halbwachs compared historical discourse to an ocean which swallowed up all the partial narratives and memories (Halbwachs 1985b, pp. 72–3). For him, histor(iography) is the universal memory of humanity, while collective memories are embodied by specific groups and therefore always

partial and biased. In order to integrate these memories into the total view of history, they have to be severed from their carriers and their social milieus. With this transformation of memory into the abstract scheme of history goes the evaporation of live experience and meaning.

Sixty years after Halbwachs, the French historian Nora has revived the concept of collective memory and dispelled lingering critical doubts about its theoretical status. He has extended Halbwachs' notion of an oral memory shared by a social group to a cultural memory of symbols, values, rites, and local traditions that provide the cohesive cement of a society. These shared elements Nora called 'lieux de mémoire,' reinterpreting a term from the tradition of Roman mnemotechnics, the 'imagines' and 'loci' of the '*ars memorativa*.' Like Halbwachs, Nora stresses the nexus between memory and group identity, and like Halbwachs, he distinguished the symbols of collective memory from the scholarly discourse of historiography. For Nora as for Nietzsche, embodied collective memory is threatened to disappear under the burden of academic historiography. Nora defines memory and history as polar opposites. For him, the decisive distinction is that between a lived past which he calls memory and an abstract past which he calls history. Memory links the past to a group, it creates an emotional and symbolically charged bond between the past and a collective identity in the present, while history severs this bond; it is universal in character, therefore it belongs to everybody and nobody (Nora 1984/1992, pp. 23–43).

Whereas the pioneers of critical historiography discredited memory as a rival of history, theorists such as Nietzsche, Halbwachs, and Nora rehabilitated memory over against the ideal of objective and abstract histor(iography). They rediscovered the social functions of memory as an important and indispensable mode of assessing the past. In spite of their evaluations of memory and history, however, they share with the representatives of critical historiography the attitude of polarizing memory and history. In their different writings, this polarization takes the following form:

Memory

- is an embodied form of memory
- stresses differences and exists in the plural
- is linked to a carrier which may be an individual, a group or an institution
- bridges the past, present and future
- is highly selective, includes forgetting
- creates a profile of values supporting an identity and providing orientation for action.

History

- is a disembodied form of memory
- provides a universal frame and exists in the singular
- is disengaged from specific carriers

- separates the past from present and future
- develops an even and impartial attention
- is in search for truth, therefore values are suspended.

2.3 Mnemohistory—New Affinities Between Memory and History

The third stage in the relationship between memory and historiography is marked by a blurring of the distinction between history and memory. This, however, does not imply a simple return to the first stage. Despite postmodern qualms and revaluations, the self-imposed disciplinary regulations that were installed with the modern institutionalisation of critical historiography as a scholarly and intersubjective project may well be considered an irreversible evolutionary achievement. Historiography, writes Le Goff, 'must aim at objectivity and must be built on a belief in historical truth' (Le Goff 1986, foreword). If the third stage is not to be described in terms of a regression, it is perhaps more adequately described as a response to certain tendencies towards an extreme historical positivism inherent in the second stage. The professional historians as exclusive guardians of the past are challenged by those who invoke memory as alternative modes claiming an access to the past. The term 'history' is a collective singular; to this concept corresponded the idea of history as a universal memory of mankind. To say that history is universal memory may be another way of saying that it is indifferent to memory. The price that had to be paid for this generalized memory was a disregard of living and preserved memories.

The object of modern historiography was a past that was neatly separated from the present and the future. The historian Koselleck (1997) has coined the terms 'space of experience' (*Erfahrungsraum*) and 'horizon of expectation' (*Erwartungshorizont*) and stressed the point that in modernity, both dimensions have become radically separated. Along the same lines, history was defined by the philosopher Blumenberg as 'the separation of the space of experience from the horizon of expectation,' i.e., as a fundamental rupture between past and future (Blumenberg 1986, p. 61). From this point of view, memory was the other of history because collective memory tends to bridge past and future, creating a usable past for a changing present.

After the long period of polarizing history and memory, new affinities were discovered between both modes of rendering the past. They were due to a number of reasons: the discontents with positivist historiography and transcendental historiosophy, a reorientation in methodology, and the establishing of new themes for historical investigation. To begin with the discontents. In the 1980s and 1990s the universalist frame of academic historiography was severely shaken

when repressive and homogenizing ideologies such as communism and colonialism lost their power. Since then, many diverse memories of political groups have asserted their right. Historians were baffled by the enormous impact of these living memories that they had considered to be a negligible entity and which had reappeared with great aplomb at historical turning points. At such moments of sociopolitical reversal, social institutions as well as the cultural frame of perception and values changed. In the course of such changes, new memories emerged and old memories were seen in a different light. It is at such historical moments that the conflict between history and memory rises to consciousness and becomes a political issue. With the reconstruction of the political framework of society, there is a chance to narrate the past in a different way. Such a moment of retrospection is often a moment of revelation; then it suddenly becomes obvious that what had been presented and taken as objective history had only been a particular and biased form of memory. The experience of a fundamental change of values exposes the contingency of earlier accounts of the past. In such situations memory becomes self reflexive, developing a sense of its own constructedness.

Another way of approximating history and memory was to be observed in changes in the methodology of the discipline. The journal *History and Memory* which was founded in 1989 signals a shift in historical interest. Two years before that date, an important debate had taken place between the nestor of German *Zeitgeschichte*, Martin Broszat, and the Jewish Holocaust survivor and Holocaust scholar Saul Friedländer. The controversy which they discussed in their exchange of letters revolved around the key concepts of memory and history (Broszat and Friedländer 1988). Broszat argued for a form of presenting history that moves from the past into the present, thus adopting the perspective of the contemporary observers who experience it as an open process. Friedländer, on the other hand, argued for a presentation of history starting from its manifest results and moving backwards to the beginnings, thus adopting the perspective of the memory of the victims. For Dan Diner, Friedländer's co-editor in *History and Memory*, historiography is no longer an objective universalist discourse but one that is ethically committed to the participatory perspective of the victims. Thus, the two terms converge, history becoming a form of memory and vice versa. In his latest book on the persecution of the Jews in Nazi Germany, Friedländer develops a form of historiography that is not confined to the description of structures and processes but also seeks to give a voice to individual victims and agents thus translating the anonymous statistics of the holocaust back into the memories of those who participated in and suffered from history (Friedländer 1998). In books such as this one a new methodology of history is being developed that seeks

to enlarge the scope of historiography by integrating personal memories into historiography. Instead of being unified by the authorial voice of the interpreting historian, historiography thus becomes multivocal, mixing history with memories. In blending what is known with what is remembered, historical discourse is opened up to include contingencies and dissonances, forgoing coherence and closure in favor of marking what otherwise remains unrepresented and unrepresentable.

A third possibility of crossing memory and history lies in the discovery of new themes for historiography. Since the 1980s, historians have become more and more interested in modes of remembering as a form of social practice. What was known of the past in the past, which images of the past were constructed, which acts of commemoration were devised? In asking such questions, memory became itself the object of historiography. This form of historiography was favored by medievalists, who had already been used to investigating social forms of religious and secular memorialization (Oexle 1995). Such interests brought about a form of historical investigation which can be termed *mnemohistory*. *Mnemohistory* is what could also be called 'the reception history of history.' 'It concentrates exclusively on those aspects of significance and relevance which are the product of memory' (Assmann 1997, p. 9). To speak in this context of memory rather than of tradition is to acknowledge the constructive as well as the distorting power of memory. It takes into account the ambivalence of the past both as a conscious choice and as an unconscious burden, tracking the voluntary and involuntary paths of memory. 'The task of historical positivism consists in separating the historical from the mythical elements in memory and distinguishing the elements which retain the past from those which shape the present. In contrast, the task of *mnemohistory* consists in analyzing the mythical elements in tradition and discovering their hidden agenda' (Assmann 1997, p. 10).

An early and impressive project in *mnemohistory* was organized by the French historian Nora who published seven volumes on what he called the French 'lieux de mémoire' (Nora 1984/1992). Nora designed this historical enterprise which he carried out as team work with eminent historians in reaction to the publications of the *Annales* school of French historiography. In the scholarly works of this school, which had concentrated on structures, figures and processes in a *longue durée* perspective, history had more and more lost its meaning and memorability. The palpable dimension of history was restored by Nora who created an inventory of *mnemo-historical* items which included historical sites such as Vézelay, rites and songs of folklore, national symbols, commemoration dates and festivals, and more than 300 types of cheese. They are the 'lieux de mémoire' that, according to Nora, have constituted the French cultural memory and identity since 1789.

2.4 Memory as an Empirical Basis for Historiography

It is not easy to predict the next phase in the relationship between memory and history. It is possible that after their convergence, new efforts will be directed at their separation. There is also a new interest in memory from the point of view of a new kind of empiricism that is probing the criteria for historical truth. While one variation of *mnemohistory* tried to integrate personal memories into history and another sidelined factual history in order to focus on a remembered, reconstructed, or even repressed and forgotten past, there are some signs now of a new anthropological empiricism which turns away from memory as a social and cultural construct and concentrates on the neurological operations of remembering and forgetting. In an article with the title *The Veil of Memory*, the historian Johannes Fried noted 'a frightening backwardness among historians as far as questions of memory are concerned' (Fried 1997, p. 8). After the growing fascination and even obsession of historians with questions of memory since the 1980s, this statement can only make sense when it is applied to questions of memory from the point of view of the human brain and the neurosciences. He is interested not in cultural memory but 'the factuality of memory, its conditions and its effects on the historian's sources' because 'the neural connections or the mental processes determine what is preserved and remembered, and thus all knowledge and all history. The same applies to forgetting. The veil of memory, the river of oblivion is constantly in flow between the historian and its object' (Fried 1997, pp. 14, 16–17). This knowledge of the operations of memory also includes the mechanisms of distortion as explored in cultural studies and the neurosciences in the last ten years (Schacter 1999, Schudson 1992). For Fried, however, the crucial question for the historian to ask is no longer: how are texts connected? but: how does memory work? At this point, we return to the lived experience of individuals witnessing history as the starting point for historiography which was treated in the first paragraph of this article. Fried, however, no longer discusses memory in an existential or in a social but in a biological and anthropological perspective. He argues that everything that the historian deals with is primarily mediated by the human memory. Therefore, the historian must have recourse to anthropological empiricism in order to 'see through all this memory-activity' and probe the truthfulness of his one and only reliable source which is, after all, human experience.

See also: Collective Beliefs: Sociological Explanation; Collective Memory, Anthropology of; Collective Memory, Psychology of; Commemorative Objects; Historiography and Historical Thought: Current Trends; History and Literature; History: Forms of

Presentation, Discourses, and Functions; History: Overview; Identity: Social; Media and History: Cultural Concerns; Memory: Collaborative; Narrative, Sociology of; Oral History; Ritual; Ritual and Symbolism, Archaeology of

Bibliography

- Arnaldi G 1974 La storiografia come mezzo di liberazione dal passato. In: Cavazza F L, Graubard S R (eds.) *Il Caso Italiano*. Garzanti, Milan, pp. 553–62
- Assmann J 1992 *Das kulturelle Gedächtnis. Schrift, Erinnerung und politische Identität in frühen Hochkulturen*. Beck, Munich, Germany
- Assmann J 1997 *Moses the Egyptian. The Memory of Egypt in Western Monotheism*. Harvard University Press, Cambridge, MA
- Blumenberg H 1986 *Weltzeit und Lebenszeit*. Suhrkamp, Frankfurt, Germany
- Broszat M, Friedländer S 1988 Diskussion um die Historisierung des Nationalsozialismus. Ein Briefwechsel. [On the historicization of national socialism—an exchange of letters.] *Vierteljahrshefte für Zeitgeschichte* 36: 339–72
- Campbell L B 1964 *Shakespeare's Histories—Mirrors of Elizabethan Policy*. Methuen, London
- Connerton P 1989 *How Societies Remember*. Cambridge University Press, Cambridge, UK
- Fried J 1997 *The Veil of Memory. Anthropological Problems When Considering the Past*. The German Historical Institute London, 1997 Annual Lecture, London
- Friedländer S 1998 *Nazi Germany and the Jews* (Das Dritte Reich und die Juden, Vol. I: Die Jahre der Verfolgung 1933–9). Beck, Munich, Germany
- Giesen B 2000 *Triumph and Trauma*. University of California Press, Berkeley/Los Angeles, CA
- Halbwachs M 1985a *Das Gedächtnis und seine sozialen Bedingungen*. Suhrkamp, Frankfurt, Germany (French original: *Les Cadres Sociaux de la Mémoire*. Paris 1925)
- Halbwachs M 1985b *Das kollektive Gedächtnis*. Fischer, Frankfurt, Germany
- Herodotus 1992 *The Histories* (Blanco W (ed.)). Norton, New York/London
- Hutton P H 1993 *History as an Art of Memory*. University Press of New England, Hanover, UT
- Kirsch J H 1998 Mythos, Geschichte und Geschichtswissenschaft. In: Tepe P et al. (eds.) *Mythologica 6. Düsseldorfer Jahrbuch für interdisziplinäre Mythosforschung*. Verlag Die Blaue Eule, Essen, Germany, pp. 108–27
- Koselleck R 1997 Vom Sinn und Unsinn der Geschichte. In: Klaus E M, Rüsen J (eds.) *Historische Sinnbildung. Problemstellungen, Zeitkonzepte, Wahrnehmungshorizonte, Darstellungsstrategien*. Rowohlt, Reinbek, Germany, pp. 79–97
- Le Goff J 1986 *Histoire et Mémoire*. Gallimard, Paris
- Lewis B 1975 *History—Remembered, Recovered, Invented*. Princeton University Press, Princeton, NJ
- Münkler H 1994 Politische Mythen und nationale Identität. In: Frindte W, Pätzolt H (eds.) *Mythen der Deutschen. Deutsche Befindlichkeiten zwischen Geschichten und Geschichte*. Leske & Budrich, Opladen, Germany
- Nora P 1984/1992 Entre mémoire et histoire. In: Agulhon M, Boutry P (eds.) *Les Lieux de Mémoire*. Gallimard, Paris, Vol. 1
- Nietzsche F 1988 Unzeitgemässe Betrachtungen. Zweites Stück: Vom Nutzen und Nachtheil der Historie für das Leben. In: Colli G, Montinari M (eds.) *Friedrich Nietzsche: Sämtliche Werke. Kritische Studienausgabe in 15 Einzelbänden*. Deutscher Taschenbuch Verlag, Munich, Germany, Vol. 1, pp. 243–334
- Oexle O G (ed.) 1995 *Memoria als Kultur*. Vandenhoeck & Ruprecht, Göttingen, Germany
- Plumb J A 1969 *The Death of the Past*. Macmillan, London
- Ricoeur P 1998 Passé, mémoire et oubli. In: Verlhac M (ed.) *Histoire et mémoire*. Centre Régional de Documentation Pédagogique de l'Académie de Grenoble, Grenoble, France
- Schacter D 1999 The seven sins of memory. Insights from psychology and cognitive neuroscience. *American Psychologist* 54(3): 182–203
- Schudson M 1992 *Watergate in American Memory. How We Remember, Forget and Reconstruct the Past*. Basic Books, New York
- Vidal-Naquet P 1981 *Les Juifs, la mémoire et le présent*. Maspero, Paris

A. Assmann

History and the Social Sciences

History and the social sciences have developed and maintained complex relationships made up of commonalities and differences as well as divergences and convergences. In principle, both disciplinary areas are concerned with the scientific study of the human life-world and the whole range of human behavior, individual and collective, in time and space as well as in the past and the present. Both disciplinary areas originated with the rise of the modern world as a common enterprise to develop systematic, secular, and empirically validated knowledge about the historical-social reality. In this fundamental sense, history can be and should be seen as part of the social sciences or vice versa the social sciences as part of the historical sciences—both constituted by substantially interrelated modes of sociohistorical research (Gulbenkian Commission 1996, Hall 1999).

Despite their fundamental commonality, however, history and the social sciences have followed quite different and diverging epistemological and methodological tracks during their academic institutionalization in Western Europe and North America in the nineteenth and early twentieth centuries. On the one hand, the social sciences (economics, sociology, and political science as well as anthropology) came to be oriented towards the successful epistemological model of the natural sciences and defined themselves primarily as nomothetic and generalizing disciplines. On the other, history predominantly influenced by the humanities and critical against speculative social laws developed traits of an idiographic and particularistic discipline. Against these traditional disciplinary definitions, a converging interdisciplinary cooperation

between history and the social sciences has developed only in the second half of the twentieth century. The social-scientific orientation of history was programmatically formulated in approaches such as the 'new social history,' 'la nouvelle histoire,' 'Social Science History,' or 'Historische Sozialwissenschaft,' developing in leading journals such as *Annales Economies, Sociétés, Civilisations* (1946–), *Past and Present* (1952–), *Journal of Social History* (1967–), *Journal of Interdisciplinary History* (1970–), *Social Science History* (1976–), and *Geschichte und Gesellschaft* (1975–). In a parallel movement, the historical orientation of the social sciences crystallized in new areas like historical sociology, historical economics, historical political science, and historical anthropology. Sometimes even the epistemological identity of history and the social sciences has been stated in a fundamental sense.

More recently, however, this *rapprochement* between history and the social sciences was halted and partly reversed. Within the study and presentation of history a 'revival of the narrative,' a 'new cultural history,' and concepts of the 'historical cultural sciences' have emerged, emphasizing what distinguishes them from the social sciences proper. But it should be stressed that, in a parallel movement, the social sciences have been similarly challenged from within by alternative cultural, interpretative-hermeneutic, and deconstructive-postmodernist epistemologies. Often the so-called 'cultural turn' in history and the social sciences is seen as a replay of the traditional juxtaposition of idiographic and nomothetic positions and conservatively used to reconfirm the traditionally separate institutionalization of both disciplinary areas. But properly understood, this cultural turn can and should be rather viewed as a challenge to the still prevailing principled opposition of generalizing or particularizing disciplinary epistemologies and methodologies. The future task, then, would be to open the historical and social sciences towards a common and multivocal, intra- and interdisciplinary discourse of sociohistorical inquiry.

1. Disciplinary Divergence

In its originally institutionalized—but nowadays increasingly challenged—meaning, history is primarily concerned with the study, description, interpretation, and presentation of historical facts or relics from the past, focusing on: idiographic methods, the analysis of historical contexts and temporal change, and the conceptualization of historical phenomena as particular individualities, their inductive interpretation, and narrative presentation (Burke 1992, Zunz 1985). In the opposite—but also increasingly questioned—direction, the social sciences such as sociology, economics, political science, and anthropology have been traditionally understood as concerned with the study,

generation, and explanation of social facts in the present emphasizing nomothetic methods, systematic analysis, and theoretical explanation by reference to general laws or social regularities (Monkkonen 1994, Skocpol 1984).

This problematically polarizing division between history and the social sciences developed with the institutionalization of the university systems during the nineteenth century under the impact of the natural sciences. Science, since then, came to be equated primarily with the natural sciences and particularly with its conspicuous inorganic and organic branches of physics and biology. Taking the natural sciences as epistemological models, the social sciences emerged as an attempt to discover the general laws and evolutionary mechanisms of modern society (Lepénies 1988, Gulbenkian Commission 1996). Thus, political or classical economy as developed by Adam Smith or David Ricardo tried to find the general laws of modern capitalist market society or in a critical direction as formulated by Karl Marx to uncover the anatomy of emerging modern capitalism. In the same steps, the early founding fathers of sociology such as Claude-Henri Saint-Simon, Auguste Comte, or Herbert Spencer intended to find the structural and evolutionary laws of modern society. At the same time, with the advent of modern mass politics in the American and French revolutions, the emerging political science as in Alexis de Tocqueville or John Stuart Mill attempted to formulate the general features and bases of modern democratic politics. Finally, with the growing awareness of non-Western societies under the impact of European colonialism and imperialism, anthropology tried to analyze the archaic or 'primitive' societies as structural counterparts and evolutionary pre-stages of modern society.

By contrast, history in view of the contingent flow of historical events and human actions as well as the role of cultural and spiritual forces came to be skeptical of the search for quasi-natural laws of the historical-social world. Instead, history followed the humanities, art, and literature as culturally oriented disciplines that seemed to be closer to the spiritual, creative, and poetic aspects of historical development. Particularly under the impact of German 'historicism,' the epistemological and methodological principles of historical knowledge developed into a radical opposition to the positivist search for historical-social facts. The principle of Leopold von Ranke to narrate history 'wie sie eigentlich gewesen ist' formulated the conviction that the historical method has to grasp the historical reality in its manifold particularities. In epistemological terms as developed by Johann Gustav Droysen and Wilhelm Dilthey, history cannot be subsumed under deterministic-mechanistic laws, but has to be understood in its concrete, experience-near, and life-full quality. On this basis, not only were the historical subdisciplines such as political, legal, cultural, social, and economic history principally viewed as hermen-

eutic-interpretative disciplines, but also economics, sociology, and political science as promoted among others by Gustav Schmoller, Werner Sombart, or Max Weber were primarily conceptualized as cultural sciences rather than as social sciences in order to grasp the individual and collective meanings structuring socioeconomic and political actions and processes. These epistemological and methodological differences and divergences were characteristic for the relationships of history and the social sciences not only in Germany, but in all Western societies. However, through the particular influence of German 'historicism' refuting on ontological grounds the applicability of explanatory social-scientific methods to historical events and cultural individualities, these divergences often developed into a principled opposition (Iggers 1993, Lepenies 1988).

The interdisciplinary bridging of this epistemological, analytical, and methodological opposition only became possible with the increasingly historical orientation of the social sciences as well as the growing social-scientific and sociological orientation of history (Abrams 1982, Lipset and Hofstadter 1970, Skocpol 1984). In the first half of the twentieth century, this converging movement was seldom although of crucial significance in preparing the ground for the later developing convergences between the historical and social sciences. Of lasting influence was, first, Emile Durkheim and the continuation of his approach by Marcel Mauss and Henri Berr, attempting to combine the sociology of social change with anthropological comparisons and historical developments (Raphael 1994, Stoianovich 1976). Of fundamental importance was also, second, the German historical-comparative sociology, particularly in the hands of Max Weber, Otto Hintze, Werner Sombart, and in their continuity also of Norbert Elias, who tried to explain the specificities of the Western civilization, its unique economic, political, and cultural order by historical and universal comparisons (Lepenies 1988). Third, strongly influenced by European sociological traditions, modernization theory developed in the USA as a universal social-scientific approach closely connected with historical-comparative modernization research on the historical prerequisites and developments of modern societies on a world scale.

2. Interdisciplinary Convergence

Influenced by these precursors, the historical and social sciences combined programmatically, first in the Annales school (Raphael 1994, Stoianovich 1976) and then later in the American Social Science History approach (Tilly 1981), some proponents of English Marxist historiography (Kaye 1984) and in their emulation in the West German 'Historische Sozialwissenschaft' (Kocka 1977, Wehler 1980). This interdisciplinary convergence led to different forms of

'social-scientific history' or 'historical social science'—despite many national particularities. They materialized in the 1960s and 1970s in a parallel movement in the Western societies (Iggers 1993). In critical opposition to the received tradition of narrative and exclusively interpretive history, the major common methodological aims of social-scientific history or historical social science can be summarized in the following principles. First, social-scientific historical approaches should be oriented to a structural and processual perspective on the various (economic, sociopolitical, and cultural) dimensions of macro-historical change. Second, they should shift historical analysis from a 'top-down' elite perspective to a 'bottom-up' popular or people's history. Third, social-scientific historical approaches should give up the traditional emphasis on political and intellectual history and turn towards economic and social history. Fourth, they should make systematic use of the available social-scientific, theoretical, comparative, and quantitative methods in history and favor systematic explanation instead of hermeneutic interpretation. Finally, this movement to social-scientific history or historical social science was often motivated by a critique of traditional elitist-conservative historiography and a commitment to emancipatory, democratic, or enlightened values.

The common methodological aims of social-scientific history or historical social science, at the same time, were colored and refracted by the respective national intellectual and cultural environments. In France, the 'nouvelle histoire' promoted within the long established interdisciplinary tradition of the Annales school (Raphael 1984, Stoianovich 1976) was characterized by a rather balanced multilevel (*longue durée, conjonctures, événements*) and multidimensional (economy, social structure, and civilizational mentalities) mode of historical analysis. In England, a strong neo-Marxist influence shaped the rise of economic and social history with a particular emphasis on social inequality and class history (Kaye 1984). In the United States the interdisciplinary social science history was influenced by the development of historical-comparative multidimensional modernization research and the special attraction of econometric, statistical, and social research methods for the quantification of historical processes (Tilly 1981). In Germany, Marxian and Weberian traditions of scholarship combined with elements of Western modernization theories and led to 'Historische Sozialwissenschaft' and 'Gesellschaftsgeschichte,' in which the analysis of social inequality and political power with a particular emphasis on the German 'special path' were treated (Kocka 1977, Wehler 1980). At the same time, these national variations reflect not only differing intellectual and theoretical traditions, but also a different emphasis on the use of qualitative and quantitative methods in historical-social inquiry. This is also indicated by the paradigmatic oscillations between the concepts of

(a) historical social science, (b) structural history, and (c) social-scientific history.

'Historical social science' ('Historische Sozialwissenschaft') serves as a loosely defined summary category. Still, the historical-comparative analysis of large-scale and multidimensional processes of social change from traditional life-worlds to modern society has been at the center. In particular, the concepts of historical-comparative modernization research, historical sociology, societal history, and synthetic structural history, with different theoretical and methodological orientations, belong to this category. In the French *Annales* school in a mixture of Durkheimian and Marxian traditions, the comparative analysis of feudalism by Marc Bloch and the social history of capitalism by Fernand Braudel (to be followed by Immanuel Wallerstein's history of the modern world system) are important examples. Among English Marxist historians, the debate on the transition from feudalism to capitalism in a line from Maurice Dobb to Perry Anderson and the social history of modern classes, developed in different directions by Eric Hobsbawm and Edward Thompson, carried the field. In the United States, modernization research guided by a mixture of Max Weber's comparative sociology and Talcott Parsons' structural functionalist theory of social change was of major importance, resulting in path-breaking studies like Neil Smelser's analysis of social change in the English industrial revolution; Shmuel Eisenstadt's civilizational comparison of empires; Barrington Moore's historical explanation of the development of twentieth-century dictatorship and democracy; Tilly's edited volume on the formation of national states in Western Europe; and Reinhard Bendix's study on corresponding historical changes in the legitimacy of political domination (see as an overview Skocpol 1984). In Germany, the modernization paradigm in a mixture of Weber and Marx directed Ralf Dahrendorf's sociological analysis of Germany's troublesome process of democratization or Hans-Ulrich Wehler's (1987–95) and Jürgen Kocka's (1993) comparative social-historical investigations of the German 'special path' and the role of the social classes within it.

'Structural history,' again, has had different meanings with different outlooks. It materialized in innovative areas of a variety of 'new histories' along the lines between the main subsystems of modern society and the related social-scientific disciplines. In the 'New Economic History' of the 1960s, the application of economic concepts and theories has given rise to the history of economic growth, technology, capitalist markets, and institutions; here, David Landes' (1969) *Prometheus* or Sidney Pollard's (1981) *Peaceful Conquest* are examples of a qualitative type, whereas Gianni Tonioli's (1991) comparative history of European industrialization—debating Alexander Gerschenkron's theory of economic backwardness—represents a quantitative type of economic history. In

the 'New Social History' in the 1960s and 1970s, sociological concepts were used in the historical analysis of demographic development, social structure, and social mobility; social groups, occupational structures, and social classes; social institutions, associations, and interest groups; as well as social movements, social protest, and political mobilization (Kocka 1977). This was combined with the application of political-scientific approaches to the history of political modernization, state formation, nation building, the development of democratic regimes and party systems, as well as political power and contestation (Tilly 1984). Further, social-scientific approaches influenced by social anthropology have also been applied to cultural phenomena and developments such as family and socialization, religion and secularization, as well as value systems and mental structures (see different essays in Rürup 1977).

The concept of 'social-scientific history' has been used primarily for the application of rigorous quantitative research methods in structural or social history (Landes and Tilly 1970, Schröder 1994). Not always, but often it has been equated with historical social research or the systematic quantitative operationalization and measurement of historical-social variables. In addition, this often combined with a deductive application of theoretical models to historical data and their explanation through statistical-mathematical and increasingly computer-assisted methods. Also this type of social-scientific historical research followed, either in close cooperation or in separate organization, the main areas of the 'new histories.' Particularly, the 'New Economic History' has been suited to cliometric, statistical, and econometric research methods as for instance in Paul Bairoch's and Maurice Lévy-Leboyer's (1981) analysis of economic growth disparities since the industrial revolution. But also in the new social history, social-demographic family, mobility, and migration studies such as Richard Wall, Jean Robin, and Peter Laslett (1983) as well as time-series of social protest such as Charles Tilly, L. Tilly, and R. Tilly (1975) are prominent examples. In the new political history, the historical change of voting behavior, in particular, became an important field of historical social research (e.g., Best and Thome 1991). But also in the cultural history of mentalities as carried through particularly in the *Annales* tradition, the application of quantitative methods were fruitfully developed (e.g., Shapiro and Markoff 1998).

3. Interdisciplinary Challenges

Successful as these different new historical sub-disciplines in their qualitative and quantitative orientations have been, the guiding vision of social-scientific approaches to history has been the combination of the various structural histories in a unifying societal

history, 'histoire totale' or 'Gesellschaftsgeschichte' (Iggers 1993, Wehler 1980). This vision follows the epistemological ideal of merging the historical sciences and the social sciences in a common paradigm—as essentially expressed in the paradigmatic concepts of social science history, historical sociology, or 'historische Sozialwissenschaft' (Abrams 1982, Smith 1991, Wehler 1980). With this unifying vision, however, the disciplinary traditions and methodological differences between historicized conceptions of the social sciences and social-scientific approaches to history become easily blurred. On the one hand, historical approaches in the social sciences basically proceed within a generalizing frame of reference (Monkkonen 1994). For instance, historical economics follows the basic idea of uncovering time/space bound economic laws or regularities (Kindleberger 1990, Monkkonen 1994, Wallerstein 2000). Also historical sociology is primarily oriented to the general features of time/space bound social processes, patterns, and figurations (as exemplary studies see Eisenstadt 1996, Mann 1986/1993, Smelser 1992). In parallel, historical approaches in the political sciences concentrate on general features of political development (Monkkonen 1994). Also, historicized versions of anthropology are interested in the general evolutionary patterns of human behavior (Monkkonen 1994). On the other hand, social-scientific approaches to history as multidimensional structural or combined societal history usually remain in the confines of a narrative frame of reference—even when applying rigorously social-scientific theoretical concepts and research methods. Thus, in spite of all interdisciplinary convergence, there have been limits to a full merger between historically oriented social sciences and social science oriented history.

A second *problematique*, expressed in the oscillating meanings of historical social science, structural history, and social-scientific history, concerns the relationships between theory and method. Structural approaches to history often use social-scientific concepts in order to systematize, interpret, and explain long-term historical processes in one or another societal dimension. Social-scientific approaches to history are often oriented to theoretical models in order to explain deductively quantified historical-developmental data. Approaches in historical social science and related synthetic orientations to history often use multidimensional theories of social change in order to explain causal connections between the various societal dimensions. Many varieties of historical social science, structural history, and social-scientific history use theories and concepts, deductive and inductive modes of explanation or interpretation, as well as qualitative, quantitative, and comparative methods of analysis. In an effort to systematize these methodological strategies in historical-social inquiry, Theda Skocpol proposed to distinguish model-theoretical, causal-analytical, and interpretive modes of

explanation (Skocpol 1984, pp. 356–91). Charles Tilly, concentrating on the types of comparison and the number of cases and variables involved, differentiated between individualizing, generalizing, encompassing, and variation-finding forms of comparison (Tilly 1984, 84ff.). Charles Ragin (2000), attempting to bridge qualitative and quantitative modes of analysis, distinguished between case-analytical and variation-oriented methodological strategies. Fruitful as these methodological distinctions are, particularly the methodological relationships between explanatory and interpretive research strategies, the theoretical sources and analytical consequences of these methodological distinctions for sociohistorical inquiry still remain unexplored (Hall 1999, p. 173).

A third increasingly salient but divisive *problematique* concerns the role of culture in the differing modes of historical social science, structural history, and social-scientific history. Not only in practice, but also often programmatically, social-scientific approaches to sociohistorical inquiry tend to favor the analysis of the economic, social, and political dimensions and to limit the analysis of culture to value systems or cultural institutions. The contents and forms of culture in their constitutive or causal role in history are often neglected and the corresponding hermeneutic or deconstructive methods to interpret their meanings, symbols, values, and identities are often seen as secondary. Against these prevailing tendencies of social-scientific approaches to culture, there emerged an increasingly influential challenge by the 'new cultural history' or cultural-scientific versions of history (Hunt 1989, Dirks et al. 1994, Mergel and Welskopp 1996). This 'cultural turn' consists of different countercurrents against the social-scientific vision of a systematic, generalizing, and explanatory structural and societal history. First, the revival of the narrative insists on the interpretive, literary, and narrative tasks of history (Stone 1979). Second, the development of micro-history or 'Alltagsgeschichte' emphasizes the complexity of life-worlds as such or as microfoundations of historical macrostructures (Luedtke 1989). Third, the 'linguistic turn' presupposes the all-pervasiveness of language in the social world, thus stressing the analytical task of deconstructing discourses and decoding linguistic-symbolic structures (Hunt 1989). Fourth, the 'cultural turn' is consciously oriented to those social groups such as women, ethnic minorities, or the underclass, seen as neglected by the predominating social-scientific approaches to history (e.g., *History Workshop Journal* 1976— or *Radical History* 1975—). Finally, some authors are influenced by a postmodernist perspective emphasizing the fragmentation and decentering against the structuring and ordering of the social world (Iggers 1993). The result is not only the generation of a whole new set of historical sub-disciplines such as women's and gender history, ethnic and minority history, regional and local history as well

as the history of identities and memory, but also an influential shift from social-scientific history to cultural-scientific history (see however Bonnell and Hunt 1999).

In a sense, the 'cultural turn' signifies a revitalization of historicism and cultural-scientific approaches to history—re-emphasizing qualitative, interpretative, ethnographical, and hermeneutic methods against systematic-explanatory orientations and, with it, reversing the causal or structural determination of history from 'society' to 'culture.' At the same time, this theoretical and analytical movement reflects a parallel shift in the social sciences from social-structural to institutional and cultural paradigms (Skinner 1985, Matthes 1992, McDonald 1996). For instance, in economics and economic sociology the new institutionalism has gained wide recognition, but also cultural neo-Weberian approaches reappear (North 1981, Monkkonen 1994, Swedberg 1998). In various domains of sociology as well, the 'new institutionalism' and various forms of a 'new cultural sociology' become influential (Monkkonen 1994, Crane 1994). Further, in political science the traditional institutional approaches become challenged by cultural paradigms (Monkkonen 1994). All this is influenced by a crucial shift in anthropology from social and structural anthropology to various forms of interpretative, ethnographic, historical, and cultural anthropology (Monkkonen 1994). As a result, social-scientific history is not merely confronted with systematic, generalizing, and explanatory forms of the social sciences, but also with individualizing, interpretative, and reflexive varieties of the cultural sciences. In other words, new opportunities of interdisciplinary cooperation emerge.

4. Interdisciplinary Perspectives

The interdisciplinary cooperation and convergence of history and the social sciences have opened up novel and dynamical fields of sociohistorical research. At the same time, the *idées directrices* of that historical-social-scientific convergence have often privileged quasi-natural-scientific epistemologies and methodologies. As a consequence, generalizing modes of sociohistorical inquiry stand in tension with particularizing modes; structural and macro-analytical theories used in historical-social analysis conflict with action and micro-analytical theories; causal explanation is often limited to the application of theoretical models and causal-analytic methodologies in opposition to interpretation and hermeneutic strategies; and, accordingly, there are tensions between qualitative and quantitative methods, including generalizing and particularizing methods of comparison. These tensions are revealed, on the one hand, within the social-scientific convergence of history and the social sciences in the paradigmatic oscillations between the guiding concepts of historical social science, structural

history, and social-scientific history. These tensions develop, on the other hand, within the wider historical sciences, in the often polarized opposition to renewed historicist and culturalist, narrative and hermeneutic or deconstructivist modes of historical-social inquiry. At the same time, these theoretical and methodological oppositions and polarizations are not solely characteristic for the historical sciences. Parallel disciplinary developments can also be observed in the social sciences and the widespread and often sharp oppositions to cultural-scientific approaches in the various social-scientific disciplines.

These epistemological, theoretical, and methodological oppositions and conflicts have often led into a stalemate in both disciplinary areas of the historical and social sciences between renewed forms of narrative-interpretative and anthropological-cultural history in contradistinction to the nomothetic social sciences. However, these oppositions in the historical as well as the social sciences can and should be used to re-energize a wider intra- and interdisciplinary cooperation between the historical, social, and cultural sciences. These renewed cooperative efforts should concentrate on the crucial epistemological, theoretical, methodological, and analytical dimensions in socio-historical inquiry within and across the established academic disciplines: on the epistemological level, combining social-scientific and cultural-scientific as well as generalizing and particularizing modes of knowledge; on the theoretical level confronting systematically social-structural and cultural approaches to historical-social reality; on the analytical level, mediating between the social-structural, institutional and cultural dimensions as well as between the global, national, regional and local macro-/microdimensions in sociohistorical inquiry; and on the methodological level, clarifying the relationships between qualitative and quantitative as well as historical, social-scientific, and cultural-scientific comparative research strategies (Bourdieu 1995, Haupt and Kocka 1996, Spohn 1998). All these renewed epistemological, theoretical, methodological, and analytical efforts of an inter- and intradisciplinary cooperation should not be done in segmented isolation, but in a multidisciplinary and multivoiced context of practiced sociohistorical inquiry. Such an opening of the historical and social sciences, however, is not only a matter of an intellectual intra- and interdisciplinary cooperation, but also demands an institutional cross-cutting between and across the segmented academic disciplines of the historical and social sciences (Gulbenkian Commission 1996).

See also: Anthropology and History; Cultural History; Demography, History of; Econometrics, History of; Economic History; Economics, History of; Family and Kinship, History of; Gender and Feminist Studies in History; Historical Sociology; History: Forms of Presentation, Discourses, and Functions; History:

Theories and Methods; Marxist Social Thought, History of; Social History; Weberian Social Thought, History of

Bibliography

- Abrams P 1982 *Historical Sociology*. Open Books, Shepton Mallet, UK
- Bairoch P, Lévy-Leboyer M (eds.) 1981 *Disparities in Economic Development Since the Industrial Revolution*. St. Martin's Press, New York
- Best H, Thome H (eds.) 1991 *Neue Methoden der Analyse historischer Daten*. Scripta Mercaturae Verlag, St. Katharinen
- Bonnell V E, Hunt L (eds.) 1999 *Beyond the Cultural Turn: New Directions in the Study of Society and Culture*. University of California Press, Berkeley, CA
- Bourdieu P (ed.) 1995 *Histoire sociale des sciences sociales. Actes de la Recherche en sciences sociales* 106/107
- Burke P 1992 *History and Social Theory*. Polity Press, Cambridge, UK
- Crane D (ed.) 1994 *The Sociology of Culture. Emerging Theoretical Perspectives*. Blackwell, Oxford, UK
- Dirks N B, Eley G, Ortner S B (eds.) 1994 *Culture/Power/History: A Reader in Contemporary Social Theory*. Princeton University Press, Princeton, NJ
- Eisenstadt S N 1996 *Japanese Civilization. A Comparative View*. The University of Chicago Press, Chicago
- Gulbenkian Commission on the Restructuring of the Social Sciences 1996 *Open the Social Sciences: Report of the University Press*. Stanford University Press, Stanford, CA
- Hall J R 1999 *Cultures of Inquiry. From Epistemology to Discourse in Sociohistorical Research*. Cambridge University Press, Cambridge, UK
- Haupt H-G, Kocka J (eds.) 1996 *Geschichte und Vergleich*. Campus Frankfurt
- Hunt L (ed.) 1989 *The New Cultural History: Essays*. University of California Press, Berkeley, CA
- Iggers G G 1993 *Geschichtswissenschaft im 20. Jahrhundert. Ein kritischer Überblick im internationalen Zusammenhang*. Vandenhoeck und Ruprecht, Göttingen
- Kaye H J 1984 *The British Marxist Historians An Introductory Analysis*. Cambridge University Press, Cambridge, UK
- Kindleberger C P 1990 *Historical Economics. Art of Science?* University of California Press, Berkeley, CA
- Kocka J 1977 *Sozialgeschichte; Begriff, Entwicklung, Probleme*. Vandenhoeck und Ruprecht, Göttingen
- Kocka J (eds.) 1993 *Bürgertum und bürgerliche Gesellschaft im 19. Jahrhundert*. Beck, München
- Landes D S 1969 *The Unbound Prometheus: Technological Change and industrial Development in western Europe from 1750 to the Present*. Cambridge University Press, Cambridge, UK
- Landes D S, Tilly C 1971 *History as Social Science*. Prentice-Hall, Englewood Cliffs, NJ
- Lipset S M, Hofstadter R (eds.) 1970 *Sociology and History: Methods*. Basic Books, New York
- Lepénies W 1988 *Between Literature and Science: The Rise of Sociology*. Cambridge University Press, Cambridge, UK
- Luedtke A (ed.) 1989 *Alltagsgeschichte: Zur Rekonstruktion Historischer Erfahrungen und Lebensweisen*. Campus, Frankfurt
- Mann M 1986/1993 *The Sources of Social Power*. Cambridge University Press, Cambridge, UK, 2 Vols
- Matthes J (ed.) 1992 *Zwischen den Kulturen? Die Sozialwissenschaften vor dem Problem des Kulturvergleichs*. Vandenhoeck und Ruprecht, Göttingen
- McDonald T J (ed.) 1996 *The Historic Turn in the Human Sciences*. Michigan University Press, Ann Arbor, MI
- Mergel T, Welskopp T (eds.) 1997 *Geschichte zwischen Kultur und Gesellschaft. Beiträge zur Theoriedebatte*. Beck, München
- Monkkonen E H (ed.) 1994 *Engaging the Past. The Uses of History. Across the Social Sciences*. Duke University Press, Durham, NC
- North D C 1981 *Structure and Change in Economic History*. Norton, New York
- Pollard S 1981 *Peaceful Conquest: the Industrialization of Europe 1760–1970*. Oxford University Press, Oxford, UK
- Ragin C C 2000 *Fuzzy-set Social Science*. University of Chicago Press, Chicago
- Raphael L 1994 *Die Erben von Bloch und Febvre*. Klett und Cotta, Stuttgart
- Rürup R (ed.) 1977 *Historische Sozialwissenschaft: Beitr. zur Einf. in d. Forschungssprache*. Vandenhoeck und Ruprecht, Göttingen
- Schröder C 1994 *Historische Sozialforschung*. Minerva, St. Katharinen
- Shapiro G, Markoff J 1998 *Revolutionary Demands: A Content Analysis of the Cahiers De Doléances of 1789*. Stanford University Press, Stanford, CA
- Skinner Q (ed.) 1985 *The Return of Grand Theory in the Human Sciences*. Canto, Cambridge, UK
- Skocpol T (ed.) 1984 *Vision and Method in Historical Sociology*. Cambridge University Press, Cambridge, UK
- Smelser N J 1991 *Social Paralysis and Social Change: British Working-Class Education in the Nineteenth Century*. University of California Press, Berkeley, CA
- Smith D 1991 *The Rise of Historical Sociology*. Temple University Press, Philadelphia, PA
- Spohn W (ed.) 1998 *Kulturanalyse und vergleichende Forschung Comparativ. Leipziger Beiträge zur Universalgeschichte und vergleichenden Gesellschaftsforschung* 8(1)
- Stoianovich T 1976 *French Historical Method: The Annales Paradigm*. Cornell University Press, Ithaca, NY
- Stone L 1979 The revival of the narrative: Reflections on a new old history. *Past and Present* 85: 3–24
- Swedberg R 1998 *Max Weber and the Idea of Economic Sociology*. Princeton University Press, Princeton, NJ
- Tilly C 1981 *As Sociology Meets History*. Academic Press, New York
- Tilly C 1984 *Big Structures, Large Processes, Huge Comparisons*. Russel Sage Foundation, New York
- Tilly C, Tilly L, Tilly R 1975 *The Rebellious Century 1830–1930*. Harvard University Press, Cambridge, MA
- Toniolo G Sylla R (ed.) 1991 *Patterns of European Industrialization in the 19th Century*. Routledge, London
- Wall R, Robin J, Laslett P (eds.) 1983 *Family Forms in Historic Europe*. Cambridge University Press, Cambridge, UK
- Wallerstein I M 2000 *The Essential Wallerstein*. New Press, New York
- Wehler H-U 1980 *Historische Sozialwissenschaft und Geschichtsschreibung*. Vandenhoeck und Ruprecht, Göttingen
- Wehler H-U 1987–1995 *Gesellschaftsgeschichte Deutschlands*. Beck, München, 3 Vols.
- Zunz O 1985 (ed.) *Reliving the Past: the World of Social History*. University of North Carolina Press, Chapel Hill, NC

W. Spohn

Copyright © 2001 Elsevier Science Ltd. 6835
All rights reserved.

History: Forms of Presentation, Discourses, and Functions

History as an institutionalized, academic discipline is a continuation of extradisciplinary forms of dealing with the past and, therefore, has never enjoyed a monopoly on its representation. Narrative representations of origins—be it in the form of myths, legends, religions, or ideologies—are found in most cultures. These different forms of representing the past offer collectives an orientation in time: they give a meaning to the present by locating the present on a time axis between a future and a past.

Next to the variety of disciplinary forms of historical representation (Stern 1970, Kelley 1991, Lorenz 1997), these extradisciplinary representations of history still exist, often competing with others and disciplinary history. The problem of how to demarcate the extradisciplinary from the disciplinary ways of dealing with the past, has therefore haunted history as a discipline from its very beginning. Traditionally, the discipline's identity has been based on its truthclaim, its critical methods, and its archival foundation (see *History: Theories and Methods*). This article will only deal with the disciplinary representations of the past in the West. It will only take extradisciplinary representations into account in order to cast light on the origins and the official borderlines of the discipline (Fain 1970, McNeill 1986).

1. Historical Representation and its Function as a Historical Problem

Although the concept of representation has a clear-cut meaning in philosophy—that of an image, in which the structure of the represented phenomenon is maintained (Mormann 1997)—in history the term is usually used as an equivalent of concepts like image or construction. In titles like 'The representation of Hitler in the US press, 1933–1945' or 'The representation of the Black Woman in the literature of the Antebellum South,' historians emphasize the constructed and interpretative character of historical reconstructions: history cannot be directly presented, but is *represented* by the historian in language. This emphasis has roots in an awareness that historical representations are not mirror images of history itself. It is usually accompanied by an awareness of the fundamental plurality of historical reconstructions. At the same time it is sometimes accompanied by an awareness of the practical functions (or the politics) of representation.

For the consciousness of this 'representational' character of history, two developments were crucial. The first development was the perception of history as an object of knowledge, independent of its individual representations. According to Koselleck this development originated in Europe, in the mid-eighteenth century, when history as a process—*historia*—was first

distinguished from the traditional *historiae*, i.e., the stories told about the past. History could only be conceived of as an independent object of knowledge when it was no longer regarded as a binding tradition or as a practical craft belonging to literature or rhetoric, i.e., as the art of telling (his)stories in order to persuade a public, as had been the case in Europe from classical antiquity until mid-eighteenth century. History's functions only developed into a problem after the traditional idea, that past experiences were relevant for present purposes and histories could teach you by example what to do and what not to do (*historia magistra vitae*), began to lose its plausibility.

The second development, that promoted the awareness of the representational character of history, was the 'discovery' of its perspectival character during the Enlightenment. Although the possibility of multiple stories about the same events was known already—and a central concern of rhetorics—Enlightenment historians were the first to recognize explicitly that the historical representations necessarily vary with the perspectives of the historians and with the development of time. The Enlightenment also stimulated the tendency to identify cognitive, scholarly activity with empirical disciplines, which was a consequence of the competition of the *Geisteswissenschaften* with the modern natural sciences (Scholtz 1991, Momigliano 1990). Koselleck dates the fundamental change in the relationship between Europeans and their sense of the past halfway through the eighteenth century; according to Pocock these changes had already started in the seventeenth century (Koselleck 1979, Pocock 1961). The traditional idea of the continuity between the past, the present, and the future was gradually replaced by the experience of permanent change, so typical of modernity (Scholtz 1991, Hardtwig 1990). The past thus gradually turned into a 'foreign country' (Plumb 1973, Lowenthal 1985, Rüsen 1983–9).

Early nineteenth century, this experience of difference between the past and the present would lead to the association of history with the notions of non-recurrence and nongeneralizability, later expressed in the concepts of uniqueness and individuality (see *Historicism and History: Theories and Methods*). History thus came to be conceived of not just as change through time—which could be cyclical and thus repeatable—but as *change as a process of irreversible and nonrepeatable development*.

With the emergence of this specific new sense of the past, the stage was set for the development of history and the social sciences into separate disciplines. The object of history was increasingly identified with nonrecurring phenomena in the human past, while the social sciences were identified with recurrent phenomena in the human present. Because recurrence is a precondition for generalization and thus for (structural and processual) theories, history and theory gradually turned into opposites, excluding one another. Thus the idea was born that the forms of

representation in history differed fundamentally from those in the social sciences. While historical representation was usually characterized as time-oriented and narrative, social scientific representational forms were characterized as analytical and theoretical (see *History and the Social Sciences*, Meran 1985, Burke 1993, Lorenz 1997). The modern question of narrative in history was thus born in the eighteenth century as a concomitant of the rise of modernity.

2. The Classical Roots of the Modern Debate

2.1 The Aristotelian Heritage

Both the Enlightenment conception of history and its critique in Herder and in later, Romantic thinkers—the two conceptions of history that still define the terms of the modern debate—must be viewed in the context of Aristotelian philosophy, because the Aristotelian definitions and distinctions of disciplines were still very influential in the eighteenth century (Aristotle 1963). In Aristotle's (ca. 384–322 BC) philosophy, there is a basic distinction between empirical knowledge of singular facts and universal knowledge of general facts. Concrete, empirical knowledge just formulates the contingent and particular characteristics of its object, while theoretical knowledge formulates its necessary and universal characteristics. These latter characteristics are viewed as recurrent and unchanging and were therefore limited to the domain of nature. Theoretical knowledge was therefore limited to the domain of nature too. Theoretical knowledge, according to Aristotle, does not only inform us *what* reality looks like, but also *why* this is the case: theoretical knowledge is always explanatory knowledge, because it informs us about the causes of what happens. Therefore, explanations in science are causal explanations and are ideally based on theoretical knowledge. This idea would dominate later philosophy of science in the form of positivism from Auguste Comte to Carl Hempel (see *History: Theories and Methods*; Meran 1985).

Within Aristotelian philosophy the domain of human activity or culture (praxis) held a fundamental different status from nature, because contingency, particularity, and change characterize the first domain. This domain, therefore, was not open to theoretical knowledge and to scientific explanation in principle (*'de singularium non est scientiae,'* as Duns Scotus wrote in the thirteenth century). History, according to the Aristotelian system, belonging to the domain of culture, could never be regarded as a science and was located in the domain of practical knowledge, more akin to poetry (including literature).

Although both poets and historians dealt with the same phenomena, i.e. human actions and events (albeit respectively in verse and in prose), history only dealt with what was real and poetry with what was imagined.

The historian, according to Aristotle, 'describes what happened' while the poet 'describes what can happen' (Aristotle 1963). This distinction was later codified in the distinction between the *res factae* [Latin: *res* = affairs; *factae* = done] or 'things done,' which was the domain of history, and *res fictae* [Latin: *fictae* = imagined], which only concerned poets.

The duty of the historian to stick to the facts and to tell the truth, the whole truth and nothing but the truth, derived mainly from the model of Thucydides. It received its canonical formulations in Cicero (ca. 106–43 BC) and Lucian (ca. 125–200 AD), who also linked truth in history directly to impartiality by avoiding moral blame and envy (*sine ira et studio*). The German historian Leopold von Ranke (1795–1886) would only repeat these classical ideals in the nineteenth century, with great force and effect (Iggers 1997).

According to Aristotle, the poet enjoyed a 'constructive liberty' not shared by the historian. While the unit of the historian is defined by what had happened to specific subjects at a specific time at a specific place, the unit of the poet was not defined in such a way. In poetry actions and events are put together by a unit of meaning, that is by constructing a whole with a beginning, middle, and an end, held together by a plot. The parts of the poetic unit all fulfill a function in—i.e. derive their meaning from—this whole. Because of its constructive quality, Aristotle held that poetry was more philosophical than history. This characterization was also based on the idea that poetry—for instance, tragedy—was concerned with general problems—types—and not only with particulars, like history.

The Aristotelian definition of history would define the parameters of the debates on historical representation up to the present. The opposition of nature and culture and the identification of nature with necessity and repetition, and culture with contingency and change, has ever since defined the methodological battlefields in philosophy of history. This opposition linked nature to (universal) theory while excluding (particular) history from theory *a priori*. As influential has been the opposition of poetry and history and the identification of history with the domain of the factual in opposition to the domain of fictionality (see *History: Theories and Methods*). Last but not least the Aristotelian opposition of historical units and poetical units has put the relationship between chronological and narrative ordering of events on the agenda of philosophy of history, i.e. the relationship between chronology and narrative. This opposition raised the question what narrative explanations are and how they relate to causal explanations.

2.2 The Rhetorical Tradition

Another fundamental root of the modern debate on historical representation can be traced back to classical rhetoric (Koselleck et al. 1982). Accounts of the past

were dealt with by classical authors like Aristotle, Cicero, Lucianus and Quintilianus (ca. 35–96 AD) primarily in the context of rhetoric, especially forensic rhetoric, i.e., rhetoric used to persuade a judicial court. So, the problem how to compose and write an account of the past was primarily dealt with as a problem in a communicative context and not as a cognitive problem, as is the case since the nineteenth century. Therefore the question how to produce a convincing history loomed larger than how to write a truthful history—which in turn explains why rhetoric later acquired a doubtful reputation until it was officially expelled from the scholarly domain by Descartes only to return late twentieth century in the train of postempiricism and post-positivism (see *History: Theories and Methods* and McCloskey and Megill 1987). The origin in forensic rhetoric also explains why the domain of history was identified with the domain of human action.

Now to be convincing, according to rhetoric, truth may help, because Plato argued that true arguments are the most convincing. Quintilianus, however, later pointed out that truth as such was neither a necessary nor a sufficient condition for being convincing. Being convincing is directly related to being perceived as probable rather than as being factually true, because some truths are hard to believe. In order to be convincing, according to rhetoric, past action primarily must be explained in (motivational) terms understandable and plausible for the audience addressed. Therefore, the reconstruction of past actions was basically conceived of in the rhetorical tradition as an act of translation of factual information by a speaker, who has a specific aim and who addresses a specific audience.

The rhetorical emphasis on probability instead of on truth was a consequence of its focus on communication and the awareness that factual truth is not a communicative category. It did not imply any plea for blurring the distinction between facts and fiction in history, nor the idea that truth in history fulfilled no practical function. On the contrary, the truth of the past was sought after, because only true accounts of past actions were useful as *exempla* for later generations (Momigliano 1990, Hay 1977). History was thus usually regarded as inherently pragmatic for some 20 centuries until the presuppositions of this pragmatic conception lost their plausibility in the second half of the eighteenth century.

3. The Modern Debate on Historical Representation

3.1 The Problem of Narrative

Basically, the theories of historical representation developed since the eighteenth century can be classified on basis of their relationship with the Aristotelian and

the rhetorical characterizations of history. There is a genus of theories, that can be seen as a continuation of Aristotelian theory, and there is a genus of theories, that opposes Aristotelian theory. Both genus contain several species.

The first genus consists of theories that have transformed the Aristotelian analysis of poetical or literary narrative into theories of historical narrative. According to these theories, historical narratives are structured by plots (P. Ricoeur, P. Veyne, H. White, D. Carr), colligatory concepts (W. Walsh), configurational ideas (J.-G. Droysen, L. Mink), or metaphors (F. R. Ankersmit). This view implies that historical narratives are not structured by a chronological order; instead they are structured by linguistic instruments, which synthesize chronologically disparate elements into a synchronic unity (as, e.g., is done by concepts like 'The Enlightenment,' 'feudalism,' or 'imperialism'). Narratives, thus, are fundamentally different from chronologically ordered historical representations, such as chronicles and annals. The fundamental difference between literary and historical narrative, according to most of these narrative theories, is the difference between fiction and fact: while the construction of literary narratives is only restricted by what is plausible, historical narratives are restricted by what is factually true. Therefore, history could and did claim a scholarly status in being empirical or a *Wissenschaft*.

This position was first developed by the Historical School from Niebuhr and Ranke onwards and is philosophically legitimized by historicism that posits a fundamental connection between the narrative form of representation and the notions of time and of individuality. The basic idea is that narrative is the linguistic medium, which shows how a phenomenon (a person, nation, idea, etc.) develops its defining characteristics—its identity or individuality—in time, just as a biography shows how an individual has developed its identity in time (Lübbe 1977, Veyne 1978, Ankersmit 1994). Although most narratives appear to follow the natural flow of time, it is, in fact, the other way around: the central idea of a narrative (e.g., the Enlightenment, the Age of Revolutions, etc.) determines which temporal units and orders (periodization, etc.) are the relevant ones for the narrative in case. Narrative time is thus always constructed time. Droysen formulated this insight in 1857 (Droysen 1977).

Within this first genus of narrative theories, there are several species which show important differences. The first species, exemplified by Louis Mink, Herman Lübbe, and Frank Ankersmit, has developed the argument that explanation by narrative is *sui generis* and is logically independent of causal explanation. All advocates of this species hold that the narrative mode of representation is explanatory in itself.

The second species, exemplified by Arthur Danto and Jörn Rüsen, holds that explanation by narrative,

although different from lawlike explanation, is based on causal explanation nevertheless. The narrative mode of representation is analyzed here as a set of causal chains. A third species of narrative theories, exemplified by Paul Ricoeur, David Carr, and Paul Veyne, has developed the argument that historical explanation has a narrative form and a plot structure. Ricoeur and Carr, moreover, argue that human experience itself is narratively structured and that there is an *a priori* fit between human experience and narrative explanation. Both individuals and communities can be said to have the same narrative form of self-constitution, because their identities can only be represented in the form of narratives.

Whatever their differences, all of these three species of narrative theories regard narratives primarily as explanatory schemes. All three try to answer the Aristotelian question how narratives explain, if not straightforward causally.

A fourth species of narrative theories is not primarily focused on issues of explanatory logic, but on the rhetorical problem why narratives are regarded as plausible and convincing. Ann Rigney and Hayden White in his *Metahistory* exemplify this type of narrative theory. First, White argues that all historical narratives fulfill an ideological function, aimed at specific audiences. Second, he argues that historical narratives contain one of four literary plot structures: comedy, tragedy, satire, or romance. Third, he argues that all history can be emplotted freely in these four manners. Narratives are accepted by an audience as probable and as plausible, according to White, when the narrative plot structure and its ideological implication are recognized and accepted as such.

The acceptability of historical narratives, so White, is thus relative to audiences and relative to ideological aims and not only dependent on the truth of the factual statements in the narrative. In this respect, White's theory has a rhetorical ring, although his fixation on plot structures is at the same time Aristotelian. However, on the issue of narrative truth, White deviates fundamentally from both Aristotelian and rhetorical presuppositions. Contrary to Aristotle, he blurs the borderline between fact and fiction in history by stating that historical narratives are 'fictions of factual representation.' Consequently, according to White, there are no epistemological grounds for preferring one historical representation over another, only political and aesthetic ones. So, along with rhetoric, he holds that plausibility bears no necessary relation to truth. However, contrary to both Aristotelian and rhetorical theory, White severs the crucial link between research and composition of historical narrative: the composition of narratives is not treated as following from—and restricted by—research, but as a fully independent process (Lorenz 1998).

The second genus of theories of historical representation somehow opposes Aristotelian theory. Historically this genus derives from eighteenth century

Enlightenment conceptions of history. Contrary to Aristotle, Enlightenment historians like Voltaire, Condorcet and Chladenius held that history could be both theoretical—in the sense of containing knowledge of a general character—and scientific—in the sense of being empirical. 'Theoretical' or 'hypothetical' history was thus claimed as a legitimate 'scientific' project, in which 'theory' and empirical knowledge could be integrated. Theoretical knowledge was no longer exclusively identified with the domain of nature and no longer exclusively interpreted as knowledge of what is necessary and universal. All knowledge of a general character was regarded as theoretical, including empirical generalizations and theories of the middle range. In the twentieth century this type of position has been defended by the proponents of the conception of history as a social science, who conceived of history as a mix of particular and general characteristics. As the general characteristics could be theorized, they emphasize the importance of theories and models in historical reconstructions (Wehler 1973, Kocka 1977, Tilly 1981, Burke 1993, Iggers 1997). This type of representation, being explicit about the concepts and arguments used, is often called analysis instead of narrative, emphasizing its affinity to the social sciences. This happened especially when 'traditional' history and its function were seriously questioned between 1960 and 1980.

4. *The Nineteenth-century Origins of the Modern Debate on the Functions of History*

From classical antiquity until the eighteenth century the pragmatic presupposition that past experience was meaningful for present purposes was not questioned seriously, nor that the (empirical) truth coincides with the (ethical) good. From the end of the eighteenth century onwards, however, historians and philosophers began to doubt the salutary nature of history seriously and therefore the function of history: when history is no longer experienced as a meaningful and salutary process, what good is the study of history then? Nietzsche's all too nasty question in *The Use and Disadvantage of History for Life* (1874), whether the study of history had any use at all, was the logical endpoint of this vanishing faith in history. European historians basically countered these doubts by two strategies of demarcation. First, by attempts to enhance the scholarly status and 'objective' credentials of history by increasingly expulsing the rhetorical and artistic dimensions from its disciplinary domain. An increasing tendency to conceive of history as a *Wissenschaft*, opposed to *Kunst* or art, by identifying the discipline with its empirical research methods (*Forschung*) at the cost of the literary and artistic methods of its presentation (*Darstellung*) was the result (see Hardtwig 1990, Scholtz 1991 and *History: Theories and Methods*).

Second, these doubts were countered by demarcating disciplinary history sharp from any substantial philosophy of history, especially from its Enlightenment and Hegelian variants (including Marxism), and from any practical function. In the famous words of Ranke historians were not to pass any judgement on the meaning of the past—as the Enlightenment had done—but only ‘to show what actually had happened.’ By this demarcation strategy the reflection on the practical functions of history was excluded from the disciplinary discourse and was expelled to philosophy of history.

This epistemological demarcation had become possible while history was professionalizing and attaching itself to the national states both institutionally and financially. Because both the state elite’s and the majority of the professional historians presupposed that education in (national) history was essential for ‘nation building,’ the practical functions of history stopped being a problem. The practical function of history was now conceived of as indirect—through individual and collective identity formation alias *Bildung*—rather than as direct through *exempla*. This view was also held by proponents of a ‘contemplative’ function of history, like Ranke.

The ideal of objectivity was silently identified in historical discourse by promoting the practical cause of the nation state. This identification was not seen as a problem, because in the dominant conceptions of history the historical process itself was identified with the genesis of nations and states. It would take two world wars before this identification turned into a serious problem due to an increasing lack of plausibility. Nietzsche’s nasty question, until then comfortably put aside by many professional historians as ‘philosophical,’ had to be faced at last: what good is the study of history anyway?

Nietzsche’s analysis and classification of the functions of history in *The Uses and Abuses of History* although framed as a critique of nineteenth century discipline of history in Germany, has proved to be very fruitful up to the present day (Rüsen 1983–89; Kocka 1989, Hardtwig 1990, Tollebeek 1990). Basically, Nietzsche distinguishes three ways to deal with the past and three types of history that correspond with three different attitudes and basic needs of human beings. Therefore, history has three basic functions, which cannot be reduced to each other and which serve as mutual correctives.

The first way to deal with the past is pragmatic, as an active *acting* being. Actors may address the past in search of examples, which, according to Nietzsche, deserve imitation. The past addressed in this way is treated as ‘a collection of monuments.’ That’s why Nietzsche calls this pragmatic type of history *monumentalistic history*.

The second way to deal with the past is full of piety, as a *contemplative* being, when one is looking for one’s roots in the past in search of one’s identity. The past

addressed in this way is treated as ‘a collection of antiquities’ or as ‘a museum’; therefore, this ‘pious’ type of history, i.e., history just for the sake of the past, is labeled *antiquarian history* by Nietzsche. Because man is not only directed towards the past, but also towards the future, it is impossible always to treat the past as a ‘museum.’ In order to create something new it is necessary to break with the past to some extent and to replace it. Therefore, it is also necessary to address the past in a critical way, as something to be criticized as fit for replacement. This type of handling the past, which Nietzsche compares to ‘cutting one’s roots loose with a knife,’ is called *critical history*.

All three types of history have their own type of dangers and deficiencies, which can only be remedied by history of the other types. Monumental history, for instance, tends to equalize the past and the present and to be very selective, obliterating differences and huge parts of the past, unfit for the monumental function. This tendency can only be counteracted by the antiquarian impulse, which in turn suffers from ‘historical myopia’ and from a ‘blind drive to collect.’ This tendency, in turn, can only be counteracted by the critical impulse, which in turn must be held in check by the other two in order to prevent critical history from degenerating into a war against one’s past.

Historiography can be seen as the combined product of these three impulses, in varying mixes. Disciplinary history since the nineteenth century, however, has usually—although not exclusively—embraced the antiquarian self-image and legitimization, because the recognition of both the monumentalistic and the critical approaches to the past would raise the problem of its practical functions and its objectivity. In accordance with this contemplative self-image of most historians, the antiquarian presentation of identity—i.e., the narrative reconstruction of the genesis of individuality in time—is usually presented as the discipline’s only practical function. As in Nietzsche’s time, the monumental and the critical functions of history are usually not discussed within the disciplinary boundaries (Angehrn 1985). At times, however, as in the 1970s, the situation has been different, because of an influential number of ‘critical’ historians in the profession (Kocka 1989).

Symptomatically for this state of affairs is that recently the monumental function of history has been delegated to a separate domain, i.e., the domain of collective memory and of *lieux de mémoire* (see *History and Memory*). The same holds for the critical function of history, which is central to projects like gender history, postcolonial studies, and Foucault’s critical history (which replaces the Enlightenment self-image of the West by an image of a disciplined ‘carceral society’). These forms of critical history were often not regarded as belonging to the domain of history proper or are at best tolerated in its margins (Foucault 1980, Scott 1988). Mainly in one historiographical domain—that of the Holocaust—reflection on the functions of

history has included both its critical and monumental functions, because the very possibility of the anti-quarian approach is fundamentally questioned here. Historical representation as such has turned into a problem, because it is doubted whether the normal categories and procedures of historical representation in order to 'objectify' the past can function in this 'unspeakable' domain. Some hold that psychoanalytic categories like trauma, repression, transfer, splitting off, and working through are more apt to characterize the relationship between the historical representations and represented historical reality. Others have suggested that the 'unspeakable' and 'overwhelming' dimensions of the Holocaust can be better characterized through the notion of the sublime, suggesting a deep analogy between the confrontation with the Holocaust and the confrontation with an earthquake or an avalanche (see *Holocaust, The*, Friedländer 1992, LaCapra 1994).

Mainly in this historiographical domain, the discussion about the functions of history is directly connected to the discussion on representation and its limits. Whether this more inclusive type of debate will alter the dominant discourse on history remains as yet to be seen.

See also: Historicism; Historiography and Historical Thought: Classical Period (Especially Greece and Rome); Historiography and Historical Thought: Current Trends; Historiography and Historical Thought: Modern History (Since the Eighteenth Century); History and Memory; History and the Social Sciences; History: Overview; History: Theories and Methods; Holocaust, The; Linguistic Turn and Discourse Analysis in History; Syntax-Semantics Interface

Bibliography

- Angehrn E 1985 *Geschichte und Identität*. de Gruyter, Berlin
- Ankersmit F R 1994 *History and Tropology. The Rise and Fall of Metaphor*. University of California Press, Berkeley, CA
- Aristotle 1963 *The Philosophy of Aristotle*. Bambrough R (ed). New American Library, New York
- Burke P 1993 *History and Social Theory*. Cornell University Press, Ithaca, NY
- Carr D 1986 *Time, Narrative and History*. Indiana University Press, Bloomington, IN
- Danto A C 1965 *Analytical Philosophy of History*. Cambridge University Press, Cambridge, UK
- Droysen J-G (1977) *Historik*. Fromman-Holzboog, Stuttgart
- Fain H 1970 *Between Philosophy and History: the Resurrection of Speculative Philosophy of History Within the Analytic Tradition*. Princeton, NJ
- Fay B, Pomper Ph, Vann R (eds.) 1998 *History and Theory. Contemporary Readings*. Cambridge
- Foucault M 1980 *Power/Knowledge*. Harvester, Brighton
- Friedländer S (ed.) 1992 *Probing the Limits of Representation. Nazism and the 'Final Solution'*. Harvard University Press, Cambridge, MA
- Gardiner P (ed.) 1959 *Theories of History; Readings from Classical and Contemporary Sources*. New York
- Gossman L 1990 *Between History and Literature*. Harvard University Press Cambridge, MA
- Hardtwig W 1990 *Geschichtskultur und Wissenschaft*. DTV, Munich, Germany
- Hay D 1977 *Annalists and Historians. Western Historiography from the 8th to the 18th Century*. Methuen, London
- Iggers G 1997 *Historiography in the Twentieth Century. From Scientific Objectivity to the Postmodern Challenge*. Wesleyan University Press, London
- Kelley D (ed.) 1991 *Versions of History from Antiquity to the Enlightenment*. Yale University Press, New Haven, CT
- Kocka J 1977 *Sozialgeschichte. Begriff—Entwicklung—Probleme*. Vandenhoeck und Ruprecht, Göttingen, Germany
- Kocka J 1989 *Geschichte und Aufklärung*. Vandenhoeck und Ruprecht, Göttingen, Germany
- Koselleck R 1979 *Vergangene Zukunft. Zur Semantik geschichtlicher Zeiten*. Frankfurt, Suhrkamp, Germany
- Koselleck R, Lutz H, Rüsen J (eds.) 1982 *Formen der Geschichtsschreibung*. Deutscher Taschenbuch, Munich, Germany
- LaCapra D 1994 *Representing the Holocaust. History, Theory, Trauma*. Cornell University Press, Ithaca, NY
- Lorenz C 1997 *Konstruktion der Vergangenheit. Eine Einführung in die Geschichtstheorie*. Cologne
- Lorenz C 1998 Can histories be true? Narrativism, positivism and the "metaphorical turn" *History and Theory* 37: 309–29
- Lowenthal D 1985 *The Past is a Foreign Country*. Cambridge University Press, Cambridge, UK
- Lübke H 1977 *Geschichtsbegriff und Geschichtsinteresse. Analytik und Pragmatik der Historie*. Basel
- McClosky D, Nelson J S, Megill A (eds.) 1987 *The Rhetoric of the Human Sciences: Language and Arguments in Scholarship and Public Affairs*. University of Wisconsin, Madison, WI
- McNeill W H 1986 *Mythistory and Other Essays*. University of Chicago Press, Chicago
- Megill A (ed.) 1994 *Rethinking Objectivity*. Duke University Press, Durham, NC
- Meran J 1985 *Theorien in der Geschichtswissenschaft*. Vandenhoeck und Ruprecht, Göttingen, Germany
- Mink L O 1987 *Historical Understanding*. Cornell University Press, Ithaca, NY
- Momigliano A 1990 *The Classical Foundations of Modern Historiography*. University of California Press, Berkeley, CA
- Mormann T 1997 Ist der Begriff der Repräsentation obsolet? [Is the concept of repression obsolete?]. *Zeitschrift für Philosophische Forschung* 51: 349–66
- Nietzsche F 1981–1874 Vom Nutzen und Nachteil der Historie für das Leben. In: Nietzsche F (ed.) *Unzeitgemäße Betrachtungen*. Insel, Frankfurt, Germany, pp. 95–185
- Novick P 1988 *That Noble Dream. The 'Objectivity Question' and the American Historical Profession*. Cambridge University Press, Cambridge, UK
- Plumb J 1973 *The Death of the Past*. Penguin, Harmondsworth, UK
- Pocock J 1961 The origins of the study of the past: a comparative approach. *Comparative Studies in Society and History* 4: 209–46
- Ricoeur P 1983–85 *Temps et récit*. Seuil, Paris
- Rigney A 1990 *The Rhetoric of Historical Representation: Three Narrative Histories of the French Revolution*. Cambridge University Press, Cambridge, UK

- Rüsen J 1983–89 *Historische Vernunft. Grundzüge einer Historik*. Vandenhoeck und Ruprecht, Göttingen, Germany
- Scholtz G 1991 *Zwischen Wissenschaftsanspruch und Orientierungsbedürfnis. Zu Grundlage und Wandel der Geisteswissenschaften*. Suhrkamp, Frankfurt, Germany
- Scott J W 1988 *Gender and the Politics of History*. Columbia University Press, New York
- Stern F (ed.) 1970 *Varieties of History. From Voltaire to the Present*. Macmillan, London
- Tilly Ch 1981 *As Sociology Meets History*. Academic Press, New York
- Tollebeek J 1990 *De toga van Fruin. Denken over geschiedenis in Nederland 1860–1960*. Weselbibliotheek, Amsterdam
- Veyne P 1978 *Comment on écrit l'histoire?* Seuil, Paris
- Walsh W H 1967 *Philosophy of History: an Introduction*. Harper and Row, New York
- Wehler H-U 1973 *Geschichte als Historische Sozialwissenschaft*. Suhrkamp, Frankfurt, Germany
- White H 1973 *Metahistory. The Historical Imagination in Nineteenth Century Europe*. Johns Hopkins University, Baltimore

C. Lorenz

History of Science

The history of science studies the emergence and development of systematic knowledge. Just what kind of knowledge that is—whether it is defined by subject matter, reliability of results, contradistinction to technical know-how, cumulative progress, disciplinary consensus, and/or method of inquiry, demonstration, or explanation—is a point of persistent debate and reflection within the history of science and its cognate fields in philosophy, sociology, and anthropology. Linguistic and historiographic traditions diverge sharply on this point. Whereas Francophone and Anglophone scholars have since the mid-nineteenth century largely reserved the honorific ‘science’ for the natural sciences and mathematics (influenced by Auguste Comte’s hierarchy of sciences in his *Cours de philosophie positive* 1830–42, in which only some had reached the stage of ‘positive’ knowledge), the German *Wissenschaft* still refers to the range of fields covered by the Latin *scientia*, from physics to philology. Moreover, recent explorations of experiment and other practices have blurred the venerable boundary between *ars* and *scientia*, technique and systematic knowledge (whether understood narrowly as ‘science’ or broadly as *Wissenschaft*); close investigation of the sciences even in the restrictive English and French usage has furthermore revealed considerable variety among methods and explanations in, for example, biology and physics. Finally, studies of pre-seventeenth-century and non-Western intellectual traditions have made the creation of categories like ‘science,’ as well as the classifications of knowledge that order such categories, themselves objects (rather

than presuppositions) of historical research. Hence this article will not attempt to delineate sharply the boundaries of ‘science,’ but rather briefly survey the (a) emergence and status of the history of science as a recognized field of inquiry; (b) principal divisions of the field and disciplinary affinities; (c) periodizations and narratives that have shaped those divisions; and (d) major historiographical developments since about 1970 (for an earlier survey see Kuhn [1968] 1972).

1. Emergence and Divisions of the Discipline

The history of science is an ancient pursuit, but a relatively young discipline. From Aristotle through the early nineteenth century, practitioners of one or another branch of knowledge have variously used the history of their field to argue for its dignity and importance, introduce it to beginners, situate it within a broader cultural milieu, summarize the literature to date, position themselves in relationship to that literature, praise and blame predecessors, give evidence of progress, extrapolate a program for future research, and draw lessons concerning the nature of knowledge and the conditions for its flourishing. In works like Joseph Priestley’s *The History and Present State of Electricity* (1767) or Georges Cuvier’s *Rapport historique sur les progrès des sciences naturelles depuis 1789* (1808), the history was inseparable from the science. Particularly in fields dominated by empirical research, the judicious sifting and ordering of past results was a precondition for coherence and a means for achieving consensus on what was reliably known and where the major challenges for future research lay. These functions are preserved today in the scientific review article and the posing of key ‘problems’ (e.g., the celebrated Hilbert problems), and practicing scientists occasionally still appeal to the past history of their field for present guidance, especially in times of crisis (Graham et al. 1983). By the mid-nineteenth century, however, histories of science had become distinct from scientific publications, although they were still written primarily by scientists, including prominent figures such as William Whewell, Marcellin Berthelot, Ernst Mach, and Pierre Duhem. Their histories often criticized the current state of science by establishing the genealogy of a controversial hypothesis (e.g., atomism), analyzing the origins of a suspect concept (e.g., absolute space) for hidden flaws, or pleading the superiority of one approach to science over another (e.g., Kantian ideas over Comtean facts). By 1900, histories of science had become a genre distinct from science, but they were still motivated by, and deeply engaged with, contemporary scientific developments (Laudan 1993).

History of science coalesced only gradually as a recognizable discipline in the twentieth century, with its own distinctive program of training, institutions (journals, professional societies, university positions), and scholarly standards. Spurred by the organiza-

tional efforts of such scholars as Paul Tannery in France and Karl Sudhoff in Germany, there was a spurt of professionalizing activity around 1900, when the first international congress on the subject was held in Paris. The German Gesellschaft für Geschichte der Medizin und der Naturwissenschaften was established in 1901; the first chairs in the history of science also appeared at European universities around this time (Kragh 1987). The Belgian historian George Sarton, who emigrated to the USA after World War I, brought with him the journal *Isis* (est. 1912), which became the official organ of the American History of Science Society after its establishment in 1924.

Although Sarton's Comtean view of the history of science as a saga of intellectual and moral progress that should evaluate the past in light of present scientific understanding (Sarton [1927] 1948) has left little trace in current historiography, his tireless attempts to institutionalize and organize the discipline in the form of university departments, specialized journals, and research tools like the *Isis Current* (earlier *Critical Bibliography*) (History of Science Society 1967) and *Dictionary of Scientific Biography* (Gillispie 1981) had a lasting impact on the field (Thackray and Merton 1972). During the 1950s graduate programs dedicated to the history of science (sometimes in conjunction with the philosophy of science) multiplied; the 1992 *Guide to the History of Science* lists 188 programs in 25 countries, led by the USA with 72 institutions granting doctorates in the subject.

By no means all history of science is done under the auspices of university departments or published in journals expressly devoted to the subject. The disciplinary affinities of the history of science are many and mutable, with distinctive national differences. Until the 1970s the primary connection, both in Europe and North America, was probably with the sciences themselves. This connection was rooted deep in the origins of the field and sustained by the continuing importance of some form of the history of science in the intellectual and social initiation of young scientists, as well as by the recruitment of historians of science from the ranks of those with scientific training. In the UK and North America, history largely replaced philosophy during the 1980s and 1990s as the discipline to which historians of science felt most akin, although interactions with sociology and, more recently, anthropology have also exerted a powerful influence. In France, Germany, and Italy, philosophy has remained a significant framework for the history of science, in diverse forms ranging from neo-Kantianism to logical empiricism. In addition to its disciplinary polyvalence with respect to choice of problems and methods, the history of science is closely linked by its subject matter to the histories of medicine and technology, albeit to differing degrees depending on the historical period in question. The phrase 'science and technology studies' bears witness to these criss-crossing ties to other

disciplines, serving as an abbreviation for the conglomerate 'history, philosophy, sociology, and anthropology of science, medicine, and technology'—which, however cumbersome, accurately reflects the ecumenical perspective of many historians of science.

2. Narratives and Periodization

Because of its links to philosophy, sociology, and anthropology, as well as to the natural sciences and mathematics, the history of science has been highly theorized for centuries. Theories about what scientific knowledge is, how it develops over time, where and when it expands and contracts, and in which ways it interacts with extrascientific contexts have profoundly shaped the way in which the history of science has been periodized and narrated. The more recent developments (since about 1970) in the historiography of science are incomprehensible without some background concerning the dominant theoretical schemes that had molded the field previously.

All history is inevitably written to some extent from the standpoint of the present, but present science—its orthodoxies and heresies, its controversies and critics—has exerted a mighty influence on the history of science since the Enlightenment. The great bulk of the history of science has been written explicitly or implicitly to illuminate, justify, or criticize the current state of science. In the Enlightenment histories, e.g., Jean-Étienne Montucla's *Histoire des mathématiques* (1758) or Jean-Baptiste Delambre's *Rapport historique sur les progrès des sciences mathématiques depuis 1789* (1810), the aim was to demonstrate that at least some of the sciences had shown steady and irreversible progress, especially since the Renaissance, and that the history of science was therefore a more edifying genre than the chronicles of human folly recounted by political and military history. The notion of progress was originally applied to art (by Giorgio Vasari in 1550) and to technology (by Francis Bacon in 1620), not to science. Once this interpretation of scientific change took hold in the eighteenth century, however, science became the premier example of progressive advance and scientific progress; a criterion by which to judge the comparative worth of civilizations. Whatever may have been the cultural achievements of the Latin Middle Ages or the Chinese empire, the stagnation of the sciences in these milieux, so ran the argument of the Enlightenment historians, revealed fundamental intellectual and social limitations—limitations that in the nineteenth and twentieth centuries came to be described as characteristic of 'traditional' or 'premodern' societies. Hence narratives of scientific progress eventually became part and parcel of larger narratives about modernization and European superiority.

Although the assumptions underlying the Enlightenment vision of scientific progress are seldom invoked

explicitly by contemporary historians of science, the reigning periodization of the field and the definition of specialties as central or peripheral still bears the traces of the traditional/modern division. Despite their glaring differences in many substantive aspects, the study of science in antiquity, the Arabic and Latin Middle Ages, and in non-Western civilizations such as China and India has often been amalgamated into a shared problematic: to what extent did the knowledge of these highly disparate cultures anticipate/approximate the science of post-Renaissance Europe? And to the extent that it did not, what were the intellectual and social obstacles that blocked the way (e.g., Huff 1993)? Hence the pioneering works in these areas (e.g., Duhem 1908, Maier 1949–58, Clagett 1959, Needham et al. 1954) foregrounded possible forgotten contributions to, and missing preconditions for, the Scientific Revolution.

Knowledge traditions that did not lead to modern science, understood to commence in early modern (a term as saturated with hindsight as ‘the Middle Ages’) Europe, were generally relegated to the periphery of the discipline, whatever their intrinsic interest and however excellent the scholarship devoted to them. Conversely, the Scientific Revolution has remained oddly central to the historiography of science: oddly, because recent scholarship on the sixteenth and seventeenth centuries has emphasized the heterogeneity and regional variations in the developments in anatomy, astronomy, natural history, mechanics, chemistry, mathematics, and physics once summarily lumped together as ‘the Scientific Revolution’ (Lindberg and Westman 1990, Porter and Teich 1992, Shapin 1996, Osler 2000), and because the steep increase of studies of nineteenth- and twentieth-century science since about 1980 has highlighted the novelty and significance of post-1800 developments, as the steady increase in the number of entries under these rubrics in the *Isis Current Bibliography* bears witness.

Yet the Scientific Revolution retains much of the original status first granted to it by the Enlightenment historians as the mold in which modernity was cast, one of the few points upon which both the apostles of progress and the critics of modernity agree. This periodization has left traces in the division of sub-specialties within the history of science, which tend to be chronological or geographic (‘ancient,’ ‘Southeast Asian’) for knowledge traditions prior to, or uninfluenced by, the Scientific Revolution, but disciplinary (‘physical sciences,’ ‘biomedical sciences’) for those following that original event—even for disciplines (e.g., astronomy) with histories that long antedate the sixteenth century and with geographical distribution well beyond Europe.

The progress of the sciences (especially mathematics and the exact sciences) set forth by the Enlightenment historians was cumulative but not revolutionary. In their accounts, knowledge steadily expanded around a

stable core of permanent truths, which may have taken millennia to discover, but which were thereafter firmly established beyond challenge. The language of ‘revolutions’ declined among early nineteenth-century scientists, perhaps in order to distance the permanent improvements of natural knowledge from the brusque reversals of politics; instead, the language of ‘law’ (e.g., Comte’s ‘Law of Three Stages’) was applied to the historical development of science (Cohen 1985). In the middle decades of the nineteenth century, however, this view of durable, inexorable scientific progress gave way to one of the rapid succession of scientific theories, even in areas (e.g., optics) long believed to rest upon unshakable foundations. Scientific progress had become vertiginous and open-ended; today’s truth might be discredited as tomorrow’s error. Under these conditions, it became increasingly difficult for historians of science to portray the steady growth of knowledge as culminating in present triumphs, since the endpoint was constantly changing. Instead, the history of science turned critical and conceptual: in the hands of Mach, Ostwald, and other scientist-historians, history of science revealed the past roots of present incoherences and taught the mutability of scientific theories. Within the revised interpretation of scientific progress, the purpose of the history of science was to use past science to change, not confirm, present science. Despite the dramatic shift in the narrative of scientific progress and in the goals of the history of science, its presentist orientation remained unchanged.

Almost all the strands of early twentieth-century historiography of science can be seen as attempts to come to grips with the peculiar character of scientific knowledge, which seemed to be simultaneously mutable and cumulative, driven by an internal dynamic and yet sensitive to ambient social conditions. These attempts often tugged in opposite directions. For example, the logical positivist program that stamped the history and philosophy of science in the middle decades of the twentieth century, especially in the modified form imported from the Vienna Circle to the UK and the USA, fixed upon the empirical findings of science as its enduring core in the Heraclitean flux of hypotheses and theories. These ‘protocol statements’ (Carnap 1928) were likened to simple observations, bringing science into the neighborhood of common sense. In contrast, the French historian and philosopher Gaston Bachelard, drawing upon the insights of psychoanalysis, insisted upon the necessity of overcoming everyday experience and intuitions in the formulation of successful scientific concepts (Bachelard 1938).

A many-cornered debate was ignited by the attempts of Marxist historians to relate the rise of modern science to specific socioeconomic conditions (Hessen 1931, Bernal 1939), out of which emerged the opposition of ‘internal’ to ‘external’ approaches to the history of science (Kuhn [1968] 1972, Shapin 1992). External approaches found a non-Marxist model in

Robert K. Merton's influential study of how science came to be valued in seventeenth-century England (Merton 1938, Shapin 1988). Informed by German and French neo-Kantian traditions in the history of philosophy, Alexandre Koyré's studies of the development of science in the broader intellectual context of early modern Europe stamped the work of historians of science, especially in France and the USA, trained in the 1950s and 1960s (Koyré 1939–40, 1961, Jorland 1981).

Common to all of these otherwise highly diverse perspectives on the history of science was an engagement with contemporary science and its lineage. Whether positivist or psychoanalytic, externalist or internalist, the history of science took the form of a genealogy leading to ideas and discoveries canonized by contemporary science. However much they may have disagreed on historical methods and forms of explanation, the combatants agreed on what and who in the history of science was worthy of explanation, namely that which had turned out to be pregnant with the future. In these histories a potent telos was at work, making present science appear to be not only the actual but also the inevitable outcome of past developments.

3. *Developments Since 1970*

Thomas S. Kuhn's *The Structure of Scientific Revolutions* (Kuhn [1962] 1970) broke decisively with this historiography, and stimulated new trends (not all to Kuhn's liking) in the history, sociology, and philosophy of science that have transformed the understanding of science and its growth in these fields since 1970. Kuhn argued for a plurality of scientific revolutions, in addition to the Scientific Revolution, which alternate with periods of 'normal science' framed by a 'paradigm' that models what can count as an acceptable problem, method, and solution in a given science in a specified period. It was less, however, Kuhn's 'structures' or patterns of long-term scientific development that exercised a deep and enduring influence on the history of science (although this feature of the book dominated its reception in other disciplines, particularly the social sciences) than his frontal attack on progressive narratives, his sketchy but evocative notion of 'paradigm' that embraced practices as well as theories, and the thoroughgoing historicism of many of his examples. All of these lines of thought were further cultivated in the history, philosophy, and sociology of science (the three disciplines were drawn into intensive interaction in part through their differentiated reception of Kuhn's book) during the subsequent three decades, often with consequences that led in directions diverging markedly from those plotted in *The Structure of Scientific Revolutions*.

Kuhn's rejection of progressive narratives in the history of science derived from his view that successive paradigms established and later overthrown by scientific revolutions were, first, internally coherent, and second, in a strong sense incommensurable with one another. This implied that science could not be cumulative in any progressive sense, one paradigm extending and building upon its predecessors. Kuhn's claim that conflicts between rival paradigms were not necessarily resolved by paradigm-independent criteria (e.g., predictive accuracy of explanatory unity) unsettled the inevitability of scientific development still further. Sociologists of science Barry Barnes and David Bloor took these claims as the basis of an approach called 'constructivism' (Golinski 1998), in which the development of scientific knowledge was to be explained without appeal to its truth or falsehood, on the grounds that the truth of a proposition is neither a necessary nor sufficient condition for its acceptance (Bloor [1976] 1991).

For historians of science, this program proved stimulating in at least three respects. First, in conjunction with Kuhn's focus on periods of conflict between rival paradigms, it prompted detailed studies of how scientific controversies were in fact resolved. Second, it enlarged the definition of what could count as 'science' for the purposes of historical investigation by ruling out hindsight judgments about what was and was not legitimate by present standards (to the benefit of specialties such as the history of premodern science and of the human sciences). Third, it encouraged a broader exploration of the grounds for scientific conviction by enforcing a 'symmetric' treatment of reasons for embracing false as well as true beliefs. This deliberate refusal to adopt the perspective of present science in the investigation of past knowledge went under the name of 'anti-Whiggery' among historians of science (in allusion to Herbert Butterfield's critique of 'Whig' political history). Among sociologists of science, the symmetry principle led to an increasingly distanced approach even to present science, to the point of adopting the ethnographer's estranged perspective (Latour and Woolgar [1979] 1986). The overall impact of the constructivists in both sociology and history of science was to highlight the contingency of the development of scientific knowledge, in stark contrast to earlier narratives of inexorable progress (Pestre 1995, Rheinberger 1997).

Kuhnian paradigms drew upon the philosophical reflections of Ludwig Wittgenstein (Wittgenstein 1953) and Michael Polanyi (Polanyi 1958) to stress the importance of knowledge that could not be reduced to rules or algorithms or even to words in a normal scientific tradition. In the history, sociology, and philosophy of science a rich literature on 'scientific practices' has taken the inadequacy of the formal aspects of theory to explain the distinctive aspects of scientific knowing as its departure point. Although scientific practices may embrace everything from a

knack for integrating equations to the deft adjustment of finicky instruments, studies of experiment have taken on an exemplary character in this branch of the field. Interactions among historians (e.g., Shapin and Schaffer 1985, Galison 1987, Gooding et al. 1993), philosophers (e.g., Cartwright 1983, Hacking 1983), and sociologists (e.g., Collins 1985) created a discussion that blended epistemological, historical, and social analyses.

Experiment and scientific practices more generally have emerged as autonomous forms of knowledge, no longer subsidiary to and in lockstep with theory, that depend crucially on local factors (including skill, material culture, values) that travel only with difficulty. This emphasis upon the locality of scientific knowledge has spawned a new set of problems: How is local science made global? In the earlier historiography, science had been assumed to be universal because nature was so. For historians of scientific practice, in contrast, further explanation is required to show the ways in which localized achievements can be made at least partially communicable among practitioners divided by geography, history, or culture (Schaffer 1992, Galison 1999). Sociologists of science have attacked this problem from the standpoint of 'actor-networks' (Latour 1987).

Both the rejection of presentist perspectives and the close study of scientific practices have pushed historians of science toward a more thoroughly historicist approach. Whereas their predecessors practiced anachronism openly and with conviction, on the view that past science must be understood by the lights of present science, most historians of science now writing insist upon some attempt to reconstruct the conceptual and cultural categories of historical actors. This historicism has expressed itself in two important movements in recent historiography of science: the aim to understand 'science in context,' and attempts to write histories of entities previously assumed to be transhistorical.

'Science in context' (which lent its name to a journal) thickens the connections of science with its historical place and time, as well as blurring the boundaries between factors previously regarded as either internal or external to a discipline. Indeed, one of the consequences of the studies of science in context has been to make the emergence, persistence, and occasional disappearance of such boundaries, topics of historical investigation in their own right. One of the most important of the new strands of contextual history of science, because of its potentially wide application, has been the investigation of how ideas and practices of gender have shaped the modern sciences (e.g., Haraway 1989, Jordanova 1989, Schiebinger 1993).

Although some critics of this embedded view have feared that context impugns scientific autonomy and therefore validity, many recent studies have depicted the social, technological, and economic milieu of science as resources for scientific creativity, not merely

as ideological contaminants (e.g., Wise 1989–90). Histories of supposedly transhistorical entities such as experience (Dear 1995), truth (Shapin 1994), and objectivity (Daston and Galison 1992, Porter 1995) make use of the tools of science in context to anchor these Kantian abstractions in specific historical circumstances; they are also at least indirectly indebted to the remarkable studies of Georges Canguilhem (Canguilhem [1943] 1966) and, especially, Michel Foucault (Foucault 1969) that challenged the universality and permanence of fundamental modern categories like normalcy and sexuality.

The most recent developments in the history of science might be briefly described as materialized and embodied. The material culture of science, ranging from the architecture of the spaces in which science is done (Rheinberger et al. 1996, Galison and Thompson 1999) to the machines that do the science, have shifted from the periphery to the center of historical attention. Hands-on experience with objects is no longer understood as a means to a greater end, but as a force that shapes the culture and penetrates into the metaphysics of science (Galison 1997). Embodiment has entered the history of science through several doors: the study of the interaction of experiment and craft traditions (Sibum 1995), the investigation of the scientific persona in historical context (Lawrence and Shapin 1998, Shortland and Yeo 1996), and the making visible of the role of women in science (Rossiter 1982, Rossiter 1995). All of these explorations of embodiment challenge the Cartesian image of the disembodied knower, as well as the hagiography of great individuals; they aim at the reconstruction of collective identities of scientists, as well as their consequences for who was allowed to do what kind of science.

Given the profusion of topics and approaches in the history of science, it is understandable that one historian might wonder 'whether the history of science is a coherent discipline or just a collection of scholars aggregated by the accidents of history and the accretion of a common historiography' (Rosenberg 1988, p. 570). This is a doubt that in large part revolves around the difficulty of defining the *sine qua non* of scientific knowledge, if it has one. Yet in this still somewhat undisciplined discipline, the reward of permeable boundaries has been a rare dialogue, ongoing and sometimes heated, among the natural sciences, social sciences, and humanities.

See also: Cognitive Psychology: History; Comte, Auguste (1798–1857); Cultural Studies of Science; Enlightenment; History of Science: Constructivist Perspectives; History of Technology; Knowledge, Sociology of; Kuhn, Thomas S (1922–96); Medicine, History of; Norms in Science; Objectivity, History of; Paradigms in the Social Sciences; Physical Sciences: History and Sociology; Progress: History of the Concept; Science and Religion; Science and Tech-

nology, Anthropology of; Science, Sociology of; Scientific Disciplines, History of; Scientific Revolution: History and Sociology; Situated Knowledge: Feminist and Science and Technology Studies Perspectives; Sociology, History of; Statistics, History of; Technology, Social Construction of; Truth and Credibility: Science and the Social Study of Science; Wittgenstein, Ludwig (1889–1951)

Bibliography

- Bachelard G 1938 *La Formation de l'Esprit Scientifique*. Vrin, Paris
- Bernal J D 1939 *The Social Function of Science*. Macmillan, New York
- Bloor D [1976] 1991 *Knowledge and Social Imagery*. University of Chicago Press, Chicago
- Canguilhem G [1943] 1966 *Le Normal et le Pathologique*. Presses Universitaires de France, Paris
- Carnap R 1928 *Die logische Aufbau der Welt*. Weltkreis-Verlag, Berlin
- Cartwright N 1983 *How the Laws of Physics Lie*. Clarendon Press, Oxford, UK
- Clagett M 1959 *The Science of Mechanics in the Middle Ages*. University of Wisconsin Press, Madison, WI
- Cohen I B 1985 *Revolutions in Science*. Harvard University Press, Cambridge, MA
- Collins H 1985 *Changing Order: Replication and Induction in Scientific Practice*. Sage, Beverly Hills, CA
- Daston L, Galison P 1992 The image of objectivity. *Representations* 40: 81–128
- Dear P 1995 *Discipline and Experience: The Mathematical Way in the Scientific Revolution*. University of Chicago Press, Chicago
- Duhem P 1908 *ΣΩΖΕΙΝ ΤΑ ΦΑΙΝΟΜΕΝΑ: Essai sur la Notion de Théorie Physique de Platon à Galilée*. A. Hermann et Fils, Paris
- Foucault M 1969 *L'Archéologie du Savoir*. Éditions Gallimard, Paris
- Galison P 1987 *How Experiments End*. University of Chicago Press, Chicago
- Galison P 1997 *Image and Logic: A Material Culture of Microphysics*. University of Chicago Press, Chicago
- Galison P 1999 Trading zone: Co-ordinating action and belief. In: Biagioli M (ed.) *The Science Studies Reader*. Routledge, New York, pp. 137–60
- Galison P, Thompson E 1999 *The Architecture of Science*. MIT Press, Cambridge, MA
- Gillispie C C (ed.) 1981 *Dictionary of Scientific Biography*. Charles Scribners' Sons, New York
- Golinski J 1998 *Making Natural Knowledge: Constructivism and the History of Science*. Cambridge University Press, Cambridge, UK
- Gooding D, Pinch T, Schaffer S (eds.) 1993 *The Uses of Experiment: Studies in the Natural Sciences*. Cambridge University Press, Cambridge, UK
- Graham L, Lepenies W, Weingart P (eds.) 1983 *Functions and Uses of Disciplinary History*. D. Reidel, Dordrecht, The Netherlands
- Hacking I 1983 *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press, Cambridge, UK
- Haraway D J 1989 *Primate Visions: Gender, Race, and Nature in the World of Modern Science*. Routledge, New York
- Hessen B 1931 The social and economic roots of Newton's 'Principia.' In: Bukharin N I (ed.) *Science at the Crossroads*. Kniga, London, pp. 146–212
- History of Science Society 1967 *Isis Current Bibliography* (originally *Isis Critical Bibliography*). University of Chicago Press, Chicago
- Huff T E 1993 *The Rise Of Early Modern Science: Islam, China, and the West*. Cambridge University Press, Cambridge, UK
- Jordanova L 1989 *Sexual Visions: Images of Gender in Science and Medicine between the Eighteenth and Twentieth Centuries*. University of Wisconsin Press, Madison, WI
- Jorland G 1981 *La Science dans la Philosophie: Les Recherches Épistémologiques d'Alexandre Koyré*. Éditions Gallimard, Paris
- Koyré A 1939–40 *Études Galiléennes*. Hermann, Paris
- Koyré A 1961 *La Révolution Astronomique: Copernic–Kepler–Borelli*. Hermann, Paris
- Kragh H 1987 *An Introduction to the Historiography of Science*. Cambridge University Press, Cambridge, UK
- Kuhn T S [1962] 1970 *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago
- Kuhn T S [1968] 1972 History of science. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan and Free Press, London and New York, Vol. 13, pp. 74–83
- Latour B 1987 *Science in Action: How to Follow Scientists and Engineers through Society*. Harvard University Press, Cambridge, MA
- Latour B, Woolgar S [1979] 1986 *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press, Princeton, NJ
- Laudan R 1993 Histories of the sciences and their uses: A review to 1913. *History of Science* 31: 1–34
- Lawrence C, Shapin S (eds.) 1998 *Science Incarnate: Historical Embodiments of Natural Knowledge*. University of Chicago Press, Chicago
- Lindberg D, Westman R (eds.) 1990 *Reappraisals of the Scientific Revolution*. Cambridge University Press, Cambridge, UK
- Maier A 1949–58 *Studien zur Naturphilosophie der Spätscholastik*. Edizioni di Storia e Letteratura, Rome
- Merton R K 1938 Science, technology and society in seventeenth-century England. *Osiris* 4: 360–632
- Needham J et al. 1954 *Science and Civilization in China*. Cambridge University Press, Cambridge, UK
- Osler M J (ed.) 2000 *Rethinking the Scientific Revolution*. Cambridge University Press, Cambridge, UK
- Pestre D 1995 Pour une histoire sociale et culturelle des sciences: nouvelles définitions, nouveaux objets, nouvelles pratiques. *Annales: Histoire, Sciences Sociales* 50: 487–522
- Polanyi M 1958 *Personal Knowledge: Towards a Post-critical Philosophy*. University of Chicago Press, Chicago
- Porter T M 1995 *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press, Princeton, NJ
- Porter R, Teich M (eds.) 1992 *The Scientific Revolution in National Context*. Cambridge University Press, Cambridge, UK
- Rheinberger H J 1997 *Towards A History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Stanford University Press, Stanford, CA
- Rheinberger H J, Hagner M, Wähgig-Schmidt B (eds.) 1997 *Räume des Wissens: Repräsentation, Codierung, Spur*. Akademie-Verlag, Berlin

- Rosenberg C 1988 Woods or trees? Ideas and actors in the history of science. *Isis* 79: 565–70
- Rossiter M W 1982 *Women Scientists in America: Struggles and Strategies to 1940*. Johns Hopkins University Press, Baltimore
- Rossiter M W 1995 *Women Scientists in America: Before Affirmative Action, 1940–1972*. Johns Hopkins University Press, Baltimore
- Sarton G [1927] 1948 *Introduction to the History of Science*. Williams & Wilkins, Baltimore, MD
- Schaffer S 1992 Late Victorian metrology and its instrumentation: A manufactory of ohms. In: Bud R, Cozzens S (eds.) *Invisible Connections: Instruments, Institutions, and Science*. SPIE Optical Engineering Press, Bellingham, WA, pp. 23–56
- Schiebinger L 1993 *Nature's Body: Gender in the Making of Modern Science*. Beacon Press, Boston
- Shapin S 1988 Understanding the Merton thesis. *Isis* 79: 594–605
- Shapin S 1992 Discipline and bounding: The history and sociology of science as seen through the externalism–internalism debate. *History of Science* 30: 333–69
- Shapin S 1994 *A Social History of Truth: Civility and Science in Seventeenth-century England*. University of Chicago Press, Chicago
- Shapin S 1996 *The Scientific Revolution*. University of Chicago Press, Chicago
- Shapin S, Schaffer S 1985 *Leviathan and the Air-pump: Hobbes, Boyle, and the Experimental Life*. Princeton University Press, Princeton, NJ
- Shortland M, Yeo R (eds.) 1996 *Telling Lives in Science: Essays on Scientific Biography*. Cambridge University Press, Cambridge, UK
- Sibum H O 1995 Reworking the mechanical value of heat: Instruments of precision and gestures of accuracy in early Victorian England. *Studies in the History and Philosophy of Science* 26: 73–106
- Thackray A, Merton R K 1972 On discipline building: The paradoxes of George Sarton. *Isis* 63: 473–95
- Wise M N 1989–90 Work and waste: political economy and natural philosophy in nineteenth century Britain. *History of Science* 27: 263–301, 391–449; 28: 221–61
- Wittgenstein L 1953 *Philosophical Investigations* [trans. Anscombe G E M]. Blackwell, Oxford, UK

L. Daston

History of Science: Constructivist Perspectives

The last quarter of the twentieth century saw the emergence of an important and challenging new perspective on the sciences and their history. Drawing largely on developments in the social studies of science, but also on new trends in labor and social history, historians began to look at science as local practice, rather than as an abstract, disembodied, and universal set of ideas. Rather than seeing science as somehow outside culture, historians started looking at science as part of culture, locally constituted in particular material,

social, and even political practices. Constructivist historians are interested in the detailed processes of scientific knowledge production, dissemination, and reception. They differ from more traditional social historians of science in their insistence that the content as well as the context of science is a proper focus of cultural historical analysis. Crucially, constructivist historians see science as being a species of skilled practice and therefore inquire into the cultural origins of those skills, the means through which they are produced, validated and disseminated, and the ways in which particular forms of inquiry—and particular kinds of knowledge—have historically been characterized as scientific.

1. Local Cultures

In their exemplary study of the early modern culture of experiment, Shapin and Schaffer lay out some of the central concerns of the constructivist history of science (Shapin and Schaffer 1985). Focusing on the dispute between Robert Boyle and Thomas Hobbes concerning the validity of Boyle's famous air-pump experiments, they use Hobbes's critique of Boyle's experiments to lay bare the interests underlying them and the strategies that Boyle and his allies adopted to convince others of the experiments' reliability. They demonstrate how even such an apparently universal cornerstone of the scientific method as experiment has its origins in particular local and contingent circumstances. Making science local and examining the interests that underlie the articulation and resolution of scientific debates have been preoccupations of constructivist history. Other pioneering studies include accounts of the ways that disillusionment with rational philosophy and mechanistic notions of causality in postwar Weimar Germany informed interpretations of quantum theory (Forman 1971), and of the role social interests played in debates concerning the anatomical structure of the brain in early nineteenth-century Edinburgh (Shapin 1979). Such studies emphasize the way in which such interests are constitutive rather than contingent features of even the most apparently abstract and technical forms of knowledge.

One important outcome of constructivist history is the recognition that even the issue of what kind of activity might count as science is often the product of locally articulated debates and controversies. Not only did Boyle and Hobbes disagree as to what might reliably be inferred from the outcome of air-pump experiments, but they also disagreed as to whether or not doing such experiments should be counted as a reliable way of generating scientific knowledge at all. Similarly, the scientific and the medical establishment who opposed the introduction of continental ideas concerning evolution and transmutation into early

Victorian England did so both because they did not recognize the practices or the practitioners involved as being properly scientific and because they disagreed with the theories themselves (Desmond 1989). As the labels 'science' and 'scientific' gained in cultural value, so too did different groups of knowledge producers disagree more vociferously as to what kinds of people and what kinds of activity should be accorded those labels (Gieryn 1999). Constructivist historians have identified the business of policing the constantly changing boundaries of science as a crucial part of science's history. Through institutions, regimes of training, control of resources, and access to networks of patronage, natural philosophers and scientists have struggled to circumscribe and discipline their own and others' knowledge producing activities.

This localist historical perspective leads to recognition of the extent to which the production and validation of scientific knowledge are contingent upon more or less informal networks of trust. In order for particular groups or individuals to be considered reliable spokespersons for nature, they have to be trustworthy. They have to be the kinds of people who can be relied upon to tell the truth. This opens up the possibility of producing a cultural history of truth and objectivity. The marks of being trustworthy vary according to time and place. In early modern England for example, trustworthiness was considered the hallmark of the gentleman (Shapin 1994). As a result, many of the informal (and more formal) rules of early modern gentlemanly behavior became enshrined as codes of conduct for the nascent scientific community as well. Some historians have argued that the increased scientific reliance on precision measurement and self-registering instrumentation since the beginning of the nineteenth century may be regarded as efforts to circumvent the need for trusting individuals with knowledge production (Schaffer 1988). Others have pinpointed the increasing quantification of science as a means of overcoming the problem of trust in the production and particularly in the transfer of knowledge (Porter 1995).

2. The Material Culture of Experiment

The constructivist history of science pays attention not only to experiment as a form of inquiry, but also to the process of experiment. Experimentation is taken to be a matter of human agents actively engaged in constructing the natural world around them (Gooding et al. 1989). From this perspective, the material culture of experiment becomes a crucial focus of attention. Historians have worked to recover the kinds of material resources required for successful experimentation. They have traced the emergence of the laboratory as a specialized space within which experiments take place. They have emphasized the ways in which successful manipulation of nature requires the

experimenter to manipulate material and social resources successfully as well (Latour 1988). Michael Faraday's discovery of electromagnetic induction, for example, involved as much his marshalling of the Royal Institution's considerable resources as it was the outcome of his artful manipulations in the laboratory (Morris 1998). Constructivist historians have identified as well the important ways in which individual scientific reputations hinge upon recognition of the property relationship between experiment and experimenter. Priority disputes in science are frequent and virulent precisely because of the ways in which experiments embody the skills and reputations of their makers (Iliffe 1992).

Seeing experiment as a skilled activity, constructivist historians have focused much attention on understanding the historical origins of the kinds of skills required in order to carry out experiments. They have investigated how particular skills have been transferred into the laboratory from other settings, such as workshops or factories. Historical experiments have been replicated in an effort to better understand the tacit, unarticulated skills required for their successful performance. Such studies have shown how iconic experiments such as James Prescott Joule's famous paddle-wheel experiments establishing the mechanical equivalent of heat during the 1840s, were contingent upon the technical know-how that Joule and his workers derived from their experiences in the family brewery (Sibum 1995). Increasingly also, historians have looked at the politics of skill in the laboratory and elsewhere, investigating the division and distribution of hierarchies of skilled activity within science at different times and places. Much attention has been devoted to the historical transmission of material resources and skilled personnel that accompanies the standardization of experimental practices. Work in this vein has shown how experiments move from being instances of highly skilled, difficult to replicate, and delicate procedures performed in highly circumscribed settings to being robust, standardized, and readily reproducible embodiments of group knowledge (Galison 1987).

Apparently self-evident features of experimentation, such as precision measurement, are similarly shown by constructivist historians to have a history. Far from being a routine and unproblematic aspect of scientific activity, measurement can be taken to be an expression of particular cultural values (Wise 1995). The rise of metrology, for example, was an expression of science's increasing embeddedness in and contribution to a rapidly bureaucratizing and industrializing Western culture. Measurement could be regarded not only as a means of disciplining nature, rendering it more amenable to standardized routines of inquiry, but also as a means of disciplining the scientific workforce. Nineteenth-century scientific managers such as George Biddell Airy, the English Astronomer Royal, regarded their laboratories and observatories as

industrial workplaces and used regimes of routinized measurement and calculation to impose an industrial discipline on their workforces (Smith 1991). In late nineteenth century laboratories at Cambridge or Glasgow, natural philosophers such as James Clerk Maxwell or Lord Kelvin saw the imposition of regimes of precision measurement in teaching and research as an expression of their affinity with industrializing society (Schaffer 1992). Increasingly, constructivist historians have investigated how adherence to cultural values like precision came to be taken as a hallmark of what it meant to be properly scientific. More recently, some historians have returned to matters of theory, using the analytical tools developed to understand experiment to look at theory as a form of cultural practice as well. From this perspective, processes of theory construction and dissemination are seen in terms of the pedagogical training and practical skills required to participate in highly specialized theoretical cultures (Galison and Warwick 1998).

3. Science as Performance

One notable outcome of constructivist historical work has been to make the identity of the scientific practitioner itself problematic. Being a scientist is not a self-evident activity: the word itself is of comparatively recent, early nineteenth century origin. Scientific practitioners in the past occupied a number of different cultural niches. Crucially, constructivist historians have shown how natural philosophers at different times and places have been engaged actively in carving out social spaces for themselves and their work. Galileo's position at the Medici court in Renaissance Florence, for example, was not ready made. Galileo had to work hard to promote the idea of a philosopher or mathematician courtier that he could then fulfill (Biagioli 1993). He had to fashion himself into the kind of person who could easily inhabit that cultural niche and then to orient his work so that it would fit into the milieu of Italian Renaissance court life. The same could be said of Humphry Davy and his efforts to forge a place for the fashionable natural philosopher in Georgian London (Golinski 1992). Constructivist approaches share with the 'new history' an emphasis on the important role of individual agents' artful engagement in constructing both themselves and their worlds.

Recent constructivist work has paid particular attention to the body of the scientific practitioner (Lawrence and Shapin 1998). This is an outgrowth of the constructivist insistence on viewing scientific practices and the knowledge they produce as being in crucial ways embodied. Scientific practitioners at different times and places have needed to be particular kinds of human beings with specific bodily attributes. The ways in which those attributes were held to be best produced—whether by birth, self-discipline, or

rigorous training—have been a focus of increasing historical attention. Historians have examined, for example, the emergence of particular regimes of mental training in Victorian Cambridge and the ways in which candidates for mathematical honors were encouraged to train their bodies into being fit receptacles for their minds (Warwick, in Lawrence and Shapin 1998). The bodily interaction of practitioners with their experimental apparatus and the spatial organization of their laboratories have come to be recognized as playing an important role in the production of knowledge. The organization of such spaces and the kinds of material interactions between practitioners, and between them and their instruments, that have been taken to produce reliable knowledge are now seen as contingent and open to historical investigation.

Not only the identity and activities of the practitioner producing knowledge, but also the role of the audience for such productions are now recognized as constitutive. Audiences for scientific knowledge do not simply act as passive receivers of settled and unproblematic matters of fact. They have an important and active role to play in the articulation and resolution of controversies. During the early modern period and well into the nineteenth century, experiments were often performed before an audience. This was not simply a matter of demonstration. Members of the audience—prestigious and often named individually—played an important role in validating the knowledge produced. Their reputations could be used to underwrite the matters of fact they had witnessed. Audiences have participated in and responded to the processes of knowledge production through a variety of media, from direct witnessing to the virtual witnessing of reading a scientific paper at first hand, to the reception of second or third-hand reports in museums, newspapers, reviews, or popular handbooks. In the hands of states, public instrumental uses of science and technology have become devices for legitimating political power (Ezrahi 1990). Public scientific performances are an important part therefore of the dynamic process of knowledge-production and validation.

4. Science at Work

Constructivist historical analysis looks as well at the many ways in which nature has been used historically as a model for social organization. The constitution of the natural economy has frequently been used as a powerful argument in favor of particular ways of ordering society. Natural theology in the late eighteenth and early nineteenth centuries, for example, was widely used as an argument in favor of existing social hierarchies. Evolutionary theories were similarly cited to support new, meritocratic visions of society. The proper social economy was often held to mirror the natural order. Constructivist historians have shown

how models of nature and society in the past were frequently mutually sustaining. They were produced together. Discussions in mid-nineteenth century thermodynamics surrounding issues of work and waste in natural economies were coextensive with arguments concerning the proper organization of work and the minimization of waste in political economy (Smith and Wise 1989). Proponents of new forms of industrial organization, for example, looked to nature for their models. At the same time, scientists looked to factory organization as a model for the best means of ordering the search for truth.

In turn, different models for the organization of scientific work often expressed different understandings not only of how nature was ordered but also of the proper politics of knowledge production. Nineteenth-century scientific gentlemen, for example, regarded scientific work as a highly disciplined, hierarchical, and increasingly specialized activity. In principle, it was open to all. In practice, however, it could only be conducted properly either by, or under the supervision of, properly trained and vocationally committed 'gentlemen of science' (Rudwick 1985). Other social groups had different notions of the politics of knowledge. The gentlemen's vision excluded women and artisans, for example, from anything other than a supportive role in making knowledge. Working-class men, on the other hand, could construct their own very different accounts of how knowledge making ought to be organized (Secord 1994).

Through constructivist historical analysis, many of the traditional analytical categories employed by historians and sociologists of science have become recognized as being themselves in need of careful historical contextualization. Pronouncements by historical actors concerning matters such as the scientific method, experimentation, objectivity, or even more novel analytical categories such as skill, are not to be taken at face value. Such pronouncements usually carry considerable weight. The issue of what kind of activity might properly be considered as carrying out an experiment, for example, is not straightforward. At different times, experimentation has been regarded as both a public and a private process. It has been characterized as both a fundamentally communal and an essentially individual activity. What it means to be a skilled worker, again, has changed a great deal since the early modern period. Skill, like method, or even science itself, is a historical construction. By looking at science as a pre-eminently cultural activity, constructivist histories of science have made it possible to raise a whole host of new questions concerning the ways in which human beings, in the past, have tried to make sense of the world around them. Rather than seeing science as a static, monolithic edifice, constructivist history recognizes it as a constantly changing process of production, reproduction, and consumption, its boundaries continually shifting and its meaning always open to renegotiation.

See also: Constructivism/Constructionism: Methodology; Constructivism in Cognitive Psychology; History and the Social Sciences; History of Science; History: Overview; Scientific Knowledge, Sociology of; Universities, in the History of the Social Sciences

Bibliography

- Biagioli M 1993 *Galileo, Courtier: The Practice of Science in the Culture of Absolutism*. University of Chicago Press, Chicago
- Desmond A 1989 *The Politics of Evolution: Morphology, Medicine, and Reform in Radical London*. University of Chicago Press, Chicago
- Ezrahi Y 1990 *The Descent of Icarus: Science and the Transformation of Contemporary Democracy*. Harvard University Press, Cambridge, MA
- Forman P 1971 Weimar culture, causality, and quantum theory, 1918–1927: Adaptation by German physicists and mathematicians to a hostile intellectual environment. *Historical Studies in the Physical Sciences* 3: 1–115
- Galison P 1987 *How Experiments End*. University of Chicago Press, Chicago
- Galison P, Warwick A (eds.) 1998 *Cultures of Theory. Special Issue: Studies in History and Philosophy of Modern Physics*. Pergamon, Oxford, UK
- Gieryn T 1999 *Cultural Boundaries of Science: Credibility on the Line*. University of Chicago Press, Chicago
- Golinski J 1992 *Science as Public Culture: Chemistry and Enlightenment Britain, 1760–1820*. Cambridge University Press, Cambridge, UK
- Gooding D, Pinch T, Schaffer S (eds.) 1989 *The Uses of Experiment: Studies in the Natural Sciences*. Cambridge University Press, Cambridge, UK
- Illiffe R 1992 In the warehouse: Privacy, property and priority in the early Royal Society. *History of Science* 30: 30–68
- Latour B 1988 *The Pasteurization of France*. Harvard University Press, Cambridge, MA
- Lawrence C, Shapin S (eds.) 1998 *Science Incarnate: Historical Embodiments of Natural Knowledge*. University of Chicago Press, Chicago
- Morus I R 1998 *Frankenstein's Children: Electricity, Exhibition and Experiment in Early Nineteenth-century London*. Princeton University Press, Princeton, NJ
- Porter T 1995 *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press, Princeton, NJ
- Rudwick M 1985 *The Great Devonian Controversy: The Shaping of Scientific Knowledge among Gentlemanly Specialists*. University of Chicago Press, Chicago
- Schaffer S 1988 Astronomers mark time: Discipline and the personal equation. *Science in Context* 2: 115–45
- Schaffer S 1992 Late Victorian metrology and its instrumentation: A manufactory of ohms. In: Bud R, Cozzens S (eds.) *Invisible Connections: Instruments, Institutions and Science*. SPIE Optical Engineering Press, Bellingham, WA
- Secord A 1994 Science in the pub: Artisan botanists in early nineteenth century Lancashire. *History of Science* 32: 269–315
- Shapin S 1979 The politics of observation: Cerebral anatomy and social interests in the Edinburgh phrenology disputes. In: Wallis R (ed.) *On the Margins of Science: The Social Construction of Rejected Knowledge*. University of Keele Press, Keele, UK

- Shapin S 1994 *A Social History of Truth: Civility and Science in Seventeenth-century England*. University of Chicago Press, Chicago
- Shapin S, Schaffer S 1985 *Leviathan and the Air-pump: Hobbes, Boyle, and the Experimental Life*. Princeton University Press, Princeton, NJ
- Sibum H O 1995 Reworking the mechanical value of heat: Instruments of precision and gestures of accuracy in early Victorian England. *Studies in History and Philosophy of Science* **26**: 73–106
- Smith C, Wise N 1989 *Energy & Empire: A Biographical Study of Lord Kelvin*. Cambridge University Press, Cambridge, UK
- Smith R A 1991 National observatory transformed: Greenwich in the nineteenth century. *Journal of the History of Astronomy* **22**: 5–20
- Wise N (ed.) 1995 *The Values of Precision*. Princeton University Press, Princeton, NJ

I. R. Morus

History of Technology

Historians interested in technology have written articles and books about the subject for at least since 1900. Only since World War II, however, has the subject taken on some of the attributes of a discipline through the establishment of graduate programs and a scholarly journal of international reputation. History of technology includes the history of invention, the history of engineering, and social history of technology. Scholars with varied backgrounds, some formally trained in the history of technology, contribute to the field.

1. Defining the History of Technology

Because technology defies simple definition, its history as a field of study is not easily defined. Editors of the five-volume *History of Technology* (1954–1958, Oxford University Press, London) broadly define ‘technology’ as ‘how things are commonly done or made ... [and] what things are done and made.’ Finding this definition too diffuse, the editors of the two-volume *Technology in Western Civilization* (1967, Oxford University Press, New York,) reject, as well, the oft-encountered definitions of technology as applied science or as a rational attempt to control nature. They associate technology with human work, ‘with man’s attempts to satisfy his wants by human action on physical objects.’

Technology and its history are as difficult to define as history of politics or history of economics which experienced scholars refuse to define except by countless examples.

It is also hard to characterize a historian of technology. Unlike general historians, who usually

have graduate training in history departments, historians interested in technology enter upon the study of its history from a variety of backgrounds, besides graduate study in the history of technology. Persons trained, for example, in engineering, history of science, economic history, general history, and sociology, do outstanding works in the field. Professional writers also contribute works addressed to a general public including essays that appear in widely circulated journals such as the *American Heritage of Invention and Technology*.

2. Handbooks and Histories of Invention and Engineering

During the nineteenth and early twentieth centuries, scholars, especially German, nurtured the history of technology by publishing information-rich encyclopedias and handbooks. Franz Marie Feldhaus’s *Die Technik der Vorzeit, der Geschichtlichen Zeit und der Naturvölker* (1914, Engelman, Leipzig) and Ludwig Darmstaedter’s *Handbuch zur Geschichte der Naturwissenschaften und der Technik* (1908 Springer, Berlin) continue to be useful sources for the history of invention and discovery.

In the early years, biographies of engineers and histories of engineering appeared alongside histories of invention. Samuel Smiles, a railroad manager and medical doctor, in his influential *Lives of the Engineers* (1861–2, Murray, London,) cast problem-solving, project-managing engineers of the British Industrial Revolution in a heroic mold. They surmounted technical, political, economic, and organizational hurdles to transform undeveloped land into industrial regions by means of canals, roads, and improved harbors. In the twentieth century engineers have played a more enigmatic role than did Smiles’s heroes. Karl-Heinz Ludwig in *Technik und Ingenieure in Dritten Reich* (1979, Princeton University Press, Kronigstein), Kendall Bailes in *Technology and Society under Lenin and Stalin* (1978, Harvard University Press, Princeton, NJ), and Loren Graham in *The Ghost of the Executed Engineer: Technology and the Fall of the Soviet Union* (1993, Harvard University Press, Cambridge, MA) follow the tragic path of engineers under totalitarian regimes.

3. Early Monographs

Lewis Mumford, a public intellectual and professional writer, remains a seminal contributor to the field. Painting with broad-brush strokes on a canvas of western history, his approach is imaginatively interdisciplinary as he draws upon secondary works in western history, sociology, and anthropology. Prolific, he appealed to an interested, educated public, and

simultaneously influenced the scholarly world with his insightful concepts. *Technics and Civilization* (1934, Harcourt Brace, New York), his most enduring history of technology, surveys the development of technology from pre-historic to modern times. He acknowledges the deterministic role technology often plays in history, but he also stresses that values shape technology. His later works, including *The Myth of the Machine: The Pentagon of Power* (1970, Harcourt Brace Jovanovich, New York, NY), lament the rise of megamachines, or large bureaucratic, military, technological systems that deny individual's freedom and sweep them towards catastrophic wars.

Swiss architect and historian Siegfried Giedion's *Mechanization Takes Command* (1948, Oxford University Press, New York) ranges from medieval to modern times. A highly original study, it seeks to explain the mechanization of dwellings, food, furniture, and other aspects of everyday life. Anticipating the attitude of post-World War II historians of technology, he laments in particular the way in which mechanization has displaced organic processes and substituted abstract reason for feeling.

Louis C. Hunter's *Steamboats on the Western Rivers, an Economic and Technological History* (1949, Harvard University Press, Cambridge, MA), and Lynn White Jr.'s *Medieval Technology and Social Change* (1962, Clarendon Press, New York) also anticipate the contextual and social history approaches of a later generation of historians of technology. Hunter shows how the natural environment of the Mississippi River shaped the design of nineteenth century steamboats. On the other hand, White believed technology brought manorialism and feudalism (see *Technological Determinism*).

4. General Histories of Technology

The publication of the information-rich *A History of Technology* (1954–8, Oxford University Press, London) edited by Charles Singer, E. J. Holmyard, A. R. Hall, and Trevor Williams gave the field substantial scholarly credentials. Museum curators, engineers, and natural scientists figure prominently among the authors of the more than 100 chapters that tell of the development of technology from its beginnings to 1900. The authors do not integrate political, social, and economic history into their deterministic account of a nature-mastering, society-shaping, western technology that, they believe, had become science-based by 1900.

Published about a decade later, *Technology in Western Civilization* (1967, Oxford University Press, New York), another collective work, differs notably in approach from *A History of Technology*. Trained in general history before turning to history of technology, editors Melvin Kranzberg and Carroll Pursell, Jr., in

contrast to the editors of *A History of Technology*, urged their 34 authors to integrate history of technology with social, economic, and political history. The majority are historians of science, historians of technology, and economic historians, and they tend to take a contextual approach and to focus upon the British and American industrial revolutions. By stressing the interplay of technological and social factors, Kranzberg and Pursell move away from technological determinism encountered in *A History of Technology*.

Kranzberg and Pursell anticipate the recent trend among historians of technology to emphasize the interaction of technology and the natural environment by including three chapters on land use and resources. In a concluding chapter about the challenge of technology for the future, they also take a more critical view of the interaction of technology and culture than did the editors of *A History of Technology*.

Other broad ranging histories of technology have emerged from several national sources. Under the general editorship of Maurice Daumas, *Histoire Générale des Techniques* appeared between 1962 and 1979 (Paris). *Propyläen Technikgeschichte* (Berlin 1990–1992), a German collective survey under the general editorship of Professor Wolfgang König, also moves from ancient times into the twentieth century. A. A. Sworykin, N. I. Osmova, W. I. Tschernyschew, and J. W. Schuchardin published a Soviet version of the general history of technology translated into German as *Geschichte der Technik* (Leipzig 1964).

Recognized as a major scholarly achievement, Joseph Needham's multivolume *Science and Civilization in China* (Cambridge, UK, 1954) integrates from a Marxist perspective the history of medicine, technology, and science in a non-Western civilization. Among national histories of technology is *Geschiedenis van de Techniek in Nederland: de Wording van een moderne Samenleving, 1800–1890* (Zutphen 1992–1993) under the general editorship of Harry W. Lintsen.

5. Society for the History of Technology

The year of 1958 brought a turning point in the development of the history of technology with the establishment of the Society for the History of Technology (SHOT) and its journal *Technology and Culture* (*T&C*). Earlier, in 1920, the British Newcomen Society for the Study of the History of Engineering and Technology began publishing *The Transactions* which attracted, in particular, engineers interested in the history of their profession. The influential French journal, *Les Annales*, occasionally carried articles about the history of technology, and in 1935 published a special issue on the subject with a wide-ranging article on the history of water mills by its distinguished editor Marc Bloch.

Originally edited by Kranzberg and subsequently by Robert Post and John Staudenmaier, *Technology and Culture* attracts essays and readers from history of science, technology, and economics. The annual bibliography of the history of technology prepared initially by Eugene Ferguson proved especially valuable for researchers in the field. Within a few years, SHOT and *Technology and Culture* became the primary organizational framework for historians of technology in the USA and Western Europe. With several thousand individual members, SHOT includes non-US members on its governing committees and regularly has annual meetings in Western Europe. The International Committee for the History of Technology (ICOHTEC) provides through its colloquia and publications, a forum not only for US and Western European historians of technology, but also those from Eastern Europe, Russia, and elsewhere.

6. Graduate Programs in the History of Technology

Not only did SHOT and *Technology and Culture* provide an institutional framework for the history of technology, but the establishment of several graduate programs did as well. In 1954 the Eleutherian Mills-Hagley Foundation of Wilmington, Delaware with its archives, library, and museum, concentrating on history of technology and business organized, in conjunction with the University of Delaware, a graduate program for historians of technology. In 1973 the University of Pennsylvania initiated a doctoral program for the history of technology within its History and Sociology of Science Department. Other US universities followed suit, including the Massachusetts Institute of Technology. In Europe, Svante Lindqvist founded a history of technology and science program at the Royal Institute of Technology in Stockholm which has become a model for other institutions of higher learning in Western Europe. The research centers at Munich's Deutsches Museum and London's Imperial College also cultivate graduate training in the field.

7. Emerging Themes

7.1 Technology–Science Relationship

Since its founding, the pages of *Technology and Culture* express thematic concerns of scholars publishing and teaching in the field. In the 1960s, they often explored the relationship between technology and science. Motivation for this interest came in part from their wish to establish an identity for their subject other than as the history of applied science, a linear model

often adopted by historians of science and science policy makers. Historians of technology point out how often scientific theory rationalizes earlier technological practice.

Historian of technology Edwin Layton argued cogently in the 1970s that engineers develop their own applicable theoretical structures from practical experience and experimentation. Recently Walter Vincenti, eminent aeronautical engineer and historian of technology, has expounded this position persuasively in *What Engineers Know and How They Know It* (1990, Johns Hopkins University Press, Baltimore, MD).

Besides writing biographies of Thomas Edison and other independent inventors who experimented beyond scientific frontiers, historians of technology and science publish works about industrial research. John Beer provided a model for such studies with his essay on [German] 'Coal Tar Dye Manufacture and the Origins of the Modern Industrial Research Laboratory,' *ISIS* 49, (1958). Histories of the General Electric, Bell, and other major industrial laboratories followed. In these, fundamental research, in contrast with pure research, brings improvements in existing technology such as the three-element vacuum tube and incandescent lamp filaments.

7.2 Invention, Development, and Innovation

Technological change and creativity, a perennially intriguing human activity, became another early thematic concentration of historians of technology, especially those interested in economic and business history who see innovation as a source of economic development. After World War II, *The Sources of Invention* (Jeukes et al., 1969, W.W. Norton, London), which includes more than 50 case histories, stimulated a shared discourse about invention, as did historian Abbott Payson Usher's *A History of Mechanical Invention* (Cambridge, MA, 1954) and sociologist S. Colum Gilfillan's *Supplement to the Sociology of Invention* (San Francisco, 1971). Sociologist of science Robert K. Merton's seminal study, *Science, Technology & Society in Seventeenth Century England* (New York, 1970), broadens the discourse by demonstrating that attitudes and values stimulate technological innovation which in turn stimulates the search for underlying scientific knowledge.

The shared discourse usually defined technological creativity as consisting of three phases: invention, development, and innovation. Professional inventors and researchers in industrial laboratories were seen as presiding over the process of invention from a concept to a physical model. Development, a collective venture, moved the simple model through testing stages until a full-scale version could be tested. Innovation brought the device, process, or system into use. *Technology and Culture* authors usually situated the

three-phase process in an economic, political, and social context. The ongoing publication of the extensive Thomas Edison papers, along with Paul Israel's *Edison, a Life* (2000, New York) based on these, provides historians with an unprecedented rich source of information on a master of invention, development and innovation.

7.3 Industrial Revolutions, Systems, and Management

Industrialization also stimulated research and teaching among historians of technology. The rise of the USA to a position of preeminence among industrial nations raised questions about the nature of the industrialization process, especially as embodied in putative industrial revolutions in the UK and the USA.

Influential scholars exploring the process tend to see it as systematic. In *Das Kapital* (1867), Karl Marx portrays a factory as a system of textile machines driven by a single steam engine. Widening his perspective, he insightfully understands that textile factories interact with transportation, energy, and other systems to produce an industrial revolution, one scholars now label the 'British Industrial Revolution.'

David Landes an American economic historian, in *The Unbound Prometheus* (1972, Cambridge University Press, Cambridge, UK), writes of an interrelated succession of technological changes culminating after 1750 in the British Industrial Revolution and its spread throughout Western Europe. He shows how the coming of the steam engine, for example, depended upon developments in coal mining, iron manufacture, and railroad transportation. Ken Alder discusses the systematic interaction of military culture, industrialization, and politics in *Engineering the Revolution: Arms and the Enlightenment in France, 1763–1815* (1997, Princeton University Press, Princeton, NJ). *Histoire des Techniques* (Paris, 1978), edited and mostly written by Bertrand Gille, presents a history of successive technological systems and industrial revolutions from the earliest times into the twentieth century.

Historians customarily deal with the industrialization of the USA in the early nineteenth century as the spreading of the British Industrial Revolution by technology transfer and adaptation. David Hounshell in *From the American System to Mass Production, 1800–1932* (1984, Johns Hopkins University Press, Baltimore) challenges this interpretation by stressing the originality of the American system of production that developed in the nineteenth century and reached its apogee in the Ford system of mass production. Historian of technology Merritt Roe Smith in *Harpers Ferry Armory and the New Technology* (1977, Ithaca, NJ) sees the early nineteenth century American system as both a technical and managerial enterprise, and

places it within a shaping community, or social, context.

Only a relatively few historians explore a second Industrial Revolution in the USA and Germany after 1870 with its technical core of electrical and internal combustion technology.

Hugh G. J. Aitken surveys the spread of radio in *The Continuous Wave: Technology and American Radio, 1900–1932* (1985, Princeton University Press, Princeton, NJ) and Susan Douglas amplifies the story in *Susan Douglas, Inventing American Broadcasting, 1899–1922* (1987, Johns Hopkins University Press, Baltimore, MD). Within the frame of a Second Industrial Revolution, Leslie Hannah gives the history of electrical supply in the UK in *Electricity before Nationalisation* (1973, Johns Hopkins University Press, Baltimore, MD). *Networks of Power: Electrification in Western Society, 1880–1930* (1983, Johns Hopkins University Press, Baltimore, MD) by Thomas P. Hughes gives a history of electric power systems since 1880 in major urban centers in the USA, the UK and Germany. He defines systems as embodying in a seamless web a multitude of technical and organizational components and follows their development through stages characterized respectively and successively by the problem solving of inventive, managerial, and financial/political system builders. Hughes's methodology has spread among historians and sociologists who characterize it as a pragmatic, systems approach.

US historian of management Alfred D. Chandler in his seminal work *The Visible Hand: The Managerial Revolution in American Business* (1977, Cambridge, MA) stimulated a number of historians of technology interested in a Second Industrial Revolution to consider his proposition that innovative management since the late nineteenth century has done more to increase productivity than technical advances. Jurgen Kocka describes in a broad context large-scale, enterprise management at the Siemens company during German industrialization in *Unternehmensverwaltung und Angestelltenschaft am Beispiel Siemens, 1847–1914* (1969, Stuttgart), James R. Beniger in *The Control Revolution: Technological and Economic Origins of the Information Society* (1986, Harvard University Press, Cambridge, MA) moves from a second to an information revolution.

David Noble in *America by Design: Science, Technology, and the Rise of Corporate Capitalism* (1977, Alfred A. Knopf, New York) takes a strikingly different approach from Chandler on the rise of modern management. Reasoning from a Marxist perspective, Noble contends that US corporate culture represented by engineers and industrial scientists co-opted invention, research, and technical education in the interest of capitalism. Philip Scranton, a historian of technology and labor, offers in *Endless Novelty: Specialty Production and American Industrialization, 1865–1925* (1997, Princeton University Press, Prince-

ton, NJ) a correction to both Chandler and Noble by pointing out that small-scale, proprietary management and batch production continued to flourish in the era usually associated with large-scale mass production.

Military technology has been a constituent factor of modern history, especially during and since the Second Industrial Revolution. William H. McNeill in *The Pursuit of Power: Technology, Armed Force, and Society Since A.D. 1000* (1982, University of Chicago Press, Chicago) integrates military technology into general history. Technological developments during World War II and after, are described in Edward W. Constant's *The Origins of the Turbojet Revolution* (1980, Johns Hopkins University Press, Baltimore, MD). Richard G. Hewlett and Oscar E. Anderson, Jr. provide a detailed account of wartime atomic energy in *A History of the United States Atomic Energy Commission: The New World, 1939/1946* (1962, Pennsylvania State University Press, University Park, PA) and Richard Rhodes dramatizes the story effectively in *The Making of the Atom Bomb* (1986, Simon and Schuster, New York). Walter A. McDougall *The Heavens and the Earth* (1985, Basic Books, New York) places space technology in a political context.

Donald MacKenzie in *Inventing Accuracy: A Historical Sociology of Nuclear Missile Guidance* (1991, MIT Press, Cambridge, MA) and Gabrielle Hecht in *The Radiance of France: Nuclear Power and National Identity after World War II* (1998, MIT Press, Cambridge, MA) place political decision-making in the shaping context of evolving nuclear technology. A number of historians are exploring the recent history of the computer and biotechnology, but the definitive history has not yet appeared. Paul Edwards *The Closed World: Computers and the Politics of Discourse in Cold War America* (1996, MIT Press, Cambridge, MA) goes beyond narrative history to suggest that information systems are increasingly the shaping context within which military and political decisions are made.

7.4 Environment: Natural and Built

During the 1990s, the history of technology and environmental history intersected in the works of a number of historians. Earlier Leo Marx in *The Machine in the Garden: Technology and the Pastoral Ideal in America* (1964, Oxford University Press, New York) discussed the ideas of early American visionaries, such as Ralph Waldo Emerson who envisioned the harmonization of nature and technology in a pastoral landscape. In *Nature's Metropolis* (1991, W.W. Norton, New York). William Cronon describes the way in which Chicago's transportation and manufacturing technologies interacted seamlessly with the exploitation of the hinterland's natural resources to produce enormous quantities of marketable goods and nurture the enormous growth of Chicago.

Richard White, a social and environmental historian, further explores the seamless web theme in an imaginatively titled book *The Organic Machine* (1995, Hill and Wang, New York). It tells of the transformation of the Columbia River Basin in the states of Oregon and Washington into an artifact blending the natural and human built. White believes that the history of technology and the history of the environment can no more be told separately than the story of a marriage.

7.5 Gender, Race, and Technology

In the 1990s historians also began to explore assiduously the mutual shaping of gender and technology. 'Gender Analysis and the History of Technology,' a bibliographic essay in a special issue of *Technology and Culture* 38 (1997) documents the extensive literature emerging in the field. Ruth Schwartz Cowan in *More Work for Mother: the Ironies of Household Technology from the Open Hearth to the Microwave* (1983, Basic Books, New York) demonstrates that the household is a context within which factory-produced goods and consumption interact. Other historians further develop the consumption junction theme by showing that consumers, male and female, are not simply passive users of technology, but shape artifacts through their consumption preferences.

Judith McGaw opens another area of gender studies in *Most Wonderful Machine: Mechanization and Social Change in Berkshire Paper Making 1801-1885* (1987, Princeton University Press, Princeton, NJ) in which she demonstrates that women have gendered niches in the workplace which often perpetuate inequalities. Other historians argue similarly that some technological developments are shaped by ideas about racial differences and conversely that technology often defines racial roles, especially in colonial and developing areas. Michael Adas in *Machines as the Measure of Man: Science, Technology, and Ideologies of Western Dominance* (1989, Cornell University Press, Ithaca) demonstrates that nineteenth century colonial ideology associated technological progress, or the lack of it, with racial characteristics.

8. Methodology: Social Construction and Actor Networks

Until the 1980s histories of technology were generally written as narratives with little formal analysis. Then, European sociologists and anthropologists began shaping the field through the introduction of social construction and actor network methodologies. (See *Technology, Social Construction of; Actor Network Theory*.) French sociologist Michel Callon in several

influential articles including 'The State and Technical Innovation: A Case Study of the Electrical Vehicle in France,' *Research Policy*, 9 (1980) presents an actor-network model. It draws upon, but differs from, the Hughes systems approach previously described.

Callon uses an abstract concept 'actor' to signify—without distinguishing between—animate and inanimate components in a network. An actor-network creates and sustains a complex technological artifact such as an electric automobile. French professor of sociology Bruno Latour in *Aramis, ou l'Amour des Techniques* (1993, Paris) utilizes actor-network theory to present the history of a failed effort to introduce a radically new subway system in Paris. Deploying relativist sociology, Latour lets the actors, animate and inanimate, speak for themselves and offer conflicting, but equally valid, accounts of the evolution of a fluid project characterized by conflicting visions and policies.

Together with the systems and actor-network, the social construction approach of Trevor Pinch and Wiebe Bijker became a widely adopted methodology among historians and sociologists of technology in Europe and the USA. Influenced by the sociology of scientific knowledge cultivated at the University of Edinburgh, Pinch and Bijker offer a seminal example of social construction in an essay entitled 'The Social Construction of Facts and Artifacts,' *Social Studies of Science*, 14 (1984). They demonstrate that social environments shape the technical characteristics of an artifact, in this case the late nineteenth century bicycle. Social groups of varying class and different gender pursuing their interests and values define and determine the solution of bicycle-development problems thus embodying meaning in bicycle design.

Essays by Pinch and Bijker, Hughes, Callon, John Law ('Technology and Heterogeneous Engineering: the Case of Portuguese Expansion'), Donald MacKenzie ('Missile Accuracy: A Case Study in the Social Processes of Technological Change'), and by others are collected in *The Social Construction of Technological Systems; New Directions in the Sociology and History of Technology* (1989, MIT Press, Cambridge, MA). From this volume a language of analysis has spread, a language which includes 'seamless web,' 'systems,' 'actors,' 'network,' 'closure,' 'stabilization,' 'conservative and radical innovation,' 'technological momentum,' 'system builder,' 'heterogeneous engineer,' 'reverse salients,' 'interpretive flexibility,' 'translation,' and 'black box.' *The Social Shaping of Technology* (1998, Milton Keynes), edited by Donald MacKenzie and Judy Wajeman, also explores methodological and social construction themes. These concepts and the new approaches to the history of technology did much to discredit technological determinism so widely accepted among the laity, so persuasively argued by Karl Marx, and so eloquently enunciated by White in *Medieval Technology and Social Change*.

See also: Actor Network Theory; Business History; Economic History; History of Science; Industrialization, Typologies and History of; Science, Technology, and the Military; Scientific Revolution: History and Sociology; Technological Determinism; Technological Innovation; Technology and Organization; Technology and Social Control; Technology, Anthropology of; Technology, Social Construction of

Bibliography

- Buchanan R A 1991 Theory and narrative in the history of technology. *Technology and Culture* 32: 365–76
- Daniels G H 1970 The big questions in the history of American technology. *Technology and Culture* 11: 1–21
- Daumas M 1969 L'histoire des techniques: son objet, ses limites, ses méthodes. *Documents pour l'histoire des techniques* 7: 5–31
- Ferguson E S 1974 Toward a discipline of the history of technology. *Technology and Culture* 15: 13–30
- Hindle B 1966 The exhilaration of early American technology. In: Hindle B (ed.) *Early American Technology*. University of North Carolina Press, Chapel Hill, NC, pp. 3–28
- Hounshell D A 1981 On the discipline of the history of American technology. *The Journal of American History* 67: 854–65
- Hounshell D 1989 Rethinking the history of 'American technology'. In: Cutcliffe S H, Post R C (eds.) *In Context: History and the History of Technology: Essays in Honor of Melvin Kranzberg*. Lehigh University Press, Bethlehem, PA, pp. 217–20
- Hounshell D 1995 Hughesian history of technology and Chandlerian business history: parallels, departures, and critics. *History and Technology* 12: 205–24
- Pursell C 1978 History of technology. In: Durbin P (ed.) *A Guide to the Culture of Science, Technology, and Medicine*. Free Press, New York
- Rürup R 1972 Die Geschichtswissenschaft und die moderne Technik. Bemerkungen zur Entwicklung und Problematik der technikgeschichtlichen Forschung. In: Kurze D (ed.) *Theorie und Praxis der Geschichtswissenschaft*. Berlin, pp. 49–85
- Science Museum 1992 *Guide to History of Technology in Europe*. Lists history of technology programs, journals, and research scholars. London
- Staudenmaler J M 1985 *Technology's Storyteller: Reweaving the Human Fabric*. The MIT Press, Cambridge, MA
- Staudenmaier J 1990 Recent trends in the history of technology. *The American Historical Review* 95: 715–25

T. P. Hughes

History: Overview

1. Anthropological Universals

Human beings must interpret their world, themselves, and their coevals in order to live. Human activities, sufferings, and desires are formed by definitions of meaning and of purpose, which incorporate such

interpretations. An important area for such interpretation is the experience of time, and here memory plays a central role. The interpretations of the present that disclose it to us as a world, and allow us to act in this world, always refer to the past; the past is differentiated from the present and is thus related to it in that the future can be shaped as a perspective on action.

On this general anthropological level, history is the totality of the cultural practices through which human beings interpret their past in order to understand the present and anticipate the future. In other words: history is the generation of sense out of the experience of time.

But not every interpretation of time is historical. What determines the historical quality of an interpretation is the fact that time confronts human beings in two ways: as *real change within the world*, in which contingent events occur and are experienced, and as an *interior, temporal extension of consciousness* (and of the unconscious). Here, intentions and objectives, guided by the elemental mental forces of fear, anxiety, desire, hope, and longing, play a central role. We interpret the 'objective' time of real change (the present) through the recollection of another time (the past); thus the 'subjective' time into which we project a future is bound to our approach to past experiences. This is human culture's specifically historical procedure for experience and interpretation. What defines the historical interpretation of time is its implicitly meaningful and thus orientative conception of the course of time. The latter comprises the past, the present, and the future, and at the same time articulates experiences and intentions; it is thus empirical and normative at the same time. The philosophy of history explicates this conception of the course of time as a particular mental structure. Even without philosophical reflection or justification, such a conception of the course of time organizes historical thought, and shapes its content. History interprets the contingent events of the past and defines their significance for the present. It organizes historical narrative as a particular kind of explanation (Danto 1965). It is this 'explanation through narration' that defines the logic of historical thought in its academic form as well.

2. Conceptions of Time

The different conceptions of the course of time important for historical interpretation employ various temporal categories in various constellations (e.g., cyclical time, linear time, the concept of the *kairos*), each encouraging particular expectations: decline, progress, permanence, recurrence, revolution, destruction, and so on. They relate real temporal processes to the active intentions and self-conceptions of the subjects that make use of them. In other words, history mediates the subjectivity of acting and suffering

human beings with the objectivity of the temporal process of their world. In early societies, this was accomplished by an appeal to divine forces, which impel the transformation of the world and the human spirit at the same time; in other societies these divine forces are replaced by humankind as a spiritual and moral being, its ability to create symbolic worlds or to plan the conquest of the natural and social world through knowledge. Finally, nature, too, can be invoked as a sort of metaphysical instance for the continuity between the two temporal dimensions of human historicity. Nature can thus be understood as that which compels the human race, by means of social antagonisms, to create culture (as in Kant, for example), or as struggle of the races in Social Darwinist terms. Psychoanalysis, on the other hand, posits an interior, psychological nature, in order to explain the historical movement of the human world by means of the dynamics of the unconscious.

The idea of a 'history' comprising all temporal change within the human world is of modern origin (Koselleck 1985). However, it has a premodern equivalent in older cosmogonies, theogonies, and anthropogonies, as well as in the concept of redemptive history.

3. History as Event and Narration

In every culture and epoch, 'history' has a double meaning. On the one hand, it designates temporally organized events in the past. At the same time, and inseparably from the first meaning, 'history' also means the account of these events that articulates their significance for our understanding of ourselves and of our world. The first meaning has to do with temporal succession and the second with its narrative representation.

Not everything that has happened can be called history, but rather only that which is somehow meaningful in the present; for example, the foundation of a city or a kingdom is significant to its inhabitants, a person's birthday is significant to him or her, and so on. Usually, this event is described as a sequence of occurrences. These may often be the actions of human beings, heroes, or divine beings, as in myth or epic; however, they may also involve more complicated phenomena, such as the development of a manufacturing process, the evolution in the meaning of a concept, the epochal transformation of a mentality, an 'axial time,' and so on.

In its second meaning, history is the representation of the past in the culture of the present. Here, 'history' is the product of historical consciousness: through memory, the past has become present and meaningful in narrative form. In this way, the past 'in itself,' that is, everything that has happened, and persists only in the form of trace, relic, memorial, tradition, ethic, and so

on, becomes history 'for us.' The basic mental procedures involved can be organized somewhat artificially into four:

(a) the *perception* of 'another' time as different: the fascination of the archaic, the obsolete, the mysterious trace, the insistent memorial, and so on;

(b) the *interpretation* of this time as temporal movement in the human world, according to some comprehensible aspect (e.g., as evidence of the permanence of certain values, as examples of a general rule, as progress, etc.);

(c) the *orientation* of human practice through interpretation—both 'outwardly,' as a perspective on action (e.g., as the increase of political legitimacy through political participation, as the restitution of the world before its destruction, as the institution of 'true' conditions against the decline of morals) and 'inwardly,' as identity conceptions (e.g., 'we are the children of the sun,' or 'we as a nation stand for the universality and fulfillment of human rights,' or 'we belong to the communion of saints,' or 'we represent true spirituality as against others' materialism');

(d) and finally, the *motivation* for action that an orientation provides (e.g., a willingness to sacrifice, even die, for the sake of historical ideas of national greatness, the missionary spirit, etc.).

4. Memory, History, and Metanarratives

The relationship between memory and historical consciousness continues to be debated. Memory in the everyday sense is directed toward the experiences of individuals in their own lives, whereas historical consciousness primarily thematizes a past that exceeds an individual life span. Both, however—the individual memory on which individuality and social affiliation are built as well as the expansion of 'history' beyond the limits of a single lifespan back into the past—are as inseparable as two sides of a coin: human beings tend to transform their own identity into temporally more inclusive structures, such as the communion of saints, the nation, the people, the culture, and so on. The basis of individual subjects' self-esteem and of their practical temporal orientation tends to exceed their lifespan or to outlive them. It therefore usually extends far back into the past.

'History' generally performs its orientative function through a 'master narrative' or 'metanarrative.' This describes the conception and evolution of the addressee's world in order to legitimate its normative framework. At the same time it provides an experiential inventory, which is available whenever problems of orientation require consensus. It legitimates not only systems of politics, economics, and the environment, but also the collective and individual subjectivity and identity of the people. Metanarratives represent social affiliation and differentiation internally and externally.

Metanarratives generally organize social inclusion and exclusion *ethnocentrically*, assigning one group a positive normative weight, and other different groups a negative weight. One's own life world appears as culture or civilization, other worlds are looked upon as savagery and barbarism. This difference can be articulated in different ways: as an opposition, as deviation from a norm, as atavism on a progressive scale, etc. Just as history presents past experiences as a 'mirror' in which the present perceives itself, the image of others in the same historical mirror has a contrastive function: it defines a self-image. Under the special conditions of an orientation crisis, the 'other' can assume admirable features, as in the image of the 'noble savage.'

5. Historical Culture

The interpretive work of historical consciousness and its product called 'history' is concretely manifested in a society's historical culture. Historical culture is multidimensional, like every other culture. It has religious, moral, pedagogical, political, and rhetorical expressions; its cognitive substance is always the knowledge of '*wie es eigentlich gewesen*,' how it really was. We can distinguish three basic dimensions of historical cultures as an ideal type, each different in its logic and thus accountable to different criteria of meaning.

(a) The *political* dimension is concerned with the legitimation of a certain political order, primarily relations of power. Historical consciousness inscribes these, so to speak, into the identity concepts of political subjects, into the very construction and conception of the I and the We, by means of metanarratives that answer the question of identity. There is no political order that does not require historical legitimation. The classical example, applicable to every culture and every epoch, is genealogy. Even the pure rule of law that appeals only to the applicability of formal decision procedures must be historically based if these procedural rules are to be plausible to the participants. Charismatic leadership also cannot do without historical elements. Generally, the vehicle of political charisma will refer to the spiritual or natural forces that, within his culture, guarantee the world's temporal coherence.

(b) The *esthetic* dimension is concerned with the psychological effectiveness of historical interpretations, or that part of its content that affects the human senses. A strong historical orientation must always engage the senses. Masquerades, dances, and music can all have historical content. Many older metanarratives are composed in poetic form and are celebrated ritually. A formal defect can destroy the effect of such a presentation, and even endanger the world's continued coherence. Historical knowledge must employ literary models to become discursive. In

many cultures, historical narrative occupies a secure place in the literary canon as a genre of its own. In modern societies, memorials, museums, and exhibitions are among the familiar repertoire of historical representation. In older kinds of social systems, objects such as relics, tombs, temples, and churches obligate the present to the legacy of the past, indeed make the present, in its relationship to the future, responsible for the vitality of historical memory.

(c) The *cognitive* dimension is concerned with the knowledge of past events significant for the present and its future. Without the element of knowledge, the recollection of the past cannot effectively be introduced into discourses concerned with the interpretation of current temporal experience. Mythical metanarratives, too, have a cognitive status, although science would eventually deprive them of it. They lose their orientative power when confronted with a science of the past that possesses a more elaborated relation to experience. Metanarratives then become prosaic, as they already did in antiquity, with Herodotus and others. In modern societies, the cognitive dimension of history has become an academic discipline.

6. Typologies of the Historical

Synchronic and diachronic typologies of the historical can be generated out of the particulars of historical thought, and inquire into its carriers, functions, class and gender specificity, and other factors within the social context's framework of determinations (Rüsen 1996). They become historically specific, however, only when they differentiate the specific interpretative work of historical consciousness and of its product—'history' as the cultural orientation system of temporal consciousness—according to the different possibilities of historical meaning construction. Such differentiation from a metacultural perspective has been little attempted. Previous efforts have usually focused on functional and formal aspects, and have thematized less the total domain of human temporal interpretation than simply the domain of historical writing, and if at all they only inspected the Western tradition. Nonetheless, such typologies can also be used as analytical tools for research in the wide areas of historical culture and for intercultural comparisons.

Nietzsche (1985), for example, differentiated the monumental, the antiquarian, and the critical uses of history in terms of their advantages and disadvantages for life:

(a) the *monumental* approach inspires action by recalling the creative deeds and great moments by which the world was actively transformed, and emphasizes the greatness of man as the agent of historical change;

(b) the *antiquarian* approach serves the preservation of tradition and the affirmation of existing conditions by its reverence for and transfiguration of their origins;

(c) the *critical* approach, finally, defeats the power of memory, breaks the control of tradition, and liberates life for new orientations.

Hayden White (1973) presented a very influential model for the analysis of historical writing, and applied it to the nineteenth century's most significant historians and philosophers of history. It consists of a complex system of emplotment, formal argument, ideological implication, and tropes of historical discourse.

(a) The narrative forms that connect a sequence of historical events can follow the patterns of romance, tragedy, comedy, or satire.

(b) Formal argument no longer concerns so much the narrative combination of events into a story or narrative, but rather the explicit explanations within these narratives. They can be formalist, emphasizing the uniqueness of an event; they can be organicist, integrating isolated phenomena into larger totalities for explanatory purposes; they can argue mechanistically, in terms of causal regularities; and they can operate contextualistically, in terms of functional relationships between events.

(c) Ideological implications relate to the standpoint within the present from which a history develops its perspective on time. Such a perspective (White is most concerned with the nineteenth century) can be anarchistic, conservative, radical, or liberal.

(d) What is decisive for the extensive and foundational meaning of a narrative is the verbal factor that brings the different elements of historical narrative into a coherent whole. White identifies and explains this factor in terms of rhetorical figures or tropes that underpin the four possibilities for historical meaning: the latter is generated out of metaphorical, metonymic, synecdochic, or ironic uses of language.

White developed this typology for the analysis of historiographical texts; nevertheless, it can be applied to other manifestations of history, such as historical museums (Bann 1984).

Rüsen (1993) also developed a universally applicable typology of historical sense generation according to its most basic principles: its particular referents, its conception of temporal movement, its communicative form, and its identity concept. All these factors are communicated through and constituted by the principle of historical sense.

(a) The *traditional type* recalls the origins grounding present relations. Traditional history represents the course of events in terms of the persistence of these origins and their normative power, and represents identity by affirming prior interpretive models of human subjectivity (e.g., social roles). Past and present fuse into the permanence of currently effective social relations that are conveyed by time yet are above transience. Traditional narratives eternalize time as meaning. This type predominates in mythical thought.

(b) The *exemplary type* recalls past facts that concretize the rules of current conditions. Exemplary

history represents a course of events in terms of the transtemporal applicability of these rules, and constructs identity by generalizing various temporal experience into rules for action (e.g., as rule-based competence). This narrative type is represented in Europe by the classical motto *historia vitae magistra*, and in China by the metaphor of the mirror. This type predominates in the historical thought of great civilizations up to the modern period.

(c) The *critical type* recalls past circumstances in order to interrogate present conditions. Critical history represents a course of events as the transformation of prior and, in the orientative framework of life praxis, effective interpretive models (i.e., as rupture or as alternation). It constructs identity by negating pregiven identity-forming interpretive models; in other words as the power to say 'no.'

(d) The *genetic type* recalls qualitative changes in the past by which other, strange conditions become present familiar ones. The genetic construction of meaning presents a course of events as evolution, in which conditions change in order to become dynamically permanent. It constructs identity as the synthesis of permanence and change; in other words as a dynamic process. This type predominates in modern societies.

7. Media

The diverse manifestation of history, in its dual significance as fact and as interpretation, depends crucially on the medium through which temporal experience and interpretation are communicated. Language is always predominant, although there are other creative areas of historical perception and representation, such as the visual and musical arts, architecture, and dance. The modes by which language orders human communication are decisive for the expression of historical consciousness.

(a) History was originally communicated orally, as it is even now—at least in the private sphere. In preliterate cultures, history becomes oral history or oral tradition (Vansina 1985). Such a history is usually politically conservative: the traditional world order becomes obligatory; new experiences have to be integrated into it so as to confirm the collective's basic norms and identity. Such representations are sensually immediate and concrete; they are communicated face to face and correspondingly have immediate effects. Cognitively, they contain the cumulative experience of many generations; experientially, they reach into a mythical time that genealogically connects the present to the all-determinative origin of the world and include the more recent generations' memory which contain real events. Its criteria of significance are of a mythical nature; in other words, topical and in the narrow sense 'historical' events are not the vehicles of orientative normativity. Its validity claims are raised and made

plausible through ritual and repetition. In modern historical research, oral history is the method for gaining historical knowledge from the recollections of acting and witnessing people. It provides an important source for research into the mechanisms of biographical memory as an element of historical culture in the intersection of the public sphere and individual biography. The memories of witnesses to historically significant events, communicated orally, acquire an important role in public historical culture—especially the testimony of survivors of great catastrophes such as the Holocaust (Langer 1991).

(b) The medium of *writing* has dominated for millennia and continues to dominate hegemonic historical culture (in which, of course, oral elements continue to be important and effective, even constitutive—at least in education). Written history frees participants from the immediacy of the communicative situation and creates a distance between history as content and history as a form of communication (Ong 1982). This distance increases considerably the experiential horizon of historical consciousness, and facilitates new methods for accumulating and managing experience. Literacy liberates memory, fixes facts, creates new communicative forms, disconnects the cosmos of temporal meaning from immediate practical relationships, and allows it to become a phenomenon *sui generis*. Along with new opportunities for objectivation in the approach to historical experience, it also discloses new opportunities for subjectivation through historical interpretation: the narrator can become an author, and readers can obtain expanded scope for critical readings. Validity claims either are raised through canonization, or become an opportunity for discursive justification. In either case, interpretation becomes a discrete practice of historical meaning construction, with its corresponding specializations and the attendant problems of mediation and popularization. The original relationship between poetry and truth dissolves; mythology is now subject to a fundamental critique. Explicit conceptualization becomes an essential cognitive element in historical interpretation, so that history can, at least in the long term, be an object of study and of theory.

Entirely new media have transformed the historical mode of thought. We cannot yet discern distinct lines of development and stable structures, but we can certainly indicate new developments from which we might expect transformations of the most basic kind.

In the historical culture of the public sphere, collective memory is being overpowered by the torrent of historical images. The forms of consciousness created by literacy—and above all the distancing effect of rationality—may quickly decrease in significance, and especially in political efficacy. The grammar of history is becoming an imagology of presentations in which every era is contemporaneous, and the fundamental idea of a single linear movement of time is

disappearing. The constitutive difference of temporality can be suspended into a universal contemporaneity that can no longer be narratively put in order. Whether there can then be a specific 'historical order' within the orientative temporal continuity between past, present, and future has at least become a problem. The very term *post-histoire*, and the related discussion of a mode of life without genuinely historical interpretation (Niethammer 1992), suggest that these questions are now open. At the same time, there has been an immense increase in access to data. New storage media allow new modes of historical experience, and radically call into question previous criteria of significance. At the same time, new media of communication such as the Internet permit no single, politically sanctioned decision. The abundance of possibilities and the diversity of new voices require new strategies, new forms, and new contents of historically grounded participation and exclusion. In each case, fixed ideas of the permanence and substance of individual and collective identity are being outstripped by the diversity of global communication, in favor of more dynamic and more open differentiations. This process then provokes reactions, often expressed through these new media, that stubbornly insist on ethnocentric distinctions.

8. Developments

The transformation in the media of historical thought sketched above suggests a transcultural, macro-historical transformation, that will ultimately leave no culture unaffected. This provides us with a support for a diachronic classification for historiographical variety. Such a classification could be organized in terms of epochal distinctions, yet still have a typological quality. For example, White's or Rüsen's typologies could both be used to characterize historical processes. White's sequence of tropes can be understood as cyclical, and Rüsen's distinction between traditional, exemplary, and genetic historical thought can describe the great transcultural epochs where the critical type of historical thought mediates the transition between the other types.

One of the most important distinctions concerns the relationship between myth and 'history,' where the latter is understood as mundane occurrence. Myth relates to a divine prehistory as the source of meaning (Eliade 1991). In contrast, 'historical thought' in the stricter sense discerns the meaning of history in the immanent dimension of a chain of events (*res gestae*). The Judeo-Christian sacred history mediates between these two modes of thought, and allows events in the world to appear as the manifestation of God's power and God's will. In other cultures, the divine and the mundane are mediated differently: mundane but somehow significant events have to be integrated into a comprehensive cosmic order. History becomes mean-

ingful through nature, which is conceived as divinely ordered. With increasing complexity in human praxis, more and more mundane incidents are integrated into this order until it is finally 'historicized' and 'denaturalized.' Yet this process is often opposed by attempts to renaturalize it as biological racism.

Comprehensive large-scale developments of historical thought have often been conceptualized as leading to the current status of historical studies. That is often criticized as too one-sided and the diversity of new media and the different dimensions of the historical culture are emphasized instead; nevertheless plausible accounts of a developmental process can be articulated (in the form of typological hypotheses). These accounts can be used to discern and analyze transformations of historical thought in the diversity of different cultures.

Max Weber's conception of a universal process of rationalization and disenchantment seems as relevant as ever (Weber 1988). The classical philosophy of history of Voltaire, Kant, and Hegel had presented these criteria in the form of the actual course of events. To the extent that historical studies acquired increased competence in investigating and representing real events in the past, the question of meaning was investigated within the framework of a logic of historical thought (Dilthey, Max Weber, Rickert). This question is now perceived primarily as the concern of the central verbal procedure of historical thought, as historical narrative (Danto 1965). Issues of scholarly standards and truth claims (Appleby et al. 1994; Evans 1999) can be obscured by only focusing on the rhetorical and poetic strategies of historical writing (Rigney 1990).

Despite this formalization of historical understanding, the structures of real historical processes remain important points of reference. Here, the global and often conflictual transformations brought about by modern industrial societies are in the foreground. Their developmental dynamics can be understood only by a concept of time in which the horizon of experience and expectation structurally diverge, and can only be reconnected by a fundamental temporalization of the human world (Koselleck 1985).

Narratives of progressive rationalization and disenchantment must, of course, be complemented by an account of a critical and transcendent re-enchantment resistant to rationality, such as the endlessly renewed rationality of sense (value rationality); it refuses to dissolve in methodical rationality of the control of experience. Historical studies, the development of which in Europe from the eighteenth century has been the most effective factor in the modernization of historical culture, must employ sense criteria that methodical rationality alone cannot generate.

The rationalization of modern historical culture is constituted by a universalization of the norms and values that make the past significant for the present.

The ethnocentric logic of identity formation always has a universalizing tendency, insofar as it describes its own world as both unique and normatively obligatory. 'Humanity' is a common frame of reference for historical identity. In the focus of historical experience, humanity is extended to the limits of the known world. This creates a fundamental, structural conflict of cultures in historical thought, which has repeatedly led to violence, to the legitimation of domination, and even to genocide.

This leads us to a fundamental problem in contemporary historical thought. It is shaped by the rationalism that culminates in historical scholarship and related human sciences (Bentley 1997). This rationalization of history as scholarship corresponds to the modern rationalization of all knowledge, without which the organization of modern societies would be inconceivable. Is historical thought inseparable from the value system of Western culture in its current form along with this methodological rationality? If so, how can it conceive of history in such a way as to account for the diversity in different cultures' articulations of 'history' as the orientative framework of human praxis? Does this diversity, once it has been researched and analyzed by scholarship, not always appear in the light of a particular interpretation—namely the interpretation to which methodological reason essentially belongs?

Current tendencies in historical thought (post-modernism, postcoloniality, the 'cultural turn') are determined to overcome modernity in this sense. They want to create a pluralism of 'histories' that can break the spell of uniformizing rationality and generalizing normativity. They thereby destroy the orientative power of modern historical thought with its notion of a unified direction of development, or 'progress.' At the same time they problematize the claims of the scholarship that describes the laws of such development along with the desire for mastering them, e.g., in Marxism.

In the age of globalization, in spite of the critique of the claim for hegemony inherent in Western historical thought, we cannot do without either methodological rationality or the universal regulative of communication. We need this regulative to understand the present by the interpretation of the past to define inclusion and exclusion, and to act in anticipation of the future. Yet, if the ethnocentric logic of such a universalization is to be broken or overcome, the truly universal regulative of historical thought and of the historical culture can only be a mutual and thereby universal recognition of difference and particularity. This idea is and will remain counterfactual, even utopian; nevertheless, it could disclose new dimensions of historical experience and lead to new and politically significant orientations.

Certainly the terrors and catastrophes of the twentieth century, such as the Holocaust, provide a new challenge to historical thought. Only with difficulty

can they be integrated into older conceptions of historical meaning. They require new strategies of interpretation and representation. The question of human 'nature' must be radically reconsidered, and senselessness itself should be taken as a constituent of a meaningful conception of history (Friedländer 1998).

Bibliography

- Appleby J, Hunt L, Jacob M 1994 *Telling the Truth about History*. Norton, New York
- Assmann J 1992 *Das kulturelle Gedächtnis: Schrift, Erinnerung und politische Identität in frühen Hochkulturen*. Beck, Munich
- Bann 1984 *The Clothing of Clio*. Cambridge University Press, Cambridge, UK
- Bentley M (ed.) 1997 *Companion to Historiography*. Routledge, London
- Danto A C 1965 *Analytical Philosophy of History*. Cambridge University Press, Cambridge, UK
- Eliade M 1991 *Myth of the Eternal Return, or Cosmos and History*. Princeton University Press, Princeton, NJ
- Evans R J 1999 *In Defense of History*. Norton, New York
- Friedländer S 1998 Writing the history of the Shoa: some major dilemmas. In: Blanke H-W, Jaeger F, Sandkühler T (eds.) *Dimensionen der Historik: Geschichtstheorie, Wissenschaftsgeschichte und Geschichtskultur heute*. Böhlau, Cologne, Germany, pp. 407–14
- Hutton P H 1993 *History as an Art of Memory*. University Press of New England, Hanover, NH
- Koselleck R 1985 *Future's Past: On the Semantics of Historical Time*. MIT Press, Cambridge, MA
- Langer L L 1991 *Holocaust Testimonies. The Ruins of Memory*. Yale University Press, New Haven, CT
- Lowenthal D 1985 *The Past is a Foreign Country*. Cambridge University Press, Cambridge, UK
- Megill A 1998 History, memory, identity. *History of the Human Sciences* 11(3): 37–62
- Müller K E, Rüsen J (eds.) 1997 *Historische Sinnbildung—Problemstellungen, Zeitkonzepte, Wahrnehmungshorizonte, Darstellungsstrategien*. Rowohlt, Reinbek bei Hamburg, Germany
- Niethammer L 1992 *Posthistoire: Has History Come to an End?* Verso, London
- Nietzsche F 1985 *The Use and Abuse of History*. [Trans. Collins A]. MacMillan, New York
- Nora P 1989 Between memory and history: les lieux de mémoire. *Representations*. 26: 7–25
- Ong W J 1982 *Orality and Literacy*. Methuen, London
- Ricoeur P 1984–87 *Time and Narrative*, 3 vols. University of Chicago Press, Chicago
- Ricoeur P 1998 *Das Rätsel der Vergangenheit. Erinnern—Vergessen—Verzeihen*. Essener Kulturwissenschaftliche Vorträge, Vol. 2. Wallstein, Göttingen, Germany
- Rigney A 1990 *The Rhetoric of Historical Representation: Three Narrative Histories of the French Revolution*. Cambridge University Press, New York
- Rüsen J 1993 *Studies in Metahistory*. Human Sciences Research Council, Pretoria, South Africa
- Rüsen J 1996 Some theoretical approaches to intercultural comparative historiography. *History and Theory*. 35: 5–22
- Vansina J 1985 *Oral Tradition as History*. University of Wisconsin Press, Madison, WI

- Weber M 1988 Science as a variation. In: Lassman P, Velody I, Martins H (eds.) *'Max Weber's Science as a Vocation'*. Unwin Hyman, London
- White H 1973 *Metahistory. The Historical Imagination in Nineteenth Century Europe*. Johns Hopkins University Press, Baltimore

J. Rüsen

History, Professionalization of

It is evident that the term professional in this connection has two different meanings, which are related to each other but are not equivalent. The first (a) is professional in the sense of being able to live by the trade (in the meaning of occupation, such as the French word 'profession'). The second (b) is professional as a qualified person who is recognized by society (not least by his or her fellows in academia) as competent to practice the trade with a closure directed against nonprofessionals, i.e., amateurs and those writing on history without being 'real' researchers.

1. The Historian's Profession as Bread-winning

In the first sense of professionalism (see *Professionalization/Professions in History*), there have been professional historians for a long time but there is no distinct collective process of professionalization. The writing of history is age-old. Some famous history writers were active in the European antiquity, and especially Herodotus and Thucydides of Greece have won lasting fame. These and other early European historians were not writing history for their living. In China from at least the second century BC, and later in Korea and Japan, historians were engaged to describe the history of their country under a dynasty. This usage involved a real employment in China and it even entailed a personal responsibility of the historian for what he wrote. Sima Qian (born 145 or 135 BC), known as the Grand Historian, was employed to make justice to the Han dynasty but he also maintained a theoretical discourse on history (Wang 1995). He and other Chinese dynastic historians were obviously fully occupied and cannot even normally have worked alone but must have had a small office with clerks at their disposal. The Chinese professionalism among its historians had no known counterparts. In Korea the first well-known historians (eighth century AD) were politicians who wrote on their own behalf. In India there were no real historians until the thirteenth century because of the dominant cyclical time perception in the Hindu culture. The late Roman and medieval European historiographers had little profit from their writing and could not live from it. 'A book belonged to its author for as long as it stayed in his

hands. After that it was free as air' (Smalley 1974). Some late medieval chroniclers in Europe however managed to live by their work as historians. Such chroniclers were employed by princes or cities in order to relate the important events that took place within the principality or city.

In Europe, the chroniclers employed by princes and cities (those active in monasteries were not employed for the purpose and formally not living by their historical writings) were followed later by official historiographers hired by kings and states for the 'proper' recording of the wars where their masters were engaged. Titles like Historiographer of the Realm (in variations all over Europe) make clear the official mission of the historian. The French *Historiographe du Roi*, was a title used from 1554 to 1824, and 104 persons bore the title. It was not so much a function as a title with a reward, and the historiographers were not highly regarded at court (Boer 1985). For most such official historiographers, the revenue from the title was only a side income and not enough to live on. Quite another type of professional historians in this sense arose with the great wars in Europe in the seventeenth century: those who lived from their reports from the war scenes. They established a form of market situation for themselves, and especially those behind the *Theatrum Europaeum* were successful in their branch, a sort of intermediary between history and journalism.

Already from the late eighteenth century another type of professional historians, in sense (a) of the word, appeared in Europe: the literary men writing history and more or less living by their literary accomplishments. Men like Voltaire, Gibbon, Robertson, and, to an extent, Schöler and Herder made themselves known as historians (though primarily as learned and literary men) and their works were sold in editions of considerable numbers. Their ambition was to a large extent to give to the reader an impression of historical personalities and their struggle with moral problems, a sort of applied political science or applied moral philosophy. These historians did not strive for new factual knowledge at first hand but rather a better insight into the moral problems of politics pursued by states and kings. Such literary professionals who worked with historical subjects were more successful than the contemporary professors of the universities of Europe, who also earned their living from history and who were mostly not well known.

Still in the (a) sense of the word professional the authors of the 'broad' literature on history which grew up from the early nineteenth century should be mentioned. Many narrators of history wanted to give to their countrymen a better understanding of the past of their nation. Publishing narratives on national history for the broad layers of the population sometimes also gave a living to the authors already in the first half of the nineteenth century. It is easy to see the continuation of such popular narrative into the twentieth century. The number of persons who have had

the possibility to earn their living by telling history in different media and not only in printed books has increased enormously. The demand for history as entertainment, originating in the eighteenth century among the upper classes, has grown and spread in society. In the twentieth century, publishers have been interested in different types of books, such as specialized historical works, textbooks for schools and universities, historical journals, historical fiction and pure historical entertainment. Adaptations of history in various forms for radio and television have also gradually increased their share in the total turnout, and art film has been used for telling stories with a historical setting. The comprehensive number of persons who make their living from these forms of writing and processing history has never been calculated.

In the twentieth century, though starting earlier, there has also been a great growth of the number of professional historians in sense (a) of the word, who have been employed in teaching. Here must be included all teachers specializing in history, both in secondary schools and at the university level. Historians working in archives and museums also belong to this category. With the great expansion of secondary and tertiary education, the sector has been steadily growing in the twentieth century, especially after World War II. For France in 1910, Boer has estimated the number of historians in these categories to 1045 (Boer 1985). On the eve of the 1950s the number of professional historians in France has been approximated as less than 3,000 but then this number does not include archivists and historians employed in museums. In 1967, the number of the same limited category of professional historians had grown to close to 8,000 (Charle 1995). German and Australian calculations go in the same direction. (Raphael 2000, Macintyre 2000).

2. Establishing Professionalism Among Historians as a System of Norms

The second sense in which the term professional is used, 'professional as a qualified person who is recognized by society (not least by his/her fellows in academia) as competent to practice the trade,' has a more complicated historical reference. Historians who were recognized as competent historians by their contemporaries have existed for long. But in regard to the recognition by fellow historians, especially those with university posts, this is something much newer. If recognition is taken to mean an inclusion into a network, group or a formal association, only the history of the period from 1750 provided any means for the realization of such professionalism. If, further, a sort of common program of the network or group is required, the field is narrowed even more.

From the middle of the eighteenth century we can discern the beginnings of a formation of schools

among academics in general and also among historians in Europe. It is difficult to say if anything comparable has been at hand in other parts of the world earlier. This was the period when German universities were flourishing and universities in other parts of Europe and its colonies were not. Because of the low level of state power in Germany not only free authors could express their views but also civil servants, such as professors. This made German universities, especially those in Halle and Göttingen, serve as a nursery for new notions in the humanities. Linguistics and history prospered. In the nineteenth century, 130 Protestant historians (plus an unknown number of Catholics) became full professors of history at German universities and for most of them this chair meant a culmination of their academic career. For the first time an occupational group of historians developed. Their competence for the professorship was mostly carefully examined and a discourse on history emerged (Blanke 1991). Mostly, however, these were individual accomplishments and the level of coordination between the learned men was low. No common rules for historical works were acknowledged, though an effort to write a manual of historiography was made by G.A. Will. Many other historians, e.g., Johann Martin Chladenius and Johann Christoph Gatterer, lectured on rules and wrote on the theory of history (Reill 1975). The process of standardization of methodological rules was, however, regarded with skepticism by some of the most important historians, such as August Ludwig Schlözer, who talked about imprisonment in scholarly rules, as Blanke has noted (Blanke 1991). In the early years of the nineteenth century at least four introductions to historical studies were published. As to rules for the critical use of sources and general methodological emphasis the introductory book by Friedrich Rühs from 1811 is noteworthy.

There were hardly any 'schools' of historians before well into the nineteenth century. Academic teaching of history and research in history were well established in many countries in the eighteenth century, though without a canon of methodology or a professed normative system, which applied to history as a discipline. History was rather regarded as a part of a wider discipline of morality, the French *sciences morales et politiques*. Past events could be used to show the consequences of different types of conduct and new enlightenment notions made it possible to analyze historical processes without referring to God or providence (Heilbron 1990). For centuries history had been used for this purpose, but the morality concerned changed its content and its stability in the eighteenth century. The new directions were not diffused to students in standard formulations or transformed into methods and manuals. The cohesion among enlightenment historians was low and this continued to be the case up to around 1830.

The change brought about primarily by Leopold von Ranke was to create the conditions for a network

of historians under the guidance of the master. He had meetings with his disciples, later known as his seminar, where problems of history were discussed. Ranke himself was a spokesman for three main principles of historical interpretation: the state as a central entity, the state system as the polity of international relations, and the necessity for the historian to evaluate each single entity for its role in history and not from its relation to contemporary politics. The first of these principles was a common ground for many thinkers of the period, but combined with the other two it formed a specific analytic system. The third principle has been partly traced back to forerunners in the eighteenth century (Meinecke 1936, Blanke 1991, Blanke and Rüsen 1984) and it created the basis of the thought system, which became known as *Historismus*.

In the German academic community of historians the combination of *Historismus* and teaching through seminars became an instrument for the professionalization of historians. A common evaluative system was created in the latter half of the nineteenth century. A methodological part was often combined with individualization of persons and collective entities of the past, especially among the historians of the Middle Ages who were concerned with the editing of the *Monumenta Germaniae historica*. This work elaborated further methodological principles for editing which had been developed in France by Jean Mabillon and others from the seventeenth century and onwards, and which gave rise to the *École des Chartes* at Paris. Methodology became important and the principles for the evaluation of sources became more and more systematized. A methodological direction of historians within the *Historismus* has therefore been discerned, with a philosophical direction as its counterpart. The former was characterized by its devotion to empirical studies with a sincere application of methodological principles in order to obtain objective knowledge, while the philosophical direction generally rejected the possibility of reaching any form of certainty and proclaimed relativity of knowledge. In fact it will seem that the differences within *Historismus* were great. On the one hand were those who saw historical relativism as fundamental and regarded history primarily as a means of orientation in the present. On the other hand were those who wanted to overcome as much as possible the constraints of the sources for understanding the past, though still with a considerable relativism in regard to values (Oexle 1996, Scholtz 1997).

In several respects a new turn of historical professionalism grew out of the empirical *Historismus*. It was important that new methodological manuals were published and widely accepted in the European community of historians. J. G. Droysen had published his *Grundriss der Historik* (1868), but its emphasis was rather on the principles of *Historismus* than on the strictly methodological side. Ernst Bernheim's *Lehrbuch der Historischen Methode* (1889) and

Charles-Victor Langlois' and Charles Seignobos' *Introduction des Études Historiques* (1898) were in this regard epoch-making. These books were in the first hand methodological (from the 1903 edition Bernheim added a section on philosophy of history), and they received a very wide circulation. They taught how to find secure knowledge in history. The authors sought somewhat different philosophical bases for their standpoints but the authors of both books turned their back to the relativism of knowledge. Their differences mattered less than their methodological aims, which found appeal among historians.

The first decades of the twentieth century witnessed a change in the general approach to historical professionalism. The emphasis had been on the theory of the state or the theory of state systems or the theory of the (individual) nation. Now the methodological element became central. Without the proper knowledge of how to evaluate sources a historian was not accepted as a professional. Several factors helped to make methodology a central element in the professionalism of historians. One was the rise of national historical associations. This was a process that took place all over Europe (and also beyond Europe in some countries) in the latter half of the nineteenth century. Philosophical and political issues divided all such associations and, thus, only methodology was really a common basis. This was even more the case with the international organization of historians, later known as the *Comité International des Sciences Historiques*, which can be traced back to 1898. In the early programs of its international congresses from 1900 and 1903 methodology played an important role and the object of the discussion was rather which were the conclusions for the philosophical foundations that must be drawn from the established methods than the other way round. After the congress of 1908, the French historian Simiand complained in a review article that methods had not been sufficiently emphasized in the congress contributions (Erdmann 1987).

In the last decades of the nineteenth century and the early part of the twentieth century the historical associations of the European countries also in most cases started journals (Stieg 1986). The influential German *Historische Zeitschrift* was, however, founded in 1859 as a private initiative by Heinrich von Sybel. The earliest still existing seems to be the Danish from 1838. During the twentieth century the creation of journals of history has become usual in all parts of the world, and some of these are much more than national journals but quite general (such as the *American Historical Review*, founded in 1895) or thematic for studies of the history a major part of the world (such as the *Journal of Asian Studies*, founded in 1957). Other journals and reviews with a thematic or rather broad methodological focus, such as the *Revue de Synthèse Historique* (1900–31); the *Annales d'Histoire Économique et Sociale* (founded in 1929, from 1946 called *Annales: Économies, Sociétés, Civilisa-*

tions); the *Past and Present* (founded in 1952); the *Journal of Contemporary History* (founded in 1966); the *Journal of Modern History* (founded in 1929), etc., were mostly later creations. In most of the journals, articles on themes of history were combined with critical reviews of books. An important element in such reviews was to examine the methods used and the shortcomings of authors in respect to methodological rules. Thus, in the form of critical reviews of the works on history, the methodological norms were spread. These norms were also commonly accepted. In the case that a historian wanted to contradict criticism he or she had to show that the principles had been observed rather than that these principles were not valid. This is something that characterizes methodology but is not valid for the kind of substantive norms that formed the major part of *Historismus*.

For the formation of the profession of historians, the debates on the application of historical methods were very important. A series of such debates took place in different countries in the first decades of the twentieth century. There was a big debate in Germany, called the *Methodenstreit*, on the correct use of methods in regard to any scientific activity. Generalization and interpretation were advanced as alternatives. Karl Lamprecht, with his concept of social and cultural history, was rather lonely in Germany and is often regarded as defeated. In fact, part of the issue was the theories of *Historismus* and the alternative professed by Lamprecht. Another part dealt with methods and their application in the strict sense, where individualization (W. Windelband, H. Rickert, M. Weber, F. Meinecke, and others) was opposed to views based on structures and collectivities (W. Sombart, H. Pirenne, F. Simiand and others) (Iggers 1984). The struggle was reflected in other countries such as Finland. In other countries, e.g., Sweden, Norway, and China, there were other straightforward struggles about the use of methodological principles and the consequences for national history. The historians who advanced new and radical consequences of methodologies, which were already well known, had a program. They wanted to establish firm knowledge in history, a knowledge that was contrasted to guesses and conjectures, not to historical relativism. In the USA, the focus of the debate was the possibility of establishing objectivity (Novick 1988).

Historical professionalism also increased through the ongoing specialization of all historical fields. In the beginning of the twentieth century, it was no longer the normal thing for a historian to take as the theme of an investigation the history of a country during a certain period. This had been quite usual in the middle of the nineteenth century, but in the first decades of the twentieth century this had changed. History had begun its separation into specific subfields, and economic, social and cultural history were regarded as major such parts. Specialized research institutes came into existence and specialized journals were published: the

Economic History Review from 1927, the *Vierteljahrschrift für Sozial- und Wirtschaftsgeschichte* from 1903, the *Journal of Contemporary History* from 1966, and the *Vierteljahrshefte für Zeitgeschichte* from 1953. Similar journals were published in many countries with specific reference to the speciality in the national context. Specialization went further, chronologically and thematically. Investigations became very specialized in regard to field and object of study, and period of investigation. There was a dramatic change in the themes regarded fit for doctoral dissertations from the middle of the nineteenth century to 1920. Dissertations then were also outcomes of extensive archival research where the author could and should show his or her (some women were achieving the doctoral degree at the time) acquaintance with sources and with the appropriate methods. Academic professionalism aiming at firm knowledge had found its way of expressing itself in such specialized studies.

At the same time as the second phase of historical professionalism in the sense of recognition among equals was triumphing in the interwar years, a distrust in the validity of its central notion, the idea of secure knowledge, was winning ground among historians. The philosophical direction of the *Historismus* had long since brought about doubts as to the possibility of reaching to security in historical research. Its proponents had allied to the old tradition of hermeneutics, grown out from Bible studies, for acquiring a system of understanding. It contained two main requirements: pre-knowledge as the conceptual apparatus and direction of interest; and empathy as the means of acquiring an understanding of human thoughts and actions. Hermeneutics was regarded as a means to acquire more or less probable statements on the past and the most general formulations included an ambition to maximize the probability through research. Collingwood (1946) worked in this direction, and after World War II, Hans-Georg Gadamer brought out his influential interpretation of hermeneutics, *Truth and Method* (1960), which was later overshadowed in some circles by Paul Ricoeur's *Time and Narrative* (1983–85).

Common to these works was that many historians accepted them as important foundations for the work of the historian. This meant that a new professional identity arose among some historians, conflicting with the earlier established group identity based on firm knowledge. As long as a basic norm in the hermeneutical camp was to increase probability of historical statements through empirical research, the conflict could be kept within close bounds. After the 1980s, this was no longer possible. Various theoretical directions have taken the relativistic claim further. The debate in the 1980s and 1990s has gone out from Hayden White and his rhetorical analysis of historiography (White 1973, 1978). The analyses of his standpoint have been numerous, above all in the USA, and have led to a variety of relativistic positions as regards the theory of

history based on the assumption that history is basically a rhetorical and aesthetic discipline. At the same time, in Europe, postmodernism has influenced other philosophers of history. A far-going relativism in different forms thus has more or less superseded the earlier basis of professional identity among historians.

Social theories in the broad sense of theories of the social sciences have since long occasionally been used by historians. More regularly has this been the case from the 1960s and onwards. In economic history it was already earlier the rule to rely on and use theories of economics, but theories of management and organizations and sociological theories have been used to a large extent by historians in historical studies only in the 1970s and 1980s. The professionalism of economic historians has come to rely on the ability to apply such theoretical instruments as well as specific economic (statistical) methodology. The analysis of social structures and social dynamics has also developed through the use of social theory. This approach had one of its origins in the *Annales* circle, then most often directed towards demography and geography and their social consequences. More explicitly, social science theory in general, including sociological and organizational theory, was proposed as the proper instrument for historical research by the so-called Bielefeld school, consisting of a group of historians who had some relation to the university of Bielefeld in the 1970s.

What they proposed was in line with what had been done by several American historians earlier, such as Merle Curti and Stephan Thernstrom, but their efforts had not been coordinated into a program. The Bielefeld historians advanced the concept of a 'historical social science' (Kocka 1977, Wehler 1980). This type of history found much appeal in the 1970s and 1980s. Although parts of this enthusiasm for social sciences were criticized later, it is rather the theories used than the close link to social science (including social anthropology, which found growing appeal among historians) that have met opposition. The use of social science theories has thus been one source of professional identity among historians in the recent decades.

The conflict between ideals inspired by traditional approaches, relativism and social science has not been resolved. In fact historians have no longer (if they ever had) one source of professional identity. This is not only a consequence of the fights between different convictions within the academic community. All through the twentieth century historical professionalism has also been challenged from political quarters. Autocratic regimes all over the world have tried to make historians into docile instruments for the propaganda of the ruling ideology. Many have suffered for their professional ideals, others have accepted the norms of the state instead of the norms of the profession.

By the end of the twentieth century, when historians tried to base their professionalism on interpretative

theories, methodological rules and social science theories, all without general acceptance, the history of historiography seems to have met a new interest. Different standpoints to historical knowledge may unite in the interest in investigating the history of historiography, and it is equally important for all different standpoints. The role of this disciplinary self-understanding for a renewal of professionalization should not be overrated, however. The history of the discipline cannot meet the requirement of providing the norms through which a closure can be upheld against nonprofessionals and, as regards norms there is no unanimity. In many countries it seems that the professional closure has become softened in recent decades and no sharp line is often drawn between historians and persons of literature. If history is at a point of crisis as a discipline, as is frequently said, it might be a consequence of two facts. No single normative theory is prevailing and the profession seems to rely on the quasi-theoretical field of its own history to reproduce a common professional identity.

See also: Historicism; History: Forms of Presentation, Discourses, and Functions; History: Overview; History: Theories and Methods; Professionalization/ Professions in History; Professions, Sociology of

Bibliography

- Bernheim E 1889 *Lehrbuch der Historischen Methode*. Dunker and Humblot, Munich, Germany
- Blanke H W 1991 *Historiographiegeschichte als Historik*. Frommann-Holzboog, Stuttgart-Bad Cannstadt, Germany
- Blanke H W, Rüsen J (eds.) 1984 *Von der Aufklärung zum Historismus. Zum Strukturwandel Historischen Denkens*. f. Schöningh, Paderborn, Germany
- Boer P den 1985 *Geschiedenis als Beroep: De Professionalisering van de Geschiedbeoefening in Frankrijk (1818–1914)* [History as a Profession. The Study of History in France]. SUN, Nijmegen, The Netherlands
- Charle C 1995 Être historien en France: une nouvelle profession? In: Bédarida F (ed.) *L'Histoire et le Métier d'Historien en France 1945–1995*. La Maison des Sciences de l'Homme, Paris, pp. 21–44
- Collingwood R G 1946 *The Idea of History*. Clarendon Press, Oxford, UK
- Droysen J G 1868 *Grundriss der Historik*. Leipzig, Germany
- Erdmann K D 1987 *Die Ökumene der Historiker. Geschichte der Internationalen Historikerkongresse und des Comité International des Sciences Historiques*. Vandenhoeck and Ruprecht, Göttingen, Germany
- Gadamer H-G 1960 *Wahrheit und Methode: Grundzüge einer Philosophischen Hermeneutik* [Truth and Method]. Mohr, Tübingen, Germany
- Heilbron J 1990 *Ontstaan van de Sociologie* [The Rise of Social Theory]. Prometheus, Amsterdam
- Iggers G 1984 The 'Methodenstreit' in international perspective. The reorientation of historical studies at the turn from the nineteenth to the twentieth century. *Storia della Storiografia* [History of Historiography], 6: 21–32

- Kocka J 1977 *Sozialgeschichte: Begriff, Entwicklung, Probleme*. Vandenhoeck und Ruprecht, Göttingen, Germany
- Langlois Ch-V, Seignobos Ch 1898 *Introduction aux Études Historiques*. Hachette, Paris
- Macintyre S 2000 History in a new country. Australians debate their past. In: Torstendahl R (ed.) *An Assessment of Twentieth Century Historiography*. Royal Academy of Letters, History and Antiquities, Stockholm, Sweden, pp. 71–86
- Meinecke F 1936 *Die Entstehung des Historismus*. R. Oldenburg, Munich, Germany
- Novick P 1988 *That Noble Dream: The 'Objectivity Question' and the American Historical Profession*. Cambridge University Press, Cambridge, UK
- Oexle O G 1996 *Geschichtswissenschaft im Zeichen des Historismus*. Vandenhoeck und Ruprecht, Göttingen, Germany
- Raphael L 2000 Flexible response? Strategies of academic historians towards larger markets for national historiographies and growing scientific standards. In: Torstendahl R (ed.) *An Assessment of Twentieth Century Historiography*. Royal Academy of Letters, History and Antiquities, Stockholm, Sweden, pp. 127–47
- Reill P H 1975 *The German Enlightenment and the Rise of Historicism*. University of California Press, Berkeley, LA
- Ricoeur P 1983–85 *Temps et Récit [Time and Narrative]*. Le Seuil, Paris, Vols. 1–3
- Scholtz G (ed.) 1997 *Historismus am Ende des 20. Jahrhunderts. Eine Internationale Diskussion*. Akademie, Berlin
- Smalley B 1974 *Historians in the Middle Ages*. Thames and Hudson, London
- Stieg M F 1986 *The Origin and Development of Scholarly Historical Periodicals*. University of Alabama Press, Tuscaloosa, AL
- Wang Q E 1995 Time perception in Ancient Chinese historiography. *Storia della Storiografia* 28: 69–86
- Wehler H-U 1980 *Historische Sozialwissenschaft und Geschichtsschreibung: Studien zu Aufgaben und Traditionen deutscher Geschichtswissenschaft*. Vandenhoeck and Ruprecht, Göttingen, Germany
- White H 1973 *Metahistory. The Historical Imagination in Nineteenth-century Europe*. Johns Hopkins University Press, Baltimore, MD
- White H 1978 *Tropics of Discourse. Essays in Cultural Criticism*. Johns Hopkins University Press, Baltimore, MD

R. Torstendahl

History: Theories and Methods

Since history turned into an academic discipline in the nineteenth century and professional historians claimed a privileged epistemological status for their approach vis-à-vis non-professional approaches to history, debates about the role of methods and theories in history have been characterized by a certain conjuncture. In periods of relative uncertainty and crisis about the status of history as a distinctive, academic discipline—such as the periods between 1880 and 1920 and 1965 to 1985 in the West—a substantial body of historians

participated in these debates. In those periods, *Methodenstreit* is more or less intense and visible, while in periods of relative certainty and academic recognition, this type of debate usually recedes into the background. So, seen from a historical point of view, reflection on the role of methods and theories in history is in a sense a symptom of crisis; at least outside the small group of historians and philosophers who permanently reflect on history in their capacity as philosophers of history. This direct connection between a 'reflective mood' among professional historians and their sense of professional uncertainty makes sense because, as will be elucidated below, method and/or theory have usually been presented as the discipline's cognitive foundations, especially its claim to produce 'objective' knowledge of the past. Therefore, doubts and debates about method and theory in history are always doubts and debates about the discipline's cognitive status, its academic credentials, and its claim to objectivity.

Seen from the point of view of the sociology of science, debates about the role of methods and theories in history have usually been directly linked to battles waged by the proponents of different conceptions of history for academic legitimacy and supremacy, that is, for academic recognition and reputation. In periods of serious competition for 'intellectual capital' (Bourdieu) between several competing groups, operating within the same 'academic field,' debates about methods and theories in history have been intense. This not only holds true for history, but also for other social sciences such as sociology, economics, psychology, etc., and, therefore, are a characteristic of pluralistic or poly-paradigmatic disciplines.

Characteristic for the periods of increased interest in theory and method is also an increased interest in the history of the discipline in case from both an intellectual and from an institutional angle. This interest in disciplinary history is usually motivated by a desire to locate the origins of the preferred conception of historical theory and method in the discipline and to generate academic credentials by locating prestigious founding fathers or precursors of the preferred conception in the discipline's pedigree. Therefore, in history, an increased interest in the history of history writing alias historiography can usually be observed alongside an increased interest in theory and method: theoretical and methodical reflections usually are accompanied by reflections on the origins, nature, and variety of historiography.

Seen from a philosophical point of view, debates about the role of theories and methods in history often have a blurred character due to the vague philosophical status of the central concepts of theory and method. Therefore, some elementary conceptual clarifications are called for in order to answer the question why theory and method in history have turned into a heavily contested area at times since history turned into an academic discipline.

1. Theory and Method in History: Some Preliminary Distinctions

In discussing the role of theory and method in history, it is necessary to introduce some basic distinctions. Due to the double meaning of history as *res gestae* and as *historia res gestae* (see *History: Overview*), theory in history can refer both to history as a process and to the knowledge of that process. Therefore, the first basic distinction to make is the distinction between:

(a) theories, which deal with characteristics of history as a process (such as, for instance, Marx's theory of history as class struggle); and

(b) theories, which deal with the characteristics of knowledge of history (such as, for instance, that scientific knowledge of history has the form of law-like knowledge or the form of concrete particulars).

Theories of type (a) can be called substantial or material theories of history, while theories of type (b) can be called cognitive theories of history, i.e., theories of historical knowledge.

Cognitive theories of history can further be subdivided into:

(a) epistemological theories and

(b) methodological theories, while substantial theories of history can also be subdivided in finer subsets (see *History: Forms of Presentation, Discourses, and Functions*).

Substantial and cognitive theories usually have been interrelated, because presuppositions about what can be known about the past are linked to presuppositions of how the past can be known. Substantial theories, for instance, which hold that the historical process is characterized by underlying mechanisms and regularities, analogous to theories of classical mechanics, have been interrelated with cognitive theories that posit that real knowledge of history implies knowledge of general laws (see *Positivism, History of*). And the substantial theory that the historical process is characterized by a fundamental chaos or contingency and not by hidden regularities alias a 'motor,' has been interrelated with cognitive theories which posit that real knowledge of history is knowledge of particular facts instead of general laws (see *Historicism*).

Substantial theories, in this way, not only have consequences for methodological positions, but also for epistemological positions. Substantial theories, which posit some kind of affinity or even identity of the knowing subject and the object of history in contrast with nature (as was the case with Hegelian idealism and its offshoots in historicism, such as Droysen and Dilthey's versions of historicism), have usually been interrelated with methodological theories. These posit a distinct kind of method for the human sciences—such as understanding, or *Verstehen*, or interpreting—in contrast with the methods of the natural sciences (usually defined as some form of causal or law-like explanation). And these distinct methodological theories have, in their turn, usually been inter-

related with distinct epistemological theories, linking methods of understanding to hermeneutic or phenomenological theories of knowledge and methods of explanation to empirical theories of knowledge. Before taking a better look at this terrain, however, it is first necessary to clarify the central concepts of method and theory as well as their interrelationships in history. When analyzed more closely, both method and theory in history appear to have a broad variety of meanings.

2. Historical Method(s): Concept and History

Analyzing the meaning of 'method' in history, it is useful to start with a clarification of the term 'method' in other disciplines first (Meran 1985). In the mathematical sciences it is common to refer to axiomatic-deductive, to infinitesimal, and to inductive methods, meaning a set of formal, technical procedures in order to solve problems within a specified range. Outside the mathematical sciences, the meaning of method is usually far less defined. In philosophy, for instance, there is a long tradition referring to methods of all sorts, ranging from Descartes' discourse on method to Kant's transcendental method, from Marx's dialectical method to Husserl's phenomenological method, and to De Saussure's structuralist method, to name just a handful. These 'methods' in philosophy, however, do not constitute defined formal rules for specified ranges of problem solving, but constitute complex sets of ideas, concepts, and rules of thumb on how to analyze specified domains, such as:

- (a) scientific knowledge;
- (b) the human consciousness;
- (c) the history of mankind;
- (d) the meaning of concepts;
- (e) the meaning of language in use; or
- (f) language as a system of signs.

'Method' in philosophy, thus, is definitely not identical with a set of formal techniques, but a loosely circumscribed research program. This diffuse and nonformal meaning of method is also characteristic of the human sciences.

In history, the term historical method was first introduced in a systematic way in the sixteenth century by Jean Bodin in his treatise of source criticism, *Methodus ad facilem historiarum cognitionem* (1566). Characteristically, Bodin's treatise intended to establish the ways by which reliable knowledge of the past could be established by checking sources against one another and by so assessing the reliability of the information conveyed by them, relating them to the interests involved. Historical method was thus at the same time intended to defeat the skeptical doubts with regard to the possibility of reliable knowledge of the past. Skepticism (alias historical Phyrhronism) had been fed by a large quantity of scriptural forgeries parading as original historical documents, primarily produced by competing churches and nobles to back

up their claims to rights and property. Therefore, the traditional claims to impartiality, truth, and trustworthiness of historians, dating back to classical Antiquity, were no longer believed, and were now replaced by a claim to impersonal and intersubjective method.

This direct relationship between history's claim to truth and method would characterize all later treatises on historical method, although what was meant by method did change over time. In the eighteenth century, historians of the Enlightenment, such as Schlözer, included not only the source critical techniques as the historical method (*Forschung*), but also the way the results of historical research were synthesized into a textual whole (*Darstellung*) and presented to an audience. This broader conception of method had its roots in the classical rhetorical tradition and would only recently be rediscovered in historical theory, that is from the 1970s onwards. The German historian and historical theorist Johann-Gustav Droysen (Droysen 1977) has become one of the pivotal historians behind this 'rediscovery' of the textual and rhetorical dimensions of history, because his treatment of historical method in the earlier and the later versions of his *Historik* embodies the transition from the broader to the narrower conception of historical method (Meier and Rüsen 1988).

The larger part of the nineteenth and twentieth centuries would be dominated by the research-oriented conception of historical method of the so-called Historical School in Germany, led by historians as Leopold Ranke and Berthold Niebuhr. Their conception of history, long been regarded as the beginning of modern, 'scientific' history, harked back to the 'narrow' conception of historical method, limiting the methodical character of history to source criticism. The historical composition and presentation—the *Darstellung*—was no longer regarded as methodical and was now seen as the literary, artistic, or aesthetic component of history. Therefore, the question of whether history belongs to the arts or to the sciences could and did arise from the nineteenth century onwards, a question that became part of the continuing *Methodenstreit* (Stern 1970, Fay et al. 1998).

In the nineteenth century, during the process of rapidly increasing disciplinary differentiation and academic institutionalization, it became customary in the human sciences to identify disciplines with one specific method, which was presented as the guarantee of its 'truth' and 'objectivity.' In this process, sociology was identified with the sociological method and history with the historical method. Canonical formulations of this conception of historical method can be found in E. Bernheim's *Lehrbuch der historischen Methode* (1889) or in Ch. Langlois and Ch. Seignobos' *Introduction aux études historiques* (1897).

Historical method in the nineteenth century usually referred to the techniques on how to conduct historical research of written materials, the so-called sources.

These techniques were traditionally subdivided in three parts, corresponding to three subsequent steps in research:

(a) heuristics, that is the techniques how to locate the relevant sources;

(b) source criticism, the techniques by which the temporal and spatial origins of the sources are established as well as their authenticity; and

(c) interpretation, by which the 'cleansed' information derived from the sources is put together by means of interpretation in order to infer 'what really happened.'

Usually a distinction was made between primary and secondary sources. Primary sources consist of the written material produced more or less contemporaneously with the object of research, ideally recorded by eyewitnesses (model: the official records of parliamentary debates). The time span between primary sources and what they represent is thus as small as possible, because distance in time is regarded as a major source of unreliability of the information conveyed. Secondary sources consist of records written a considerable time after the fact and not by eyewitnesses (model: 'hearsay'). As the information in secondary sources is based on other sources, it is, at best, 'second hand'; therefore, it is regarded as less reliable than primary source information. According to the dominant, Western conception of history since the nineteenth century, serious academic history can only be based on primary sources, because those are the only reliable source of information on the past. Therefore, the archives (see *Archives and Historical Databases*) were usually considered the workshop of the professional historian. According to the same nineteenth century Western conception, non-scriptural or 'low' cultures were primitive by definition and simply had no history comparable with 'high cultures.' In this way the scriptural, source based conception of historical method defined, and limited, the object of academic history to the history of 'high cultures,' especially the history of the West. It would last until the second half of the twentieth century before these Western biases in the very conception of historical method would be seriously criticized, especially by postcolonial theory (see *Subaltern History*).

Another traditional distinction was that between internal and external source criticism. Internal source criticism is actually textual criticism, i.e., the hermeneutic procedures by which a text is reconstructed (if it is passed down in an incomplete form or in several variants) as a whole and the textual procedures by which authorship and textual authenticity are checked and its meaning established. External source criticism consists of those procedures by which a source is checked by information derived from other sources, written and material. The so-called auxiliary sciences, such as paleography, chronology, sillography, toponymy, numismatics, and diplomatics, usually play an important role in this phase when historians are

dealing with older periods. And when modern historians are dealing with nonscriptural, oral cultures or with cultures that left few or no scriptural traces, archeology and anthropology often function as auxiliary science. Reference to multiple methods within one discipline became the rule only in the course of the twentieth century, thereby reopening the problem of truth and objectivity and paving the way for a new wave of skepticism. For if the method can no longer be regarded as the procedural guarantee for the truth and objectivity of the discipline, what could, so an influential argument ran (Bernstein 1983, Megill 1994)? Recognition of plural methods within one discipline, therefore, has often been regarded as a threat of the disciplines academic credentials.

From the late 1960s onwards, the process of methodological differentiation within all of the human sciences took an unprecedented pace, leading to a temporary fragmentation of most disciplinary fields and a questioning of most traditional disciplinary boundaries and self definitions. Significantly, the new, 'poly-paradigmatic' self-image was at the same time regarded by many as a crisis of their discipline. This also held for history, in 1960s and 1970s regarded by many as one of the social sciences (Novick 1988, Appleby et al. 1994, Iggers 1997 see *History and the Social Sciences*).

Historians identified the quantitative method (see *Quantification in History*), the biographical method (see *Biography: Historical*), the comparative method (see *Comparative History*), the ideal typical method, understanding as a method (*Verstehen*), the causal method (*Erklärung*), discourse analysis as a method (see *Linguistic Turn and Discourse Analysis in History*), the micro-historical method and a psychoanalytical method in history (see *Psychohistory*). Other historians claimed that the concept of method had become severely overstretched, holding that most advertised methods in fact were theories about historical knowledge, which could only be checked by the historical method of source criticism. Therefore, a closer look at the relationship between historical method and theories is necessary.

3. Theories of Historical Knowledge

3.1 Empiricism and Anti-empiricism

The connections between historical method and historical theory can be located both on the level of epistemology and on the level of methodology. As epistemology is the theory of knowledge, historical epistemology deals with the theory of historical knowledge, i.e., the conditions of its possibility, range, sources, and foundations. Two epistemological positions have been defended in relation to historical knowledge from the nineteenth century onwards, empiricism and anti-empiricist idealism.

The problem of whether and how it is possible to acquire knowledge of other minds, i.e., the mental states of others, such as thoughts and intentions, separated from the knowing subject in both time and space—the typical problem of the historian—has been a fundamental one since mid-nineteenth century. According to empiricism, mental states could only be known as far as they correlated with empirically observable epiphenomena. The empirical dispositions or expressions of other minds are seen as the only access to other minds (Hempel 1966).

The proponents of phenomenological and hermeneutic views (hermeneutics is the theory of interpretation, originally of text interpretation), to the contrary, have always affirmed this possibility of a non-empirical access to mental phenomena. Phenomenology claimed introspective methods as the access to the mental universe; hermeneutics, far more influential in theory of history than phenomenology, claimed that 'reading' other minds could be conceived of as analogous to reading texts.

The fundamental cognitive activity in all hermeneutic theories is the interpretation of meaningful symbols and symbolically structured phenomena (like actions and cultural objects), whereas all empiricist theories regard the observation of objects as the fundamental cognitive activity. Just as it is possible to know the ideas expressed in a text by reading and interpreting, it is possible to know the ideas expressed in actions and in all human made objects by inferring the unobservable ideas from their observable expressions through re-enactment and interpretation by the historian (Droysen 1977, Dilthey 1981, Collingwood 1993). This possibility was guaranteed, according to the hermeneutic views, by the basic unity of humanity, implying the basic unity of subject and object of knowledge in the human sciences. Therefore, each alter was an alter ego, accessible through a process of intuitive identification and transfer in thought or in idea (therefore: idealism). So in historical epistemology, just as in the other human sciences, the basic divide between 1850 and 1970 has been the one between empiricist and anti-empiricist, idealist positions. It would last until the 1980s, with the rise of postmodernism and its offshoots in the discipline of history, before the unitary presuppositions of hermeneutic theories would be seriously criticized and questioned. Western (white) male was no longer accepted by all as the model of all humankind. This has led to a renewed emphasis on the basic diversity of knowing subjects and their epistemological and methodological relevance for disciplines like history (Scott 1991) (see *Gender History; Subaltern History and Linguistic Turn and Discourse Analysis in History*).

3.2 Unity vs. Duality of Method

Between 1850 and 1970, this basic divide in the domain of epistemology roughly corresponded with the basic

divide in the domain of historical methodology, i.e., the divide between the proponents of the theory of the unity of scientific method and the proponents of the theory of a fundamental methodological dualism of the human and of the natural sciences.

Methodology as such aims at the rational reconstruction of the normative rules, which are constitutive for a specific discipline and its development. Rational reconstruction, in contrast with historical reconstruction, presupposes that disciplinary practices and their development can be reconstructed in terms of logical and empirical arguments; this presupposition is, of course, an idealization of the rationality of real disciplinary practice, because apart from rationality, power and privilege also play their role in reality. Methodology abstracts from those non-rational factors in disciplinary life and aims to reconstruct the cognitive criteria and ideals, by which the practitioners of a discipline judge one another's products and evaluate their relative quality. Methodology, therefore, presupposes the existence of a basic cognitive consensus and rationality among the practitioners of a discipline. This presupposition, of course, has been criticized both for the natural as for the human sciences (Goodman 1978, Rorty 1986, Novick 1988).

Next to the basic ideal of truth, other cognitive ideals of a discipline are reconstructed by its methodology, such as the explanatory ideals of the discipline and its criteria of progress. At this level the basic methodological divide between the proponents of the unity of scientific method and the proponents of the dualism of scientific methods must be located, because the divide revolves around the question whether there is one explanatory ideal for all disciplines or not. This issue has dominated the methodological debates for more than a century since 1850 and is known under the labels of the explanation vs. understanding controversy (Gardiner 1959, Hausmann 1991). According to the unity of science view, also often labeled as positivism, there was just one model or logic of explanation and development for all empirical disciplines. Classical mechanics in the nineteenth century was usually considered the paradigm for all other disciplines, including human sciences like history. The only difference between the various disciplines was their phase of development. Underdeveloped disciplines (like history) showed little similarity to Newtonian mechanics while developed ones did. Immature disciplines only produced 'explanation sketches' (Hempel 1966), which had to be developed into real explanations. The only difference between disciplines was thus a matter of scientific maturity and not a matter of explanatory logic.

Real scientific explanation boiled down to subsuming an event to be explained under a general law. Therefore, the event to be explained can be logically deduced from the general law in combination with the given circumstances (the so-called initial conditions) and is at the same time, given the law and the initial

conditions, predictable. According to this view on the logic of scientific explanation, that is also known under the headings of the subsumption model, the deductive nomological model, or the covering law model of explanation, explanation and prediction in science are basically the same; explanation is really prediction in the past alias retrodiction. Scientific explanation, in this view, is showing why what happened, was—given the known laws and circumstances—expected to happen because it was bound to happen: explanation boiled down to show what happened was logically necessary. Given the fact that the general laws were usually (though not exclusively) interpreted as causal laws, formulating universal and invariant connections between causes and effects, scientific explanation was usually seen as causal explanation and explanatory necessity also as causal necessity: scientific explanation specifies the causes of what happened by formulating its causal determinations.

Now although this unitary theory of scientific explanation was amended and modified in time, in order to account for statistical and probabilistic forms of explanation in the natural sciences, its status was based on the assumption that 'positivist' theory at least provided an adequate reconstruction of natural science and its development from the seventeenth century onwards. This assumption has been increasingly undermined by postempiricist and postpositivist philosophies of science roughly since the 1960s and 1970s. Both the empiricist underpinnings of the covering law model and its logical claims were undermined by historical examples, arguing that the supposedly paradigmatic cases in the history of science did not fit the basic logical models of philosophy of science. Therefore, this 'turn' in the philosophy of science is often called the historical turn, because history was mobilized against explanatory logic (Lakatos and Musgrave 1970). However, before examining the implications of postempiricism and postpositivism in the philosophy of science for the opposition between classical unitary and dualistic methodological theories, an overview of the dualistic positions and arguments is called for.

3.3 *The Autonomy of History*

The philosophical defense of the autonomy of historical explanation vis-à-vis the explanatory models of natural science usually has started off with a critique of the unitary claim. From Droysen and Dilthey, through Windelband, Rickert, Weber, Collingwood, and Gadamer, to White and Ricoeur, the argument has been developed that historical explanation is not only empirically different, but logically different from the covering law model. This claim of autonomy has

traditionally been defended in three variants: an ontological version, an epistemological version, and a methodological version.

The ontological argument for methodological dualism was first systematically developed by Droysen and later by Collingwood and his school. It usually harks back to Vico and Hegel, because it is based on an idea of a subject-object identity in history: in history, humankind confronts themselves, and history, therefore, is the autobiography of humanity. History is fundamentally the process in which humankind expresses and develops themselves in the form of diverse cultures; and because humankind has made history, they are capable of knowing history in a direct way, according to this argument, just as the writer of an autobiography knows the history in case in a direct way. Historians try to reconstruct cultures and the particular ideas expressed in its cultural practices (such as religions and rituals) and its material artifacts (such as its temples and sacred objects). When historians offer explanations for the past, they construct intelligible and plausible connections between the culture and the ideas expressed in it on basis of their research of the sources in case. When they succeed in this task, they offer an understanding of the phenomena in case (*Verstehen*). Understanding history is conceived of as similar to understanding a text; the major difference is that the texts of history have more or less disappeared and therefore must be reconstructed first based on their traces by research. This type of intelligibility is possible, because in history, human beings confront themselves and therefore intuitive understanding is possible in contrast to the situation in the sciences of nature, where the knowing subject confronts an external object.

The intelligibility sought for by historians—historical understanding—is thus centered on connecting ideas to their expressions in particular spatio-temporal contexts, and proceeds by going back and forth between parts and whole in the so-called hermeneutic circle or spiral. Historical understanding is therefore logically different from scientific explanation: the type of connections, which historians present, is not causally determined and cannot be deduced from general laws. Historical connections are not necessary but contingent connections, which always could have been different from their actual form.

Neither are they based on empirical observation, but on plausible inferences, based on critical research of the sources and on analogical reasoning. Historical understanding at best produces plausible and probable knowledge and never certain or founded knowledge, as scientific explanation, based on sensory experience and logic presumably does. And being dependent on intellectual, interpretative activities by the historian, which can be trained but not formalized, historical understanding, thus, remains 'subjective,' i.e., embedded in the culture, time, and place of the historian. Based on this argument, Gadamer in his

Truth and Method (1972) would radically historicize the idea of historical understanding.

The second type of argument in defense of methodological dualism is not ontological, but epistemological. It was first systematically developed by Wilhelm Dilthey and bears a resemblance to Droysen's argument, because it also harks back to Hegel's idea of a subject-object identity in history with its implied 'history as the autobiography of mankind model.' However, Dilthey does not ground historical method directly in the subject-object structure of the historical process itself, but in the 'inner experience' (*innere Erfahrung*) of historical facts as 'mental facts' (*geistige Fakten*). The 'mental facts' of history and of other *Geisteswissenschaften* consist of feelings, thoughts, and strivings of human beings living in the past, which are only accessible through the mind and the inner experience of the historian. Like Droysen, Dilthey posits that the historian makes the mental facts of the past intelligible as expressions of ideas, by putting them in a particular spatio-temporal context alias in a *Lebenszusammenhang*. And, like Droysen, he contrasts historical understanding with scientific explanation by rejecting causal, law like reasoning for history. But unlike Droysen, Dilthey emphasizes the contrast of the 'inner experience' of the historian with the 'outer experience' of the scientist as the foundation of methodological dualism.

The third type of argument in defense of methodological dualism is of a methodological character. It is not based on an ontological claim that history deals with an object fundamentally different from nature, and neither is it based on an epistemological claim that the historical facts are different from natural facts, being accessible to an 'inner' way of knowing. According to the methodological argument, as developed by Windelband, Rickert, and Weber, two fundamental ways of knowing or methodological approaches can be distinguished, which are connected by two fundamentally different aims or knowledge interests. The first way to approach an object is a generalizing method; in this case, one abstracts from the particularity of the objects of knowledge and tries to detect and explain its general characteristics. According to Windelband, this generalizing method was specific for the nomothetic sciences (Greek: *nomos* = law, *tithemi* = to posit), such as the natural sciences.

The second fundamental approach to an object is the particularizing method; in this case one abstracts from the general characteristics of an object of knowledge and tries to describe its specific characteristics. According to Windelband this particularizing method was specific for the ideographic sciences (Greek: *ideos* = particular, *graphein* = to describe), such as history. However, both the generalizing and the particularizing methods are independent of the nature of their objects of knowledge; nature can be studied in an ideographic way (for instance, in geology) and history can be studied in a nomothetic way

(for instance, in developmental psychology). It all depends on one's methodological interest in case.

Rickert and Weber would develop Windelband's basic ideas by introducing the dichotomy between sciences of nature (*Naturwissenschaften*) and sciences of culture (*Kulturwissenschaften*). Rickert introduced the idea that knowledge interests behind disciplines were guided by values (*Wertbezogenheit*); Weber pointed out that values usually change through time. According to him, the change of the guiding values behind the sciences of culture could explain why those disciplines did not accumulate knowledge in the systematic form like the natural sciences, and why they seemed to enjoy an 'eternal youth' (Weber 1973). At best, the sciences of culture could accumulate knowledge at the intermediate level between individualizing descriptions and universal explanations in the form of ideal types, according to Weber. He suggested comparison as the method at this intermediate level. Lübke, Veyne, and Mink would develop the Weberian methodological position later, identifying history with the particularizing methodological interest (Lübke 1977, Veyne 1978, Mink 1987).

All in all, different conceptions of historical method are connected to different theories of historical knowledge. Since the characteristics of historical knowledge are usually defined in contrast with scientific knowledge as the major 'contrast class,' theories of natural science are also presupposed in theories of historical knowledge. Fundamental changes in theories of natural science, therefore, affect theories of historical knowledge by detour. The rise of postempiricism, postpositivism, and of the historical turn in the philosophy of natural science, thus have undermined most traditional, classical defenses of methodological dualism in an indirect way by undermining the classical unitary model of science. A few remarks must suffice.

Postempiricism has effectively undermined the presupposition that scientific knowledge can ever be grounded on certain foundations in sensory experience. The opposition between grounded and certain knowledge at the one side—traditionally attributed to the natural sciences—and inferential and probable knowledge at the other—traditionally attributed to the human sciences—has thus vanished. This opposition has been replaced by the insight that all sensory experience is guided and framed by concepts and theories. All empirical knowledge, therefore, is based on interpretative activities of the knowing subject because it is 'theory-laden.' Therefore, all knowledge is fallible and open to later reinterpretations. Postempiricism, thus, has dissolved the presupposed absolute contrast between a 'subjective' kind of knowledge, based on interpretation—characteristically for the *Geisteswissenschaften*, the *Kulturwissenschaften*, or the historical sciences—and an 'objective' kind of knowledge, based on empirical observation and experiment, not in need of interpretation. Postempiri-

cism has thus relativized both the distinction between theoretical knowledge and empirical knowledge, at the one side, and the distinction between interpretative knowledge and empirical knowledge, at the other. Thus, it has brought conceptualization and interpretation to the center of the stage in all disciplines. In doing so, postempiricism has undermined traditional underpinnings of the idea of fundamentally distinct scientific and humanistic methods and the idea of a fundamental contrast between the 'objectivity' of science and the 'subjectivity' of the humanities. Both scientific domains aim at empirical argument and inter-subjectivity (Megill 1994, Oexle 1998).

Postpositivism, based on postempiricism, in turn, has effectively undermined the presupposition that all scientific explanations can be reduced to one basic model, i.e., the subsumption model of explanation. By discrediting this presupposition for natural science, (Salmon 1990), postpositivism *a fortiori* has discredited the subsumption model for the humanities. Moreover, the fact that the explanation of mental states by nonmental states has not succeeded up to the present has also discredited unitary views. (Bohman 1991).

In the theory of history, postempiricism and postpositivism have led to a renewed reflection on the nature of historical knowledge in postmodernism and in a renewed interest in approaches inspired by literary theory (see *History and Literature*). Now the classical science model was discredited, the roots of history in rhetoric and literature were rediscovered, leading to a renewed interest in the linguistic forms in which history is represented, i.e., in theories of historical representation. The postmodern vanguard, oft identified with the linguistic turn (see *Linguistic Turn and Discourse Analysis in History*), pushed both theory and method in history into the background, locating historical representation beyond epistemology and methodology, on the terrain of aesthetics, rhetoric, and politics (White 1973, Ankersmit and Kellner 1995, Fay et al. 1998).

This move, in fact, harks back to the premethodical and predisciplinary roots of history, cutting history loose from its claim to constitute a truth-seeking discipline and to produce 'objective' knowledge (Rüsen 1983–89, Lorenz 1997). The history of theory and methods in history, thus, has come full circle at the turn of the twentieth century. Nevertheless, the pendulum is already swinging back, leading to broader approaches of history, which keep up the cognitive claims of the discipline based on both research (*Forschung*) and composition (*Darstellung*), while integrating the literary and the practical dimensions of history (see *History: Forms of Presentation, Discourses, and Functions*).

Bibliography

Ankersmit F R, Kellner H (eds.) 1995 *A New Philosophy of History*. London

- Appleby J, Hunt L, Jacob M 1994 *Telling the Truth about History*. New York
- Bernheim E 1889 *Lehrbuch der historischen Methode*. Leipzig, Germany
- Bernstein R 1983 *Beyond Objectivism and Relativism: Science, Hermeneutics and Practice*. Oxford
- Bodin J 1595 *Methodus ad facilem historiarum cognitionem*. Geneva, Switzerland
- Bohman J 1991 *New Philosophy of Social Sciences: Problems of Indeterminacy*. Cambridge, UK
- Collingwood R G 1993 *The Idea of History*. Oxford, UK
- Dilthey W 1981 *Der Aufbau der geschichtlichen Welt in den Geisteswissenschaften*. Frankfurt, Germany
- Droysen J-G 1977 *Historik*. Stuttgart, Germany
- Fay B, Pomper P, Vann R (eds.) 1998 *History and Theory. Contemporary Readings*. Cambridge, UK
- Gadamer H G 1972 *Wahrheit und Methode. Grundzüge einer philosophischen Hermeneutik*. Tübingen, Germany
- Gardiner P (ed.) 1959 *Theories of History: Readings from Classical and Contemporary Sources*. New York
- Ghandi L 1998 *Postcolonial Theory: A Critical Introduction*. New York
- Goodman N 1978 *Ways of Worldmaking*. Indianapolis, IN
- Hausmann T 1991 *Erklären und Verstehen. Zur Theorie und Pragmatik der Geschichtswissenschaft*. Frankfurt, Germany
- Hempel C 1966 Explanation in science and history. In: Dray W (ed.) *Philosophical Analysis and History*. New York, pp. 95–127
- Iggers G 1997 *Historiography in the Twentieth Century. From Scientific Objectivity to the Postmodern Challenge*. Hanover, USA
- Koselleck R 1989 *Vergangene Zukunft. Zur Semantik geschichtlicher Zeiten*. Frankfurt, Germany
- Kuhn T 1962 *The Structure of Scientific Revolutions*. Chicago
- Lakatos I, Musgrave A (eds.) 1970 *Criticism and the Growth of Knowledge*. Cambridge, UK
- Lorenz C 1997 *Konstruktion der Vergangenheit. Eine Einführung in die Geschichtstheorie*. Cologne, Germany
- Lübbe H 1977 *Geschichtsbegriff und Geschichtsinteresse. Analytik und Pragmatik der Historie*. Basel, Switzerland
- Megill A (ed.) 1994 *Rethinking Objectivity*. Durham, USA
- Meier C, Rüsen J (eds.) 1988 *Historische Methode*. Munich, Germany
- Meran J 1985 *Theorien in der Geschichtswissenschaft*. Göttingen, Germany
- Mink L 1987 *Historical Understanding*. Ithaca, NY
- Nagl-Docekal H 1982 *Die Objektivität der Geschichtswissenschaft*. Vienna
- Novick P 1988 'That Noble Dream.' *The 'Objectivity question' and the American Historical profession*. Cambridge, UK
- Oexle O-G (ed.) 1998 *Naturwissenschaft, Geisteswissenschaft, Kulturwissenschaft; Einheit—Gegensatz—Komplementarität?*. Göttingen, Germany
- Rickert H 1986 *Kulturwissenschaft und Naturwissenschaft*. Stuttgart, Germany
- Ricoeur P 1983–5 *Temps et récit* (3 Vols). Paris
- Rorty R 1986 *Philosophy and the Mirror of Nature*. Oxford, UK
- Rüsen J 1983–9 *Historische Vernunft. Grundzüge einer Historik* (3 vols.). Göttingen, Germany
- Salmon W 1990 *Four Decades of Scientific Explanation*. Minneapolis, MN
- Scott J 1991 Women's history. In: Burke P (ed.) *New Perspectives in Historical Writing*. Cambridge, UK, pp. 42–66
- Seignobos C 1897 *Introduction aux études historiques*. Paris
- Stern F (ed.) 1970 *Varieties of History. From Voltaire to the Present*. London

- Veyne P 1978 *Comment on écrit l'histoire?* Paris
- Weber M 1973 *Gesammelte Aufsätze zur Wissenschaftslehre*. Tübingen, Germany
- White H 1973 *Metahistory. The Historical Imagination in Nineteenth Century Europe*. Baltimore, MD
- Windelband W 1915 *Geschichte und Naturwissenschaft. In: Präludien. Aufsätze und Reden zur Einleitung in die Philosophie und ihrer Geschichte*, Vol. 2, Tübingen, Germany, pp. 136–61

C. Lorenz

HIV and Fertility

It is widely recognized that HIV had a catastrophic impact on mortality from the early 1990s in much of sub-Saharan Africa, and significant impacts in a few high prevalence countries of South East Asia. The interactions between HIV and fertility are not so well known, but they deserve the special attention of epidemiologists, demographers, and behavioral scientists since they influence our understanding of the epidemic trends and the sexual behavior which underlies this. Limited evidence of the impact of HIV on fertility is available from developed nations with low-level epidemics concentrated in high risk groups such as intravenous drug users, recipients of blood transfusions, and men who have sex with men, but this review concentrates on the situation in African countries with generalized, heterosexually transmitted epidemics. The fertility effects in these populations are much more clear cut, both because the high numbers involved make it relatively easy to collect representative data, and because historically low rates of contraceptive use make it feasible to distinguish between involuntary, biological effects and behavioral responses to the epidemic.

1. The Nature of HIV and Fertility Interactions

Early speculation on the likely effects of the HIV epidemic on fertility tended to concentrate on change in intentional behavior, and predicted resulting fertility increases. Some authors believed that increases in child mortality might fuel increases in childbearing activity, as parents would want to have many children to ensure that at least a few survived (Caldwell et al. 1992). Others feared that in their quest for safe, uninfected partners, men would seek partnerships with progressively younger girls, thereby lowering the starting age for childbearing. Later writing (e.g., Gregson 1994) has given much more weight to biomedical factors and to unintended consequences of changing behavior.

1.1 Biomedical Factors

HIV lowers fertility directly in several ways, not all of them yet fully understood. Most importantly, it causes

spontaneous abortions and stillbirths (Brocklehurst and French 1998), and in the later stages of the disease women may suffer from amenorrhea (Widy-Wirsky et al. 1988). In men, it may reduce spermatogenesis (sperm production) (Krieger 1991), and lower coital frequencies have been reported by couples in which the male partner has begun to suffer from opportunistic infections.

Fertility impairments are also caused by other sexually transmitted diseases (STDs), such as syphilis, chlamydia and gonorrhea. HIV is much more prevalent among those who have other STDs, partly because of common behavioral risk factors, partly because HIV-positive individuals are more vulnerable to these infections, but most importantly because the presence of ulcerative STD infections increases the likelihood of HIV transmission (Fleming and Wasserheit 1999). These direct biological factors affect only a minority of couples in the population, namely those in which one or the other partner is already infected with the HIV virus. Therefore they may not have a large impact on fertility at the population level, but they are important determinants of fertility differentials between HIV-positive and HIV-negative women.

1.2 *Biosocial Responses to the Epidemic*

A range of behavioral responses to the epidemic may alter fertility in the whole population, even though they affect HIV-positive women more than the HIV-negative. Studies in Zimbabwe have indicated higher rates of widowhood, higher divorce rates as a result of knowledge or suspicion of HIV infection in either partner, and lower remarriage rates (Gregson et al. 1997). Increased condom use was reported in Uganda, especially in casual relationships, and there was a decrease in extramarital sexual activity (Kilian et al. 1999) and a postponement of entry into sexual unions by adolescents. These factors would tend to decrease nonmarital fertility. Other behavioral changes in response to the epidemic may tend to cause increases in fertility—for example, in Zimbabwe, some women have chosen to curtail breastfeeding, since it is becoming more widely known that the virus may be transmitted to infants through breastmilk (Gregson et al. 1997). Traditional periods of postpartum abstinence may be shortened to curb the tendency for men to indulge in extramarital affairs—Cleland et al. (1998) showed that in Benin this was more likely to happen when the period of abstinence was lengthy.

1.3 *Behavioral Change Consciously Intended to Affect Fertility*

There is little evidence that insurance or replacement effects motivate fertility increases: on the contrary,

opinion surveys in Tanzania, Rwanda, and Zimbabwe show that the majority believe HIV-infected women should refrain from childbearing in order to limit vertical transmission of the virus and to avoid leaving behind large numbers of orphaned children (Setel 1995, Keogh et al. 1994, Gregson et al. 1997). Followup studies of women in Zaire who were aware of their HIV status have indicated more contraceptive use by the HIV-positive (Ryder et al. 1991), and higher rates of induced abortion have been reported in some developed countries (Thackway et al. 1997).

1.4 *Changes in Population Structure*

Changes in the birth rate can come about not only because of changes in age-specific fertility rates but also because of structural changes in the population. Gregson et al. (1999) showed that HIV-related mortality will tend to make the population structure younger, and with proportionately more women in the peak childbearing ages of 20–29 this would tend to push up the crude birth rate. Similarly, over time, the excess mortality suffered by HIV-positive women with primary or secondary sterility due to coinfection with other STDs would tend to raise the proportion of fecund women in the population and thereby raise the overall birth rate, even if the fertility rates of HIV-positive and HIV-negative women did not change. Fecundity could also rise if interventions to treat classical STDs were successful.

2. *The Scale of the Effect of HIV on Fertility*

Empirical evidence for fertility differentials between HIV-positive and HIV-negative women has accumulated from various longitudinal studies. It is summarized below. It is rather more difficult to find unambiguous direct evidence for population level effects, though approximate effects can be estimated by assuming that in the absence of HIV infection all women would experience the fertility patterns currently experienced by the HIV-negative. However, this does not allow for the fact that HIV-positive women may be selected for high levels of other STDs and recent involvement in sexual activity.

2.1 *Differences in Fertility Between HIV-positive and HIV-negative Women*

Figure 1 is based on data from the longest running observational study of HIV in a developing country setting, in the Masaka district of Southern Uganda (Carpenter et al. 1997). It shows the typical relationship between fertility in HIV-infected and uninfected women. At ages under 20, HIV-infected women have higher fertility than the uninfected, but at age 20 and

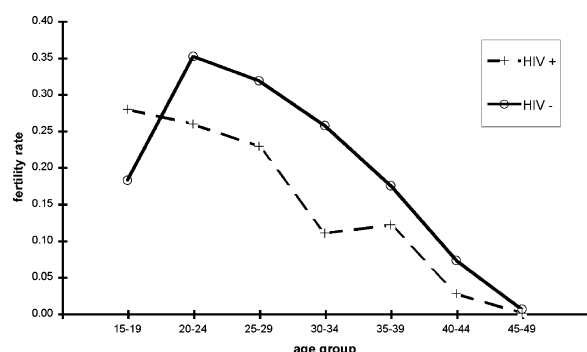


Figure 1

Age specific fertility by HIV status, Masaka, Uganda, 1989–1996

over HIV-positive women have significantly lower fertility than the HIV-negative. This is because in the youngest age group those women who have borne children are selected for an early start to sexual activity—virgins are not at risk of childbearing nor are they exposed to the risk of sexually transmitted HIV infection. At later ages, when almost all women are—or have been—sexually active, these selection effects are not so important, and the fertility-reducing effects of HIV, described above, clearly predominate.

2.2 Population Level Effects

Table 1 (adapted from Zaba and Gregson 1998) indicates the proportionate reduction in fertility attributable to HIV in a number of severely affected populations in Eastern Africa. HIV prevalence in these populations ranges from 8 percent to 30 percent, and fertility in the population as a whole is 4 percent to 12 percent lower than the fertility of HIV-negative women. This represents an average reduction of around 0.4 percent in total fertility for every percentage point increase in HIV prevalence. These observations are all based on studies in populations with low acceptance of family planning—the populations in Table 1 all have contraceptive prevalence rates of less than 15 percent.

Table 1

Population level effects of fertility reduction associated with HIV

	Mwanza, Tanzania 1992–1994	Kisesa, Tanzania 1994–1997	Masaka, Uganda 1989–1996	Rakai, Uganda 1994–96	Ft Portal, Uganda 1994–1995	Lusaka, Zambia 1995–1996	Mposhi, Zambia 1995–1996
HIV prevalence (%)	15.6	8.4	11.8	21.2	26.0	29.9	17.4
Total fertility							
HIV-women	4.29	6.41	6.80	4.75	7.95	5.69	7.40
all women	4.09	5.89	6.54	4.34	6.97	5.09	6.86
% attributable change	–4.7	–8.1	–3.8	–8.7	–12.3	–10.6	–7.3

There is some indication (Zaba et al. 2000) that fertility differentials between HIV-positive and HIV-negative women are not as pronounced in African countries with higher rates of contraceptive use, such as Kenya and Zimbabwe, where over 30 percent of women use modern contraception. However, in these populations fertility had already begun to fall before the HIV epidemic took hold, and it would appear highly likely that the impetus towards lower fertility in both HIV-positive and HIV-negative women would be reinforced by the behavioral changes described above. It would be difficult to ascertain whether fertility would have fallen more slowly in the absence of the epidemic. The ‘attributable change’ statistic is an indicator of the overall effect of differential fertility. It does not take account of the common secular trends for infected and uninfected women.

3. Consequences of HIV and Fertility Interactions

As a result of the overall fertility reduction in populations severely affected by the HIV epidemic, population growth rates will be even lower than would be expected on the basis of the dramatic increases in mortality. A fall of 10 percent in total fertility would cause a reduction of the order of half a percentage point in the growth rate, making it more likely that some of the worst affected countries could experience periods of negative growth. Since falling fertility causes population aging, this will counter the tendency of HIV-related mortality to make the age structure of these populations even younger.

4. Measurement of HIV Prevalence in Sentinel Surveillance Populations

One of the more problematic effects of the relationship between HIV and fertility is on the measurement of HIV prevalence. The most widely used source of prevalence data—indeed the source on which all national-level estimates of HIV prevalence are based (US Bureau of the Census 2000)—is sentinel surveillance in antenatal clinics. Since HIV positive women have lower fertility, they will be underrepre-

sented in antenatal clinics, so that the resulting estimates of HIV prevalence underestimate the true extent of infection among women of childbearing age. Where community-based data are available it has been noted that HIV prevalence among adult males is generally lower than that in the female population (UNAIDS 2000), but conventionally the estimates based on surveillance of pregnant women continue to be used to represent prevalence for adults of both sexes. To obtain reliable estimates of HIV prevalence levels in women of childbearing age, one needs to adjust prevalence data collected in antenatal clinics (Zaba et al. 2000), particularly if one seeks to derive estimates of HIV incidence from trends in prevalence in young women. Prevalence estimates for young women may also be affected by changes in the age of sexual debut.

4.1 Trends in Orphanhood and Child Mortality

Reduced fertility in HIV-positive women implies lower effects of HIV on child mortality than were originally anticipated, and fewer orphans will be left behind at the death of an HIV-positive mother than for one who is HIV-negative. At the individual level, many of the fertility reducing factors (such as HIV-associated illnesses or the likelihood of divorce and widowhood) strengthen with duration of infection. As a result, the downward pressure on fertility, and the associated secondary effects on orphanhood and child mortality, will change with the duration of the epidemic, as the balance between women who have been recently infected and those who have been infected for longer periods shifts in favour of the latter. Gregson et al. (1999) have modeled the effect that such trends will have on the number and age distribution of orphans and child deaths.

4.2 Family Planning in the Era of AIDS

Family planning programmes in sub-Saharan Africa have traditionally promoted hormonal contraception, namely oral contraceptive pills, injectibles, and implants. Barrier methods are regarded with suspicion, as they are considered to be less reliable than hormonal methods and because of their association with commercial and illicit sex. However, effective and systematic use of condoms for disease prevention would also make them highly effective contraceptives. Conversely, an acceptance of condoms for family planning purposes in stable marital unions would destigmatize them and make them more acceptable in other circumstances also. The AIDS pandemic will require a major shift in the contraceptive mix offered by family planning programs and a new willingness to discuss STD prevention in the context of family planning services.

See also: AIDS, Geography of; Fertility: Institutional and Political Approaches; Fertility: Proximate Determinants; Infant and Child Mortality in the Less Developed World; Reproductive Rights in Developing Nations

Bibliography

- Brocklehurst P, French R 1998 The association between maternal HIV infection and perinatal outcome: A systematic review of the literature and meta-analysis. *British Journal of Obstetrics and Gynaecology* **105**: 836–48
- Carpenter L, Nakiyingi J, Ruberantwari A, Malamba S, Kamali A, Whitworth J 1997 Estimates of the impact of HIV infection on fertility in a rural Ugandan population cohort. *Health Transition Review* **7**(suppl.): 113–26
- Caldwell J C, Orubuloye I O, Caldwell P 1992 Fertility decline in Africa: A new type of transition? *Population and Development Review* **18**: 211–42
- Cleland J, Ali M, Capo-Chichi V 1998 Post-partum sexual abstinence in West Africa: Implications for AIDS-control and family planning programmes. *AIDS* **13**: 125–31
- Fleming D T, Wasserheit J N 1999 From epidemiological synergy to public health policy and practice: The contribution of other sexually transmitted diseases to sexual transmission of HIV infection. *Sexually Transmitted Infections* **75**: 3–17
- Gregson S 1994 Will HIV become a major determinant of fertility in sub-Saharan Africa? *Journal of Development Studies* **30**: 650–79
- Gregson S, Zhuwau T, Anderson R M, Chandiwana S K 1997 HIV-1 and fertility change in rural Zimbabwe. *Health Transition Review* **7**(suppl.): 89–112
- Gregson S, Zaba B, Garnett G 1999 Low fertility in women with HIV and the impact of the epidemic on orphanhood and early childhood mortality in sub-Saharan Africa. *AIDS* **13** (suppl. A): S249–58
- Keogh P, Allen S, Almedal C, Temahagili B 1994 The social impact of HIV infection on women in Kigali, Rwanda: A prospective study. *Social Science and Medicine* **38**: 1047–53
- Kilian A, Gregson S, Ndyabangi B, Walusaga K, Kipp W, Salmuller G, Garnett G, Asiimwe-Okiror G, Kabagambe G, Weiss P, von Sonnenburg F 1999 Reductions in risk behavior provide the most consistent explanation of declining HIV-1 prevalence in Uganda. *AIDS* **13**: 391–98
- Krieger J N 1991 Fertility parameters in men infected with HIV. *Journal of Infectious Diseases* **164**: 464–69
- Ryder R W et al. 1991 Fertility rates in 238 HIV-1 seropositive women in Zaire followed for 3 years post partum. *AIDS* **5**: 1521–7
- Setel P 1995 The effects of HIV and AIDS on fertility in east and central Africa. *Health Transition Review* **5**(suppl.): 179–90
- Thackway S V, Furner V, Mijch A, Cooper D A, Holland D, Martinez P, Shaw D, van Beek I, Wright E, Clezy K, Kaldor J M 1997 Fertility and reproductive choice in women with HIV-1 infection. *AIDS* **11**: 663–8
- United States Bureau of the Census 2000 *HIV/AIDS Surveillance Data Base* (see also the companion web site, <http://www.census.gov:80/ipc/www/hivaidsn.html>). US Bureau of the Census, Washington, DC
- UNAIDS 2000 *The Status and Trends of HIV/AIDS in the World*. Technical report of the Monitoring the AIDS Pandemic Symposium. Durban, 13th International AIDS Conference, July 9–14, 2000

- Widy-Wirski R et al. 1988 Evaluation of the WHO clinical case definition for AIDS in Uganda. *Journal of the American Medical Association* **260**: 3286–9
- Zaba B, Gregson S 1998 Measuring the impact of HIV on fertility in Africa. *AIDS* **12**: S41–50
- Zaba B, Carpenter L, Boerma T, Gregson S, Nakiyingi J, Urassa M 2000 Adjusting ante-natal clinic data for improved estimates of HIV prevalence among women in sub-Saharan Africa. *AIDS* **14**: 2741–50

B. Zaba and S. Gregson

HIV Risk Interventions

Human Immunodeficiency Virus (HIV) infection, the cause of Acquired Immune Deficiency Syndrome (AIDS) has become a significant threat to global public health faster than any previous epidemic (Mann and Tarantola 1996). The genetic nature of HIV evades the development of a preventive vaccine and a cure for HIV infection remains a distant hope. HIV is transmitted through direct contact with HIV infected blood, semen, and vaginal secretions. Although HIV is transmitted during birth from mother-to-infant and through contaminated blood products the majority of AIDS cases in the world have resulted from HIV transmission between adults engaged in high-risk practices. Behavioral interventions therefore remain the most realistic means for curtailing the spread of HIV infection. Effective HIV risk reduction interventions target two principle behaviors: (a) sharing HIV contaminated drug injection equipment and (b) decreasing exposure to HIV infected semen, vaginal secretions, and sexually derived blood. Interventions to change injection equipment sharing and high-risk sexual practices can, therefore, dramatically effect the spread of HIV. In this article, factors associated with HIV transmission risks and interventions directed at reducing risks associated with injection drug use and sexual relations are examined.

1. Factors Associated with HIV Transmission Risks

Effective HIV risk reduction interventions are tailored to the immunological, epidemiological, cultural, and social correlates of HIV transmission risk behaviors. HIV transmission risk is affected by both biological and sociocultural co-factors.

1.1 Biological Co-factors

A variety of genetic subtypes of HIV exist, with each having its own unique features relevant to trans-

mission susceptibility. Genetic subtypes of HIV also differ in their geographic distribution. Therefore, genetic subtypes of HIV that are most efficiently transmitted during vaginal intercourse are more prevalent in some regions of the world than others and partially account for the regional differences in epidemic patterning (Mann and Tarantola 1996). Sexually transmitted infections, particularly syphilis, chancroid disease, and gonorrhea, cause ulcers and degradation of mucous membranes increasing both HIV transmission and susceptibility. Another biological factor associated with HIV transmission risk is the infectious state of infected persons. Individuals with lower blood concentrations of HIV, or viral load, may be at lower transmission risk than persons with greater concentrations. Therefore, the availability and use of HIV suppressing medications may influence the transmission of HIV. These biological influences on HIV transmission risks will likely interact with specific prevention strategies targeted by given risk reduction interventions.

1.2 Sociocultural Co-factors

Culture and social structures create a context for HIV transmission risk. Gender biases, power in sexual relationships, social norms supportive of risk and protective behaviors, poverty, drug addiction, and homophobia are known factors that influence HIV subepidemics (Kalichman 1998). HIV transmission occurs in personal relationships. Therefore, the social connections between persons, or social networks, play a critical role in both risk producing and risk reducing behaviors (Friedman et al. 1999). Social environments necessitate tailoring behavioral interventions to sub-population characteristics. Also, behavioral interventions must adapt over time to the ever-changing social context of HIV-AIDS.

2. Interventions for Injection Drug Use Risks

HIV and other diseases are spread through sharing contaminated needles and syringes for injecting drugs. Tainted injection equipment, therefore, not the drugs themselves transmit HIV. The use of sterile needles and syringes assures prevention against HIV and other diseases. Syringe access and exchange programs have demonstrated substantial reductions in new HIV infections among drug injectors (NIH Panel 1997). Another effective means of preventing HIV transmission via injection drug use is availability of drug treatment on demand. Obviously, when one stops injecting drugs the risks for injection-related HIV also ceases. Yet despite the known public health value of providing access to syringes and to drug treatment, political forces and limited economic resources have

prohibited policies for wide-scale implementation of these intervention strategies (National Academy of Sciences 1995). Cleaning injection equipment is considered a reasonable alternative when sterile syringes are unavailable. Behavioral interventions have shown success in training injection drug users in proper needle and syringe hygiene skills.

HIV risk reduction for injection drug users, therefore, occurs along a spectrum of health objectives. Drug treatment is the ideal intervention because it addresses the risks posed by drug addiction as well as HIV infection. Persons who do not receive drug treatment, however, can reduce the harm of drugs by using sterile injection equipment. When sterile equipment is unavailable, persons are advised to practice needle and syringe cleaning. The philosophy of accepting incremental steps to reduce risks when risks cannot be eliminated is called harm reduction and has been highly effective in framing the expected results of HIV risk reduction strategies.

3. Interventions for Sexual Relationship Risks

The vast majority of AIDS cases in the world have resulted from exposure to HIV in sexual relationships. Interventions designed to reduce sexual risk practices that permit HIV transmission—vaginal, anal, and oral intercourse without condoms or other barriers—have occurred at three primary levels: individual risk reduction counseling, small group workshops, and community mobilization campaigns. In each case, effective interventions that reduce high-risk behavior almost universally include three primary elements: educational information about HIV and how it is transmitted and prevented; motivation to induce individuals to make behavioral changes; and interactive exercises to improve risk reduction behavioral skills, such as sexual assertiveness skills and condom use skills (Fisher and Fisher 1992).

3.1 Components of Sexual Risk Reduction Interventions

Skills enhancement interventions consist of education, sensitization, and instruction in specific behavioral skills to reduce risk for HIV infection. Educational exercises and activities designed to help individuals take steps toward risk reduction are tailored to reflect the context, values, customs, and language of targeted populations. Therefore, similar intervention content is used for men and women of various backgrounds, but the contextual issues and examples vary accordingly.

3.1.1 HIV risk information and motivation enhancement. HIV prevention interventions provide basic in-

formation related to HIV transmission, illness, and prevention. Risk education can include facts about viral transmission, local prevalence of HIV and AIDS, clarification of misconceptions, dispelling AIDS myths, and description of HIV antibody testing. Discussions often couch information in personal life situations, relationships, risk behaviors, and increased motivations to reduce risk. Education also encourages individuals to assess their own behavioral risk histories and initiate discussions of individual life-situations that have placed them at risk. Theoretically, increased understanding of risk and personalization of vulnerability to contracting HIV are believed to increase one's readiness to change and enact behaviors to reduce risks.

3.1.2 Behavioral self-management skills. Behavioral self-management skills related to HIV risk reduction are acts that individuals can perform to reduce their risks for HIV infection with minimal involvement of their sexual partners. Individuals are instructed in techniques for examining their risk and cues that precede sexual risky situations. For example, persons are instructed in ways to identify environmental and cognitive-affective cues that serve as 'triggers' for high-risk situations, including mood states, substance use, settings, and sexual partner characteristics. Strategies for avoiding or managing high-risk triggers and situations are discussed with individuals generating their own risk avoidance behaviors. Coping strategies for high-risk situations may include performing safer sex acts, redirecting sexual activities toward safer sex alternatives, carrying condoms, and avoiding sex after substance use. Identification of barriers to risk reduction, such as substance use and access to condoms are placed in the context of problem solving.

A key to behavioral self-management skills is condom use during sexual intercourse. Interventions educate individuals about the efficacy and proper use of condoms, but the skills training component goes beyond education by including safer sex exercises. Instruction, modeling, practice, and guided feedback desensitizes individuals to condoms, capitalizes on changing negative attitudes toward condoms. Increased experience with condoms has been shown to decrease rates of condom failure, a potential benefit of condom skills training. It should be noted, however, that condom use is a behavioral skill that includes interaction with one's partner, and in the case of women in heterosexual relationships as well as men who are the receptive partner during anal intercourse, use of male condoms can require techniques for putting condoms on partners as well as discussing and persuading partners to use condoms.

3.1.3 Sexual communication skills. Sexual negotiation, sexual assertiveness, refusal of high-risk sex,

and nonverbal communication are the most common types of communication skills included in HIV risk reduction interventions. Increasing skills for resisting partner coercion to engage in sexual intercourse without condoms, and increasing comfort discussing safer sex with partners in advance of sexual activity occur through instruction, modeling, and practice in communicating sexual decisions and discussing individual limits on behavior. Role playing situations allows for experience in what it may be like to talk with partners about risk and risk reduction. Modeling effective responses, practice, and guided feedback again form the core for the skills training model. Individuals learn techniques to communicate feelings and sexual behavior limits through practice sessions conducted to increase proficiency in communication skills and comfort in discussing sexual alternatives with partners.

4. Levels of Sexual Risk Reduction Interventions

HIV risk reduction intervention elements that include risk education, risk sensitization, and behavioral and communication skills building are found in all effective HIV risk reduction interventions delivered to individuals, groups, and communities. The most common HIV prevention activities occur in face-to-face counseling sessions, usually in conjunction with HIV antibody testing. Risk reduction counseling can be effective when it includes information, motivation, and skills building components, even when delivered in as short a time as 20 minutes (Kamb et al. 1998). However, there is now compelling evidence that HIV testing delivered with information-based counseling that does not include behavioral skills building components is ineffective at reducing HIV risks (Weinhardt et al. 1999).

Workshops delivered to small groups of individuals, typically 6 to 12 persons, that include information, motivation, and skills building activities have been shown effective in rigorously controlled studies with men who have sex with men, heterosexual men, women, and adolescents (Kalichman 1998). Finally, HIV prevention efforts delivered at the community level have also been shown effective when they include information, motivation, and behavioral skills. For example, peers and natural opinion leaders trained in skills for disseminating information and motivational messages can successfully change patterns of risk behaviors for entire populations (Kelly et al. 1997). Thus, sexual risks for HIV infection can be effectively reduced when they provide information, risk sensitization, and behavioral skills building opportunities in highly interactive formats.

See also: AIDS (Acquired Immune-deficiency Syndrome); AIDS, Geography of; Drug Use and Abuse:

Cultural Concerns; Drug Use and Abuse: Psychosocial Aspects; Health in Developing Countries: Cultural Concerns; Mortality and the HIV/AIDS Epidemic; Public Health; Sexually Transmitted Diseases: Psychosocial Aspects

Bibliography

- Fisher J D, Fisher W A 1992 Changing AIDS-risk behavior. *Psychological Bulletin* **111**: 455–74
- Friedman S R, Curtis R, Neaigus A, Jose B, Des Jarlais D 1999 *Social Networks, Drug Injector's Lives, and HIV/AIDS*. Kluwer Academic/Plenum, New York
- Kalichman S C 1998 *Preventing AIDS: A Sourcebook for Behavioral Interventions*. Erlbaum, Hillsdale, NJ
- Kamb M, Fishbein M, Douglas J, Rhodes F, Rogers J, Bolan G, Zenilman J 1998 Efficacy of risk-reduction counseling to prevent Human Immunodeficiency Virus and sexually transmitted diseases. *Journal of the American Medical Association* **280**: 1161–7
- Kelly J A, Murphy D, Sikkema K, McAuliffe T, Roffman R, Solomon L, Winett R, Kalichman S C 1997 Outcomes of a randomized controlled community-level HIV prevention intervention: Effects on behavior among at-risk gay men in small US cities. *Lancet* **350**: 1500–5
- Mann J, Tarantola D J 1996 *AIDS in the World II*. Oxford University Press, New York
- National Academy of Sciences 1995 *Preventing HIV Transmission: The Role of Sterile Needles and Bleach*. National Academy Press, Washington, DC
- NIH Panel 1997 *National Institutes of Health Consensus Development Statement on Interventions to Prevent HIV Risk Behaviors*. NIH Office of Medical Applications Research, Bethesda, MD
- Weinhardt L, Carey M P, Johnson B, Bickman N 1999 Effects of HIV counseling and testing on sexual risk behavior: A meta-analytic review of the published research, 1985–97. *American Journal of Public Health* **89**: 1397–1405

S. C. Kalichman

Hobbes, Thomas (1588–1679)

Thomas Hobbes was born on Good Friday, April 5, 1588, in Malmesbury in Wiltshire, UK. It was not until some months later that the Spanish Armada appeared in the Channel, but the threat that it would do so was widely known and there was widespread apprehension. Hobbes was fond of referring to this conjunction of his birth with the Armada:

For Fame had rumour'd, that a Fleet at Sea,
Wou'd cause our Nations Catastrophe;
And hereupon it was my Mother Dear
Did bring forth Twins at once, both Me, and Fear.
(*The Life of Mr. Thomas Hobbes of Malmesbury, written by himself.*)

This statement of Hobbes' is often referred to because of the role that fear played in his account of human nature and the genesis of political order.

Hobbes was born of a poor family, but obscure birth was no bar to the career of a clever young man at that time. His education made him a fluent linguist and translator, and though he wrote a considerable amount in Latin, which was still the common language of the learned, he became the first philosopher to write his most significant works in English. After study at Magdalen Hall in Oxford, he was recommended by the Principal of his hall to take the role of tutor and secretary in the household of William Lord Cavendish, who ten years later became, Earl of Devonshire. Hobbes spent almost all of the rest of his life in the employment of the Earls of Devonshire or their close relatives the Earls of Newcastle.

As a member of such households, Hobbes several times took sons of his employers on tours of Europe, which gave him a probably unparalleled chance to meet intellectuals: he met Galileo, whose work greatly influenced him as a response to the widely accepted scepticism of the time, Gassendi, Mersenne, and later Descartes, amongst others. These acquaintanceships continued through correspondence and when he fled to France in 1640 in reasonable fear that his political writings might result in threats to his life or liberty. In England, too, he met important people and intellectual leaders; Aubrey records that at one time he acted as secretary to Francis Bacon and also helped to translate some of Bacon's essays into Latin, with Bacon commenting that he liked Hobbes best of all those who had taken dictation from him because Hobbes understood what it was that he was writing.

Modern science was in a formative stage at that time, with no generally accepted scientific method. As an intellectually lively man, Hobbes in middle life became caught up in the arguments about this set of problems about the investigative methods appropriate to science, and Aubrey records him as saying that it was first reading Euclid, whom he took to produce definitive proofs, which made him alive to the problem and possibility of establishing a scientific method. He was also quite explicitly influenced by Galileo's *Dialogue Concerning the Two Chief Systems of the World*, published in 1632. Part of his concern with method came out in his dispute with Boyle (an account of which is given in Shapin and Schaffer 1985) about the significance of experiment, which Hobbes downplayed. Hobbes distinguished quite sharply between knowledge and opinion, knowledge requiring the sort of certainty he found in geometry. He held to the view that people were not capable of genuine knowledge of the external world or its characteristics, but that we can know, with the certainty required for knowledge, the contents of our minds. Genuine knowledge will therefore be of ideas and the relations between them. Observation can give us opinion, but cannot provide genuine knowledge. Nevertheless, Aubrey says, 'He

was wont to say that he had rather have the advice, or take Physique from an experienced old Woman, that had been at many sick people's Bed-sides, then from the learnedst but inexperienced Physitian,' so he did not ignore the importance of experience. His interest in optics and ballistics, which developed his interest in mathematics, arose from the Newcastle family's interest in the production of an effective telescope for military purposes, and he generally took science to be important because of its practical consequences.

This practical approach probably explains at least part of the dispute about the historians' Hobbes and the philosophers' Hobbes. During the 1960s and 1970s, Quentin Skinner produced a set of articles (1964, 1965, 1966, 1972) about the interpretation of Hobbes's writings on politics, the writings for which Hobbes is still remembered. Skinner's thesis was that more detailed consideration needed to be given to Hobbes's relationship to other writers of his time and the responses that he and they showed to the political events they lived through. Philosophers, it was implied, had taken Hobbes ahistorically and hence misinterpreted him. The most detailed reply (Warender 1979) does not deny that Hobbes was concerned to contribute to practical political debate of his time, but does deny that the differences between philosophers in their interpretations of Hobbes simply reflect a failure to put Hobbes in his historical context: they are, rather, differences about the account to be given of the philosophical backing for Hobbes's position. Skinner is clearly right in his understanding of why Hobbes took up the issues he took up, why he thought that certain problems needed to be dealt with, what allusions he might be making, and what he was trying to contribute to the current political debate, etc., that require the historical context. However, it is also true that Hobbes was not simply contributing to the practical debate but thought that he had a special contribution to make because he was producing a science of politics, the science involving timeless truths. He thought that his views were just as relevant to French disturbances of the time as to those in the UK. There is philosophers' Hobbes to study, and omission of the philosophical aspect of his work leads to misinterpretation.

The events at issue in the dispute he saw himself as entering, though Hobbes did not see himself as only, or always as primarily, concerned with politics, were political events. Hobbes was born in the reign of Elizabeth I and died on December 3, 1679, in the reign of Charles II, during which time he had been a tutor in mathematics to the Prince of Wales (accepting the position in 1646), had been awarded a pension by Charles II, was probably invited to stand as a candidate for the Short Parliament, and had lived through the English civil wars of the 1640s and the Commonwealth. He had had his books burnt in the University of Oxford and had been forbidden to publish his ideas in England, had (with some reason),

felt threatened with accusations of heresy, and had seen much turmoil in Europe. The issues that concerned him in politics were basically issues of stability. He was concerned with what was required for a state to remain stable and peaceful so that people could live their lives with the security they needed in order to be able to pursue their projects.

Hobbes's initial intellectual interests were in such fields as optics, ballistics, and mathematics, especially geometry. He is remembered by many for his attempts to square the circle, and he took part in vituperative disputes with the Savilian Professors of Mathematics at Oxford and especially with Wallis. His enduring contribution to our intellectual heritage, though, is in the fields of morality and politics, which he tried to establish in a scientific manner. His main publications in these fields were *The Elements of Law; De Cive* (originally in Latin, but there soon appeared an English translation about which there is some disagreement concerning the extent to which Hobbes approved of it); *Leviathan* (originally in English but translated with changes into Latin so that it could be read elsewhere); and *De Homine* (which still lacks a complete English translation, though there is a partial translation in Gert 1972). The translations show Hobbes's keenness to make his ideas available as a contribution to the current political debates in a number of countries, and this is also reflected in the changes in his styles of presenting his ideas (see, for example, Johnston 1986). His *Behemoth*, an account of the causes of the English civil wars of the 1640s, also reflects his ideas on politics. He discusses methods of proof and the nature of science and true knowledge frequently, the most extended discussion being in *De Corpore*, an English translation of which appears in volume 1 of the Molesworth edition of *Hobbes: English Works*.

Hobbes's best-known work is *Leviathan*, the last of his English writings directly on the science of politics. It appeared in 1651, at the time of Hobbes's return to England. There are two different stories about how it fits into the historical context and what it was intended to achieve. The Royalist Earl of Clarendon (in *A Survey of Mr Hobbes his Leviathan*) reports Hobbes as saying

... that he knew when I read his Book I would not like it, and thereupon mentioned some of his Conclusions; upon which I asked him, why he would publish such doctrine: to which, after a discourse between jest and earnest upon the Subject, he said *The truth is, I have a mind to go home.*

Many Royalists did take the book that way, as a defence of Cromwell and his Commonwealth. Aubrey, on the other hand, says

He wrote and published the *Leviathan* far from the intention either of disadvantage to his Majestie, or to flatter Oliver (who was not made Protector till three or four yeares after) on purpose to facilitate his returne; for there is scarce a page in it that he does not upbraid him

and goes on to suggest that the purpose was to provide justification for those who had fought on the Royalist side until all was lost and now faced the prospect of losing all their property if they did not recognise Cromwell's authority. The most that can be done to sort out this disagreement is to see what the arguments of *Leviathan* will support.

Hobbes's moral and political theory is presented as an account of how and why people emerge from their natural condition and live together in civil society. This involves giving an account of human nature, of the natural condition of mankind, and of the laws of nature that make the transition possible. There is still argument about just what Hobbes's position is on each of these matters. It is clear that Hobbes takes the natural condition of mankind (he rarely uses the expression 'state of nature') to be a state of insecurity and constant fear. Though not always a condition of actual fighting, it is a condition in which people will take an aggressive stance towards each other and in which nobody can be trusted and fighting is always likely to break out. The reasons for this are to do with human nature, and the solution is the making of a contract to establish a sovereign.

Hobbes has more to say about the actual origins of states in *A Dialogue Between a Philosopher and a Student of the Common Laws of England*, but what he is intending to do with his story of the contract is to set up a model for the establishment of a state. This, on his account, is what is involved in explaining something or giving its causes. Establishing the cause of something, on Hobbes's account, is not the searching after particular matters of fact in the situation that we might go into in investigating a traffic accident. He gives an example to explain his method in chapter 1 of *De Corpore*. A figure placed before us might look like a circle, but mere observation cannot assure us that it is one. Knowledge of its cause assures us of that.

For let it be known that the figure was made by the circumduction of a body whereof one end remained unmoved, and we may reason thus; a body carried about, retaining always the same length, applies itself first to one *radius*, then to another, to a third, a fourth, and successively to all; and, therefore, the same length, from the same point, toucheth the circumference in every part thereof, which is as much as to say, as all the *radii* are equal. We know, therefore, that from such generation proceeds a figure, from whose one middle point all the extreme points are reached unto by equal *radii*. And in like manner, by knowing first what figure is set before us, we may come by ratiocination to some generation of the same, *though not perhaps that by which it was made, yet that by which it might have been made*; [italics added] for he that knows that a circle has the property above declared, will easily know whether a body carried about, as is said, will generate a circle or no. (Hobbes 1839, Vol. 1, p. 6)

The last sentence here makes clear that Hobbes does not seek causes as we usually do at the beginning of the twenty-first century, but seeks a model that will explain

what it is to be a circle, or a civil society. He is not committed to the idea that states actually do begin from contract, and he certainly recognizes that many begin with, or are expanded by, conquest. A civil society, then, will be something that has the features of an institution established by contract. It will involve obligations on people, and those obligations will depend on reciprocity. One of the distinguishing features of Hobbes's account is that the reciprocity would not primarily be between sovereign and citizen, but between citizens (as it is in such present-day theories as Rawls 1972); the sovereign, importantly, is outside the contract. Were he not outside it, then somebody else would have to hold the authority to determine whether the sovereign was behaving properly, and that person would then have the ultimate authority and be the actual sovereign. This is the absolutism of the Hobbesian sovereign to which so many have objected. Hobbes, both in not depending on an actual contract and in making the mutual obligations between citizens rather than between citizen and sovereign, was moving beyond the other contract theorists of his time.

Why would the natural condition, in which people have no common sovereign, be the unpleasant state that Hobbes describes? He gives several significantly different examples in discussing the natural condition, American Indians and the relationships between sovereign states amongst them, which can be confusing. The core of the notion is that people are in their natural condition if they have no common sovereign, so that somebody might be in his natural condition with respect to one person but not with another. It is not that the whole world must be in either the natural condition or civil society or that an individual person must be either in the natural condition or not. Within a family we might not be in our natural condition with respect to each other, but each of us might be in the natural condition with respect to people outside our own family.

Why should people who lack a common sovereign be unpleasant to each other? One thesis is that it is because they are self-interested and will therefore naturally clash in the absence of a coercive power to prevent them. This is the interpretation that has led to games theorists taking up Hobbes (e.g., Gauthier 1969, Hampton 1986, Kavka 1986) and to his being taken as providing a 'realist' model for international relations (e.g., Bull 1980) with reasons of state and pursuit of national self-interest at the center of concern, that is, of international politics as power politics. This is, perhaps, a plausible picture of international politics. Along with this idea goes the interpretation of the laws of nature as policies to be followed out of self-interest, fitting with Hobbes's description of them as rules of prudence.

Bernard Gert (1972) argues strongly against the claim that Hobbes is a psychological egoist who believes that people always seek their own narrow

interest. Certainly it seems clear that Hobbes did not believe that people could do no other than try to preserve themselves when their lives were threatened: not only did he comment on those who give their lives too cheaply in duels because of a confused notion of honor, but he admired some of those who gave their lives in defence of the Royalist cause in the civil wars and made laudatory remarks about one of them in dedicating *Leviathan* to his brother. This position, then, requires a different account of the laws of nature.

Hobbes was clear that the laws of nature are effective only in conditions of reciprocity and preferably conditions in which there is a sovereign to guarantee reciprocity. This at least looks as though it supports the idea that the only reason for obeying them is self-interest and particularly fear of the sovereign, but drawing that conclusion means ignoring the fact that Hobbes treats prior performance by one party to a contract as making an obligation effective on the other party just as much as does having a sovereign who will force one to comply. The laws of nature, he says, obligate *in foro interno*, in conscience, in the natural condition, that is, they bind in the intention, but they bind *in foro externo*, in the act, only in civil society. Virtues are more complex than they are sometimes treated as being: Hobbes's point is that benevolence, for example, is a good thing because of the role it actually plays in civil society. Handing a knife to somebody who wants to kill me is not benevolent; it is simply stupid. Handing over money to a grocer who refuses to provide me with the goods I want is not required by justice. The virtues actually require a setting of relatively peaceful and secure life if they are to be possible to exercise, and Hobbes's stronger thesis that they can be exercised only by people who are subject to a common sovereign obviously bodes ill for a morality of international relations and explains why Hobbes has been taken up by some proponents of world government. (On this interpretation of Hobbes, see, for example, Bull 1980 and Ewin 1991.) On this account, the story Hobbes tells of people agreeing on principles of peace is a model of the sort he describes in the account of the circle, and the model is for virtues. Certainly, for all his objections to Aristotle, his objections to what Aristotle said about virtues is that he identified them as means, and not in terms of their being, as Hobbes put it, 'not properly Lawes, but qualities that dispose men to peace, and to obedience' (Chap. 26 of *Leviathan*). That is, he suggests agreement with Aristotle's general position about morality, but disagrees with the principles by which virtues are to be identified. And this fits, not with games theory, but with the accounts of ethics that now seem to be emerging from evolutionary theory about the characteristics that make our social species capable of surviving (see, e.g., Alexander 1987).

If people have these virtues as part of human nature, even if they do not have them perfectly, why would the

natural condition of mankind be as horrible as Hobbes suggests? Because, in the absence of a sovereign, the virtues cannot be operable; the laws of nature will not bind *in foro externo*, that is, to any overt action. We are all equal in the natural condition not only in that we are all vulnerable to sudden death, but in that there is no particular reason why any of us should subject his or her reason to that of another. If we differ and neither of us can persuade the other, then there is no reason why either should submit. And in this world we do differ: we see the world in different ways from different backgrounds, are struck by different features of the situation, and we calculate differently. We have differences about the nature of justice, and we differ about who did what to whom. When we differ about something like justice, there is no reason why one should submit to the view of another and do what he simply regards as unjust. It is differences of judgment, rather than clashing interests, that the sovereign must deal with, providing an artificial reason to replace the right reason that fails to do the job. Those who insist on right reason in a dispute, Hobbes says, are simply insisting on being judges in their own case, for each man takes for right reason what is his own.

His first argument for the sovereign in *Leviathan* appears early in Chap. 5 in the discussion of disagreement and the need to establish an arbitrator or judge to determine disputes (though the arbitrator or judge might be a meeting of the populace voting on the matter or any number of other things, and is best thought of as a procedure rather than as a person), an arbitrator whose judgment we agree to accept before we know what it is, thus establishing rights. Without those rights the virtues will not be operable, and without the virtues, at least to a degree, the rights could not be established. The recognition of authority, on Hobbes's account, is something without which humanity could not survive. Our moral views have the status of private views in which we can clash, but people have a capacity for politics that makes it possible for them to establish a public morality (or law) of rights that determines the sphere within which each person's private judgment can operate.

This shows some of Hobbes's relationship to modern liberalism. And because people do have failings, and because we need security even though we know that there are dangerous people whom we could not recognize at first sight, the sovereign must have coercive powers including the powers to tax people to cover the costs of his activities; there will be clashes of interest as well as clashes of judgment. The sovereign must be absolute in that his function is to resolve disputes, and he cannot do that if his authority depends on his agreeing with both parties. But that, Hobbes suggests, is a logical truth about any system of government and is compatible with our choosing government even by a meeting of all the citizens; it does not require subjection to the arbitrary will of one person, but does require subjection to the decisions of

the community however they are determined. This brings out some of Hobbes's relationship to modern democratic theory. The sovereign makes it possible to resolve disputes peacefully and be sure that the resolution will be enforced, so he makes it possible for us to act as one; he determines for us the will of the community, and thus makes us one person. This Hobbesian idea of sovereignty as ultimate and absolute remains influential.

See also: Altruism and Self-interest; Democratic Theory; Liberalism: Historical Aspects; Sovereignty: Political

Bibliography

- Alexander R 1987 *The Biology of Moral Systems*. Aldine de Gruyter, Hawthorne, NY
- Aubrey J 1898 *Brief Lives*. Clarendon Press, Oxford, UK
- Boonin-Vail D 1994 *Thomas Hobbes and the Science of Virtue*. Cambridge University Press, Cambridge, UK
- Bull H 1980 *The Anarchical Society*. Macmillan, London
- Ewin R 1991 *Virtues and Rights*. Westview, Boulder, CO
- Gauthier D 1969 *The Logic of Leviathan*. Clarendon Press, Oxford, UK
- Gert B (ed. and intro.) 1972 *Man and Citizen*. Harvester Press, Brighton, UK
- Hampton J 1986 *Hobbes and the Social Contract Tradition*. Cambridge University Press, Cambridge, UK
- Hobbes T 1839 *English Works* (ed. Molesworth W). John Bohn, London
- Hobbes T 1961 *Opera Latina*. Scientia, Aalen, Germany
- Hobbes T 1968 *Leviathan* (ed. Macpherson C B). Penguin Books, Harmondsworth, UK
- Hobbes T 1969 *The Elements of Law* (ed. Tonnies F.). Cass, London
- Hobbes T 1998 *On the Citizen* (eds. Tuck R., Silverthorne M.). Cambridge University Press, Cambridge, UK
- Johnston D 1986 *The Rhetoric of Leviathan*. Princeton University Press, Princeton, NJ
- Kavka G 1986 *Hobbesian Moral and Political Theory*. Princeton University Press, Princeton, NJ
- Rawls J 1972 *A Theory of Justice*. Clarendon Press, Oxford, UK
- Rogow A 1986 *Thomas Hobbes: Radical in the Service of Reaction*. W. W. Norton, New York
- Shapin S, Schaffer S 1985 *Leviathan and the Air Pump*. Princeton University Press, Princeton, NJ
- Skinner Q 1964 Hobbes's *Leviathan*. *The Historical Journal* 7: 321–3
- Skinner Q 1965 History and ideology in the English Revolution. *The Historical Journal* 8: 151–78
- Skinner Q 1966 The ideological context of Hobbes's political thought. *The Historical Journal* 9: 286–317
- Skinner Q 1972 Conquest and consent. In: Aylmer G E (ed.) *The Interregnum*, Macmillan, London
- Tuck R 1989 *Hobbes*. Oxford University Press, Oxford, UK
- Warrender 1979 Political theory and historiography. *The Historical Journal*, pp. 931–40

R. Ewin

Holocaust, The

'The Holocaust' is the term most widely used in reference to the systematic extermination of the Jews of Europe by the Nazis during World War II. Usually, the term also alludes to the historical background of the extermination and to related contemporary circumstances. The first section of this entry will deal with some aspects of the historical background, the second will briefly present the immediate context of the 'Final Solution to the Jewish Question,' while the third and central part will address the extermination *per se* and the reactions to it. This chronological presentation does not imply an unavoidable passage from one phase to the next. However, while the link between the evolution of German history during the first two decades of the twentieth century and the onset of the Nazi regime was in no way evident, the causal relation between the anti-Jewish policies of the early war period and the extermination was much closer.

1. The Background

The historical origins of the Holocaust remain an object of ongoing debate. Whereas some scholars consider the extermination of the Jews as the consequence of a specific evolution of German history, at least from the end of the nineteenth century onwards, others interpret the Holocaust as an outgrowth of modernity, the implementation of a pitiless instrumental rationality at the service of a totalitarian racial utopia. Both approaches are not incompatible.

The rise of racial anti-Semitism, the spreading of social Darwinism, and the publication of the first theories of racial anthropology and eugenics turned the late nineteenth century into a breeding ground of ideological trends that were to become central tenets of the Nazi worldview. And, although political anti-Semitism declined in the Wilhelminian Reich at the beginning of the 1900s, an increasingly vocal anti-Jewish hostility spread throughout wide sectors of German society. In this evolution, the outbreak of World War I became a major turning point.

After a brief period of national unity, the unforeseen prolongation of the conflict and the growing political polarization which it induced triggered a radicalization of anti-Jewish attitudes, mainly in the nationalist-conservative camp. The defeat of Germany, the revolutionary upheavals that followed, and the birth of a republic hated and rejected by an important sector of the population, as well as the vast majority of the traditional elites, fanned the already virulent anti-Semitism even further. The Jews were accused of bearing the responsibility for the national humiliation (the 'stab in the back' myth), of controlling the new

republic, of instigating bolshevism and, above all, of conspiring to dominate the world, as described in the most widely read anti-Semitic forgery of the century, *The Protocols of the Elders of Zion* (1919).

It was in this postwar atmosphere that Adolf Hitler began his political career by joining the tiny German Workers' Party (DAP) in Munich, in late 1919. In 1920, the DAP became the National Socialist German Workers' Party (NSDAP) with anti-Semitism figuring prominently in its twenty-five point program. Notwithstanding the failed putsch of 1923 and Hitler's brief incarceration, the resumption of his political agitation in 1925 along with systematic organization, unrelenting propaganda, and, more than anything else, the impact of the Great Depression, propelled the party into prominence on the national scene.

Anti-Semitism lay at the center of Hitler's worldview from the outset of his political career and remained so after his accession to power on January 30, 1933. Reduced to its core, this anti-Semitism viewed historical evolution as determined by a merciless racial struggle dominated by the confrontation between Aryan and Jew. The outcome of this struggle was to decide the fate of mankind and particularly of the Aryan race. Hitler's faith had an apocalyptic dimension, presented in terms of 'scientific' theories. But how far this worldview concretely led to the systematic planning of the anti-Semitic policies of the regime, or, obversely, to what extent these measures were the result of uncoordinated initiatives stemming from a growing radicalization of the party and its agencies, remains an open question. One could argue that the ideological guidelines indicated a general framework for action, but that the 'twisted road' followed by the implementation of the anti-Jewish policies was dictated by changing circumstances. In any case, during the 1930s, no plan for a physical extermination of the Jews existed; the ever-more extreme anti-Jewish policies of the prewar years had two successive and identifiable aims: the exclusion of the Jews from the national racial community (the *Volksgemeinschaft*) and, thereafter, their expulsion from the Reich.

From the outset of the new regime the Jews were dismissed from the civil service, from a growing number of professions, and from all major fields of German cultural life. Moreover, all that belonged to or had been influenced by the 'Jewish spirit' had to be cleansed from the German cultural sphere: Freud's or Einstein's theories, Heine's poetry, Mendelssohn's music, books, works of art, memorial plaques, films, libretti—anything. A stream of anti-Jewish incitement was daily spewed out by all the propaganda channels of party and state, under the guidance of Propaganda Minister Joseph Goebbels and chief party ideologist, Alfred Rosenberg; it found its most rabid expression in *Der Stürmer*, the 'specialized' anti-Jewish organ of the Gauleiter of Franconia, Julius Streicher. The entire education system soon absorbed the anti-Semitic

message, as did the Hitler Youth (HJ) and the League of German Girls (BdM). The indoctrinated youngsters of the 1930s were to become the soldiers of the 1940s.

In September 1935, the Nuremberg Laws added strict biological separation to social and cultural segregation by forbidding marriage and sexual relations between Jews and Germans (the Law for the Protection of German Blood and Honor). The Reich Citizenship Law reduced the Jews to second-class citizens. Finally, the first 'Supplementary Decree' to the Reich Citizenship Law defined who was a Jew or, alternatively, a 'mixed breed' of the first or second degree (November 1935). Any transgression of the 'Blood Law' was considered as 'race defilement' (*Rassenschande*), and the offenders (usually the Jewish male offenders) were sent to jail or a concentration camp.

In 1936, the *Reichsführer SS*, Heinrich Himmler, was put in command of all German police forces. This reorganization of the control and terror apparatus led to the appointment of SS General Reinhard Heydrich, previously chief of the security service of the SS (*Sicherheitsdienst* or SD) to the command of both the SD and the newly established Security Police (*Sicherheitspolizei*), which included the Gestapo and the Criminal Police. After the outbreak of the war, these agencies were further centralized within the Reich Security Main Office (*Reichssicherheitshauptamt* or RSHA). While the planning of direct anti-Jewish terror was mainly Himmler's and Heydrich's preserve, the plundering of Jewish assets and enterprises was orchestrated by Hermann Goering's Four Year Plan Administration, and, after the dismissal of the conservative minister of the economy, Hjalmar Schacht (who had managed to protect some measure of Jewish economic life), by his Nazi successor, Walther Funk.

After the annexation of Austria, in March 1938, the drive to compel the Jews to leave the 'Greater German Reich' ('Old Reich' and Austria) became overwhelming. During that year, 'compulsory Aryanization' and, ultimately, the prohibition of any Jewish economic activity destroyed the material basis of Jewish existence in Germany. Notwithstanding the widespread reticence of most countries to take in Jewish refugees (easily identifiable since, at the demand of the Swiss government, their passports were marked with a 'J'), forced emigration and even expulsion over the borders became standard Nazi practice. Finally, on November 9 and 10, following the killing of a German diplomat by a Jewish youth in Paris, a nationwide eruption of state-organized violence, the so-called 'Night of the Broken Glass,' destroyed the remnants of Jewish life in Germany. The Jews of the Reich now desperately attempted to flee. The panic soon extended to the Jews of the 'Protectorate of Bohemia and Moravia,' after the German occupation of Prague in March 1939. Yet, it was the German attack on Poland, on September 1, 1939 and the outbreak of World War II that entirely changed the fate of European Jewry.

2. *The Immediate Context of the 'Final Solution' (1939–41)*

During the first two years of the war, the goal of Nazi anti-Jewish policies still appeared uncertain. Jewish emigration from the Reich and occupied Europe remained a possibility, although wartime conditions soon reduced it to a trickle. The goal frequently alluded to was a territorial concentration of Europe's Jews, first in the easternmost part of Poland, in the Lublin area of the newly established 'General Government' (the name used for the nonannexed area of Poland which was put under the authority of the Nazi 'Governor General,' Hans Frank), then, after the victory over France, in the French island colony of Madagascar, and, finally, after the attack on the USSR, in Arctic Russia or Siberia. None of these plans ever materialized, except for some short-lived attempts to deport Jews from Vienna and the Protectorate to the Lublin area, in the fall of 1939 and in early 1941.

A few months after the occupation of Poland, the Nazis started to deport hundreds of thousands of Jews and Poles from territories annexed to the Reich, such as the Warthegau (a part of western Poland), into the General Government; they drove the Jews into ghettos in larger towns and even dumped Jews from the Reich into Vichy France, in the fall of 1940. The deportations within ex-Poland may have been linked in part to the regime's policy of ingathering ethnic Germans from the Baltic countries, the Russian occupied territory of Poland, and the Balkans in the 'Great Germanic Reich.' Thus, the deportations of Jews and of Poles from the Warthegau were meant to make space for the incoming *Volksdeutsche*. The policy was pursued under the orders of Himmler, as newly appointed 'Reich Commissar for the Strengthening of Germanism' (RKF). It was meant to be the first stage of what became known as 'General Plan East' (*General Plan Ost*), aimed at establishing German domination in the East for centuries to come. The enslavement of entire nations and the death of tens of millions of local inhabitants was taken into account. This policy was probably not directly linked to the genesis of the 'Final Solution' since it cannot explain the deportations from the west (to be described later on).

Simultaneously, from October 1939, an entirely separate but not unrelated operation had started in the Reich: the systematic mass murder of the mentally ill, the so-called euthanasia or 'Aktion T-4': personnel were trained and killing methods were developed that were to be used for the mass extermination of Jews after the scaling down of T-4, in the summer of 1941, to a less visible 'euthanasia' operation (Aktion 14f13).

In Western Europe, the much smaller Jewish populations were not secluded (except for the establishment of a semi-ghetto in Amsterdam) or as brutally treated as the Jews of Poland. But their exclusion from public service and the liberal professions, growing curtailment of their economic acti-

vity, the registration of persons and property, limitations on travel, and police surveillance, turned them into pariahs. Their persecutors were not only the Germans, but also satellite regimes such as Vichy France.

After the German occupation of Yugoslavia and Greece in April 1941, Serbia remained under direct military control and Croatia became a satellite; in both countries, harsh anti-Jewish measures were introduced. Slovakia's anti-Jewish policy followed strict German orders. Romania implemented fierce anti-Semitic measures of its own, and only Hungary and Bulgaria maintained semi-independent policies in this domain.

In order to facilitate their control over the Jews—beyond territorial concentration—as well as to degrade them, the Nazis compelled all those over the age of six to display a 'Jewish star' on their clothing, first in Poland, then, from September 1941 on, in the Reich and the Protectorate, and, in the spring of 1942, in all countries of occupied Western Europe. But the principal instrument of Nazi domination was the 'Jewish Councils.' These councils, composed of representatives of the Jewish communities, received German orders, conveyed them to the Jewish populations, and were responsible, upon pain of death, for their implementation. Moreover, in Eastern Europe the Jewish Councils became agencies of a Jewish self-administration of sorts, struggling against impossible odds; justly or unjustly, they often became targets of sarcasm, resentment, and even hatred in their own communities. Some historians have adopted a highly critical stance toward the councils, considering them as instruments in the hands of the Nazis for the destruction of their own people. Others have questioned this overall characterization as the differences in the behavior of the councils were considerable. All in all, the councils' members were ordinary people confronted with entirely unprecedented circumstances: their own reasoning was incompatible with the murderous strategy of the Nazis.

3. *The Extermination*

Systematic mass extermination of Jewish populations began with the German attack on the USSR, on June 22, 1941. Whereas in most newly occupied Soviet territory mainly Jewish men were executed until August 1941, in the Baltic countries and in Transnistria (where the Romanian army was the principal ally of Germany), entire Jewish communities were butchered from the outset. The killing of Jews in party or state positions had been ordered by Heydrich on the eve of the campaign; the perpetration of pogroms by local inhabitants was also explicitly fostered by the Germans. Whether the killing that engulfed the entire

Jewish population from August 1941 onwards was independent of any instrumental goals, or whether, as with the starving to death of two million Russian POWs before the spring of 1942 (around 3.3 million died before the end of the war), it derived from difficulties experienced by the Wehrmacht in securing sufficient food supplies for the army is not clear. In any case, the mass murder of Jews on Soviet territory (around 500,000 victims by the end of 1941) was not perpetrated solely by the 'intervention units' of the Security Police and the SD (*Einsatzgruppen*), by local auxiliaries, and by the Romanian army in the south, but also by tens of Reserve Police Battalions and regular Wehrmacht units. In other words, not only the 'troops of the ideological war' (the SS) but ordinary men, possibly ten of thousands of them, were involved, often willingly so, in the mass murder of the Jews.

Few historians would now argue that the decision to exterminate all the Jews of Europe predated the attack on the USSR. Some contend that this decision was taken by Hitler soon after the German onslaught, during the period of 'euphoria,' due to the apparent success of the invasion, and that it found its expression in Goering's letter to Heydrich of July 31, 1941, ordering him 'to take all necessary measures to prepare the Final Solution of the Jewish Question in Europe.' Others argue that such an order was issued in the early fall, when doubts may have risen in Hitler's mind about the prospects of a quick victory in the East. Others still link the order to the USA's entry into the war, in other words to the transformation of the European war into the 'world war,' which, as Hitler had 'prophesied' in his Reichstag speech of January 30, 1939, would see 'the extermination of the Jewish race in Europe.' This last interpretation seems bolstered by documentary evidence. On December 18, according to Himmler's appointments diary, the SS chief met with Hitler and wrote down: 'Jewish question/eradicate as partisans.' The reference to 'partisans' is clearly a way of describing an enemy lurking behind the lines, an internal foe, who, just as in 1917/18, would stab Germany in the back if given the chance. Finally, some historians doubt the existence of any clear order given at the highest level and consider total extermination as resulting from the cumulative effects of local initiatives.

By October 1941, the killings had extended to Jews from the Greater Reich and the Protectorate who were deported mainly to Lodz, Riga, Kovno, and Minsk. In Riga and Kovno, thousands were shot on arrival. The others were exterminated later with the inhabitants of their ghettos. They had also spread to Serbia, where the Wehrmacht executed some 6,000 Jews in reprisal for partisan attacks. During the same month, Jews were forbidden to leave the continent, 'in view of the coming final solution of the Jewish Question' and the construction of the Belzec extermination camp (possibly that of Sobibor too) began in the General Government.

In early December, the first extermination site became operational at Chelmno, near Lodz, principally for the murder of Jews from the Warthegau. On January 20, 1942, in a meeting in Berlin (*am Grossen Wannsee*), Heydrich informed an array of state secretaries, senior SS officers and other selected officials of the general extermination program that, according to his computations, was to apply to 11 million European Jews. From then on total extermination became a problem of logistics and of some measure of haggling between the Germans and various governments or local authorities about handing over their Jews. In almost all cases, foreign Jews were arrested and delivered first, local Jews followed (as will be indicated further on, in some rare cases Jews were not handed over at all).

The arrests of Jews, mainly in Western Europe, and the logistics of the deportations were coordinated between the Higher SS and Police Leaders (HSSPF), Himmler's delegates to each occupied country or major area, local authorities and police forces, and the Jewish Affairs section at the RSHA (IVB4) under Adolf Eichmann. Furthermore, the deportations demanded joint planning with the Reich Railways Authority (*Reichsbahn*) and its territorial subdivisions (*Ostbahn*, etc.). Once the Jews reached the camps, they passed under the command of the SS Economic and Administrative Main Office (*SS Wirtschafts- und Verwaltungshauptamt*) under SS General Oswald Pohl (except for the *Aktion Reinhard* camps). In the camps, the Jews were destined for extermination; however, mainly in Auschwitz, selections separated the healthiest and the strongest from the mass of the inmates, submitting them to slave labor before sending them to their death.

The first stage of the extermination chiefly included the Jews of the Warthegau and of the General Government who, through 1942, were killed in the *Aktion Reinhard* camps (Belzec, Sobibor, Treblinka, and Majdanek), as well as in Chelmno. The deportation of the Jews of Warsaw to Treblinka began in July 1942. Simultaneously, deportations from Western Europe (France, Belgium, and Holland) started. By then, the principal Auschwitz extermination camp (Auschwitz II or Birkenau) had become fully operational. The killing method shifted increasingly from the use of carbon monoxide gas to that of Zyklon B, a powerful pesticide utilized in the fumigation of army barracks and ships' holds. Before they were killed, the victims' valuables and clothes were taken away and their hair was shorn; after the gasings, gold teeth, and artificial limbs were removed from the corpses. The German economy recycled the 'booty': gold was smelted and delivered to the *Reichsbank*; hair was used for mattresses, slippers, etc.; clothes were distributed by the *Winterhilfe* (the German winter-relief organization). Homes, furniture, and other belongings left behind by the deportees, were distributed to deserving *Volksgeossen* or acquired, usually at dirt-cheap

prices. This allocation of spoils spread to occupied countries in the West. Particularly valuable objects, mainly art, looted all over Europe by the *Einsatzstab Reichsleiter Rosenberg* or ERR, were offered to Nazi leaders (Hitler and Goering headed the list), to German museums, and various other German institutions.

In 1943, the last batch of German Jews was deported to the extermination camps; the deportations from Western Europe continued, and, simultaneously, the Jews of the Balkans and, by the end of the year, of Italy (following the German occupation of the greater part of the peninsula) were included in the killing process. From March 1944 onward, after the occupation of Hungary by German troops, this country's Jews became the last large national group of victims (over 400,000) to be sent to Auschwitz-Birkenau. On November 2, 1944, Himmler ordered the cessation of the mass exterminations; his order was frequently disregarded, even after the liberation of a growing number of camps by the Allies. 'Death marches,' organized by camp commanders or SS personnel, went on throughout the winter of 1944-5 and practically until the German surrender in May 1945. By the time the war in Europe drew to a close, approximately 5.5 million Jews had been exterminated.

The exterminations on Soviet territory had become widely known by the fall of 1941. Information about the mass killing of European Jewry beyond the Soviet border spread throughout the continent, and in Allied and neutral countries from the second half of 1942 onward, although the technical details of the 'Final Solution' may not have been widely known until 1944. The overall reaction was one of indifference and inaction, sometimes of outright collaboration. While Vichy authorities and police collaborated with the Germans in the arrests and deportations, Switzerland hermetically closed its borders to the Jews who attempted to flee. In general terms, the Christian churches intervened feebly, mainly on behalf of converted Jews. Thousands of Jewish children were hidden in Christian institutions; in most cases, their lot was conversion. The silence of Pope Pius XII, notwithstanding some help from the Vatican in hiding Jews in the fall of 1943, has, to this day, remained a matter of heated debate.

In some cases, the reactions were different. Spain and Portugal allowed the transit of Jews through their territory. Sweden attempted to take in the Jews of Norway when the deportations started and it took in almost all the Jews of Denmark (most of whom were transported to safety by the Danes themselves). Protests against the anti-Jewish measures were staged in Holland, in early 1941; the Italians tried to protect the Jews in their own country as well as those under their control. Bulgaria refused to deliver its own Jewish citizens, although it did deliver the foreign Jews. Finland rejected all German demands regarding its small Jewish population.

'The abandonment of the Jews' by the Western Allies is largely undisputed nowadays. Rejected or fatally delayed rescue plans involved transfers of ransom money to the Germans or their allies, or the bombing, in 1944, of the Auschwitz gas chambers and of the railway line connecting the camp with Hungary. Such refusal to help, whether known or merely sensed, could but contribute to the feeling of total isolation of the Jews of Europe in the face of the German extermination process.

The concrete possibilities of Jewish resistance were almost nil. Moreover, by the time most European resistance movements became active, the bulk of European Jewry had been exterminated. Hence Jewish armed revolt in the ghettos (the best-known instance being the Warsaw ghetto revolt of April 1943) or even in the camps, although hopeless, carried the utmost symbolic significance. As for the immense majority of European Jews, they had practically no options even when impending murder was obvious: flight and hiding were limited options; despair and numbness, overwhelming.

See also: Anti-Semitism; Collective Memory, Psychology of; Contemporary History; Dictatorship in History; Ethnic Cleansing, History of; Genocide: Historical Aspects; Historiography and Historical Thought; Jewish Tradition; Judaism and Gender; National Socialism and Fascism; Racism, History of; Second World War, The; Violence, History of; War Crimes Tribunals

Bibliography

- Aly G 1995 *Endlösung*. S. Fisher, Frankfurt am Main, Germany
 Bankier D 1992 *The Germans and the Final Solution*. Oxford University Press, Oxford, UK
 Bauer Y 1994 *Jews for Sale?* Yale University Press, New Haven, CT
 Browning C R 1992 *Ordinary Men*, 1st edn. HarperCollins, New York
 Burrin P 1994 *Hitler and the Jews*. Edward Arnold, New York
 Friedländer S 1997 *Nazi Germany and the Jews*, 1st edn. HarperCollins, New York
 Gerlach C 1997 *Kalkulierte Morde*. Hamburg, Germany
 Gutman Y 1982 *The Jews of Warsaw 1939–1943*. Indiana University Press, Indiana, IN
 Herbert U (ed.) 1998 *Nationalsozialistische Vernichtungspolitik*. Frankfurt Mainam, Germany
 Heer H, Naumann K (eds.) 1995 *Vernichtungskrieg*. Hamburg, Germany
 Hilberg R 1993 *Die Vernichtung der Europäischen Juden*, 3 Vols. Fischer, Frankfurt, Germany
 Longerich P 1998 *Politik der Vernichtung*. Piper, Munich, Germany

- Marrus M R, Paxton R O 1981 *Vichy France and the Jews*. Basic Books, New York
 Mommsen H 1991 The Realization of the unthinkable. In: Mommsen (ed.) *From Weimar to Auschwitz*. Princeton University Press, Princeton, NJ
 Trunk I 1972 *Judenrat*. Macmillan, New York
 Wyman D S 1984 *The Abandonment of the Jews*, 1st edn. Pantheon Books, New York

S. Friedländer

Home in Anthropology

1. Definition and Meaning of 'Home'

The term 'home,' being of Anglo-Saxon and Germanic origins, often carries rational and irrational connotations at the same time: house, landscape, habitat vs. place of belonging, object of nostalgia, place of origin. The German equivalents, *Heim* and *Heimat*, originally meant 'house' and 'owned land,' but later on denoted ideal and material spaces offering homeliness or German *Gemütlichkeit*. Not coincidentally, the term *Heimat* was re-introduced into the German written language through the reception of British gothic ideas and romanticism. Moreover, *Heimat* was adapted by many other languages as a lean word signifying a mixture of connotations all tending towards an emotionalized interpretation of the place a person feels adapted to: 'my home is my castle.' While loaded with emotional charge, the term 'home' usually also carries allusions to very profane and unromantic connotations, such as in the euphemistic names given to institutions catering for the poor, old, and disabled, or in the meanings expressed by the untranslatable German term *unheimlich* (the 'uncanny' of Freudian philosophy).

Anthropologically relevant references to 'home' initially appeared during the nineteenth century in Anglo-Saxon and German romantic and gothic styles of writing in which it denoted *Ur-Heimaten* or certain groups of humankind, such as the 'Aryan' origins in Asia or people's origins in other 'cradles' (Kuper 1988). Later on the term 'home' acquired a more specific political meaning as in the case of the 'homelands' assigned by the South African apartheid regime to ethnicized groups of 'black' citizens, which had the purpose of containing them in reservations or similar constructions of 'home rule.' More recently, the term has gained some significance again in anthropological studies of migrant, diaspora, and fugitive populations (Malkki 1995, Vertovec and Cohen 1999).

The related discussions move between two poles, one pole being represented by the 'invention of

tradition' approach (Hobsbawm and Ranger 1982), and the other pole tending towards sociobiological and ethological claims about the existence of a human instinct which orientates us towards distinct 'environs' or natural habitats (Wilson 1975). From an anthropological perspective, the often rather unreflected and untheorized use of the term 'home' in scientific texts offers a good opportunity for checking the process of translation between European and non-European cultures. For the discussion of universalistic assumptions about a human drive for being 'at home' involves very heterogeneous forms of belonging, alienation, or indifference towards a 'home.' Often the 'drive' to be at home seems quite easily contented with illusions of a spiritual home; yet in some cases historical processes lead to violent conflicts about presumed 'home countries' between certain groups.

Mary Douglas (1966) introduced anthropologists to a world of symbolic attributions, in which the body can signify everything, and everything can mean the body. The same applies to homes, when looked at as a rational construction of territorial belonging on the one hand, and as an image of embeddedness in a particular social and geographical landscape, or nostalgia about such a situation, on the other.

2. Ways of Being 'At Home'

Obviously, the portable shelter constructed by a group of Kung hunters and gatherers is not the same as the proverbial British or German 'castle.' Anthropologists observed, however, that so-called bush people in Africa, traditional Australian aborigines, as well as groups inhabiting the rainforests of South Asia and South America, have developed elaborate systems of navigation in their regions, based for example on landmarks or tactile qualities of the ground (Widlok 1997). In the nineteenth century, Australian aborigines imagined their ancestors to wander around and with their actions invent the current topographical shape of their habitat. In so doing they related to the 'ground' in a way not at all alien to European conceptions of identity defined by geographical, social, genealogical, and individual physical characteristics of a given group in a given area (Spencer and Gillen 1938). Attracted by some of the benefits of provincial town life or driven away by colonial powers from their places, their twentieth-century descendants may insist on returning to their promised lands, if their means of living allow such a step (Berndt 1977). Welfare programs can thus paradoxically lead to a partial return to traditional ways of living and especially of nurturing their ancestors' grounds. However, often we are no longer able to discern whether indigenous nationalists adhere to old projections of a social and spiritual life onto certain 'hunting grounds' and places of survival through gathering activities, or whether this just

follows Western concepts of belonging, grounded and capitalizing on geographical features of a given area.

There is varied ethnographic and mythological evidence of responses to loss of old nurturing grounds, including the wish to re-enchant the relationship between humans, heaven, and nature, although so far there is not enough comparative research available on the symbolism of land and house. Yet there is some evidence that human groups that succeed in storing greater amounts of food develop more closely-knit local mythologies bound up with the topography of landmarks, human dwellings, places for growing and harvesting food, as well as holy places. Horticulturists and cattle-breeding populations living in segmentary societies or low-grade forms of centralization tend to organize the household, the surrounding area, and far-away places that are of interest to the group by socially constructing ancestors and territorial spirits or divinities (Lienhardt 1961, Lévi-Strauss 1955). The 'juridic domain' (Fortes 1969) concerning the simple facts of life—dwelling together, dividing up food, working a territory, keeping a herd, and maintaining property rights—is thus organized by an explicit ideology of belonging, descent, holy grounds, and the profane (Kramer and Kramer-Marx 1993). A supernatural interpretation of shrines and natural features, even of artificially organized biotopes, can be the territorial expression of a temporal cult of vegetation which unites social groups in a harmonious view of the ego, the household, the family, and other related persons with its ground, including all the plants and animals living or to be raised therein (Turner 1969).

A 'system of total exchange' (Mauss 1960), of generalized reciprocity, seems to pervade and bind human groups and their natural surroundings together. Paradoxically, this can privilege those groups which have reasons to claim to have been the first humans in a particular region and can give them a superior status, heavily defended in conflict, or lost and exchanged with the status of a pariah. The conflict between Hutu and Tutsi in Central Africa has had such a component (Maquet 1961). In a time of international arms traffic, and armed interventions into local cultures, tensions of this kind can lead to severe cases of mass violence.

There is a definite difference between such closely knit groups or networks of groups and the nation states of world society, insofar as the 'total social fact' (Mauss 1960) of traditional societies is organized into a tight overlapping of the domains of landscape, social structure, cosmology, etc. (Fortes 1969). The father or uncle can be the earthmaster, celebrating a holy shrine, and the next moment he can be the leader of a group of young warriors. This does not mean, though, that the social reality of these groups is more homogeneous than ours. Whilst subscribing, like their fellows, to a certain degree to the governing harmonious ideology of space, time, and home, some outsiders, skeptics,

and dissidents always disagree and hold rational or disenchanted views of their group's attempts to maintain a homeland (Radin 1957).

By exchanging 'stigma' for 'charisma,' such outsiders can strongly affect or change their society. Thus, the classical 'big men' of New Guinea who constructed the 'pig festival' through the vegetational and reproductive cycles draw on a whole system of credits and cliental relations countering the heroic culture of the 'great men' who search for the spiritual gratification of the warrior (Godelier 1982). Early imperialist states seem to have been based on a similar mixture of exploiting nature, fellow, and other groups, on the one hand, and heroic fantasies about extending a holy landscape to other regions, on the other (Wolf 1999).

The study of nomadism might be particularly instructive with respect to human ideas about finding or having a home (Mauss 1960). For the layperson, nomads just seem to wander around without clearly relating to certain features of the landscapes. The experience of wandering and fugitive groups in Europe, such as the Sinti, Roma, and many others, may support this conception of a 'wandering' part of humanity that does not feel bound to certain cycles of exchange. Wanderers are always suspect to sedentary populations, and this resentment can remain through long periods of sedentarization, if it is supported by political powers. One of the starting points of German anthropology in the nineteenth century was a series of heated discussions about whether or not Jews felt any obligation of fairness and civil behavior towards non-Hebrew populations (Gilman 1990).

In the anthropological data, nomadic constructions of a unified horizon of landscapes abound, as does information on nomads treating their tents, carriages, or cars as 'mobile homes' (Rao 1987). Motifs in Oriental carpets, manufactured by nomads, often appear to carry the signum of descent and territorial rights as well as some kind of report of the group's history and historical claims for grounds and herds (Schuster and Carpenter 1996).

Not only nomadic populations but many other groups convey an ideal of homogeneity, tradition, and *mise en scène* of an environment of 'being at home' through their material culture, such as dress and costumes. This continues right up to European medieval, Ottoman, and today's international regulations of clothing and corporeal signs. Also the often painfully achieved imprints on the skins of members of more marginal populations seem to have the same rational and spiritual aim of demonstrating a form of belonging together (Clastres 1973). It is the invention of an 'environment,' in the sense that this term is used in modern art. Human environmental creativity ascribes to a certain space the property of being a home. These autopoietic processes of representing and creating an order of time, space, and body are perhaps most obvious in rural societies with their regional and

class-bound costumes, their festivals, shrines, and clear-cut regional politics of power.

But even in the case of feudal societies, which often enough did not allow their underclass members to leave their 'homes'—even in this continuum of slavery on the one hand and enthusiasm for one's own 'ground' on the other—we can find distinct groups who, through historical experience or the subtleties of property rights, seem to have acquired a very distanced attitude towards their houses, fields, and cattle (Lopreato 1967). In times of famine or collective violence, these groups were probably among the first to leave their grounds and wander to the cities or the promised lands, over the mountains and across seas, as soon as the opportunity was offered by the modernization of the world economy into the global capitalistic project.

3. Modern and Postmodern Homes

In cities and other modernized landscapes, anthropologists have observed distinct techniques deployed in designing the 'home' of migrated groups or former peasant inhabitants of the area (Hannerz 1980). Conscious city design, the building of ghettos for reasons of controlling the 'bio-power,' and the process of identity formation in subgroups of a given population all go hand in hand here. Homeland associations as well as informal networks amongst industrial workers and employees ascribe to their private and sometimes also their professional life characteristics of tribal life. The elite of industrialized areas tend to form closely-knit networks, sometimes secret, sometimes formally recognizable, which transform their actors into something reminiscent of the earthmasters and segmentary war chiefs as encountered in traditional societies.

In the third millennium, two-thirds of the world population are expected to live such a metropolitan way of life, and from what we know about the older metropolitan areas in the world, we can deduce that the postmodern metropolis will be partly made by neighborhoods defining and defending themselves like traditional 'villages in the city,' like those in Naples, Cairo, Rio de Janeiro, or New York. These 'lieux' and 'milieux' of trust and similarity will be defined by criteria of class and taste as well as ethnicity and religious confession. Homelessness will abound as well as those of the better off desperate to (re)create a home as an anchor to a very mobile style of life (Dehavenon 1999). The inhabitants of big cities try to overcome the anomic atmosphere of 'non-lieux' by staging and topographically creating city life through festivals and customs, charity and sport events, as well as a home life through decoration, invitations of friends, meals, and so on (Augé 1992, Cieraad 1999).

In addition to the withdrawals and the small refuges of identity, however, an extension of the word 'home' has become important. With the advent of the nation states, millions and even billions of humans can unite in very large conglomerations of landscapes, and by means of national symbolism they try to find not only a common political but also symbolic denominator. An organicist idea of the state—'a sound family makes a sound state' (Norman 1991)—may help to transfer feelings which were originally connected with house, family, regional polity, and local religion to the higher spheres of a fatherland united in the sign of a mother language, a home country (Gellner 1998).

This is where ethnicizing or nationalizing politics have often overlapped with anthropological science, a science often intended historically as a kind of school for learning how to invent nations. Not only under Hitler and Stalin did the 'nations of anthropology' (Borneman 1995) stand in close conjunction with their host nations, so did American cultural anthropology and other well-respected disciplines.

Today's 'United Nations' comprise groups that are divided by very different models of inventing traditions as well as interior constructions. Some of them are founded on the basis of the one and only territory alone, others claim that a common language, cultural heritage, race, or historical traumatization provide the fundament of their existence. However, the whole system of nations nowadays demands that no single person, house, or territory can be without a host nation. This is supported by older ethnographic constructions, which always attributed a group of people to a certain territory. Today, such origins of ethnic topographies frequently turn into a topography of terror, while at the same time international movements of refugees and migration make such attributions obsolete. Another form of homebuilding, however, will apparently continue to exist even in the most technologically advanced systems on Earth. Against all the prognostics of the social sciences, the 'home' seems to survive even its most important carrier, the family. Statistical evidence indicates that single households abound in all metropolitan areas of the world, yet this 'yuppie' culture maintains that people ought to belong to some house, some household (often lovingly decorated and furnished), a certain area, a group of friends and colleagues. Big festivals like Christmas—which is, indeed, the only feast of a truly international character celebrated today, not only in the West but in most parts of the world including China and Africa—provide this 'home' with an anchor in time and space. Groups of families and friends reunite during these days to have traditional forms of communality, exchange, and mythopoesis about a bearded old man leaving his occult home in the last area on the world that still proves unreachable for globalization, the polar circle, carrying with him gifts and goods that will help to make people feel more at home (Miller 1993).

4. Final Thoughts

In the era of 'moving targets' (Marcus 1999), anthropological fieldwork has partly lost its central referent: the small group inhabiting a restricted area since immemorial times—the field. However, it seems that humans all over the world are struggling to recreate homes under the circumstances created by today's form of globalization. The anthropological 'field' therefore re-emerges over and over again, sometimes 'under fire' (Nordstrom 1995) and sometimes in a peaceful dialogue with the anthropologist, that 'professional stranger' who leaves his home to study others' homes.

See also: Collective Beliefs; Sociological Explanation; Collective Memory, Anthropology of; Community, Social Contexts of; Community/Society: History of the Concept; Community Studies: Anthropological; Everyday Life, Anthropology of; Family, Anthropology of; Family as Institution; Folklore; Household in Anthropology; Household Production; Kinship in Anthropology; Nationalism, Sociology of; Place in Geography; Tradition, Anthropology of

Bibliography

- Augé M 1992 *Non-Lieux. Introduction à une anthropologie de la surmodernité*. Éditions du Seuil, Paris
- Berndt R 1977 *Aborigines and Change: Australia in the 70s*. Australian Institute of Aboriginal Studies, Canberra, ACT
- Borneman J 1995 American anthropology as foreign policy. *American Anthropology* 97: 663–72
- Cieraad I (ed.) 1999 *At Home: An Anthropology of Domestic Space*. Syracuse University Press, New York
- Clastres P 1973 De la torture dans les sociétés primitives. *L'Homme* 13: 114–20
- Dehavenon A L 1999 *There's No Place Like Home. Anthropological Perspectives on Housing and Homelessness in the United States*. Bergin and Garvey, Westport, CT
- Douglas M 1966 *Purity and Danger: An Analysis of Concepts of Pollution and Taboo*. Routledge and Kegan Paul, London
- Fortes M 1969 *Kinship and the Social Order*. University of Chicago Press, Chicago
- Gellner E 1998 *Nationalism*. New York University Press, New York
- Gilman S 1990 *Jewish Self-Hatred: Anti-Semitism and the Hidden Language of Jews*. Johns Hopkins University Press, Baltimore, MD
- Godelier M 1982 *La Production des Grands Hommes*. Fayard, Paris
- Hannerz U 1980 *Exploring the City*. Columbia University Press, New York
- Hobsbawm E, Ranger T (eds.) 1992 *The Invention of Tradition*. Cambridge University Press, Cambridge, UK
- Kramer F W, Kramer-Marx G 1993 *Zeitmarken. Die Feste von Dimodonko*. Trickster, Munich, Germany
- Kuper A 1988 *The Invention of Primitive Society: Transformations of an Illusion*. Routledge, London
- Lévi-Strauss C 1955 *Tristes Tropiques*. Plon, Paris
- Lienhardt G 1961 *Divinity and Experience. The Religion of the Dinka*. Oxford University Press, Oxford, UK

- Lopreato J 1967 *Peasants No More: Social Class and Social Change in an Underdeveloped Society*. Chandler, Scranton, PA
- Maquet J 1961 *The Premise of Inequality in Ruanda*. Routledge, London
- Marcus G (ed.) 1983 *Elites. Ethnographic Issues*. University of New Mexico Press, Albuquerque, NM
- Marcus G 1999 *Ethnography Through Thick and Thin*. Princeton University Press, Princeton, NJ
- Malkki L 1995 *Purity and Exile. Violence, Memory and National Cosmology among Hutu Refugees in Tanzania*. University of Chicago Press, Chicago
- Mauss M 1960 Essai sur le Don. In: Mauss M (ed.) *Sociologie et Anthropologie*. Presses Universitaires de France, Paris
- Miller D (ed.) 1993 *Unwrapping Christmas*. Clarendon Press, Oxford, UK
- Nordstrom C (ed.) 1995 *Fieldwork Under Fire. Comparative Studies of Violence and Survival*. University of California Press, Berkeley, CA
- Norman K 1991 A sound family makes a sound state. Ideology and upbringing in a German village. *Stockholm Studies in Social Anthropology*, 24. Almqvist and Wiksell International, Stockholm
- Radin P 1957 *Primitive Man as Philosopher*. Dover, New York
- Rao A 1987 *The Other Nomads: Peripatetic Minorities in Cross-Cultural Perspective*. Böhlau, Vienna
- Schuster C, Carpenter E 1996 *Patterns that Connect. Social Symbolism in Ancient and Tribal Art*. Abrams, New York
- Spencer B, Gillen FJ 1938 *The Native Tribes of Central Australia*. Routledge, London
- Turner V 1969 *The Ritual Process. Structure and Antistructure*. Routledge and Kegan Paul, London
- Vertovec S, Cohen R (eds.) 1999 *Migration, Diaspora and Transnationalism*. Elgar, Cheltenham, UK
- Widlok T 1997 Orientation in the wild: The shared cognition of Hai//om bush people. *Journal of the Royal Anthropological Institute* 3: 317–32
- Wilson E 1975 *Sociobiology*, abr. edn. Belknap, Cambridge, MA
- Wolf E 1999 *Envisioning Power. Ideologies of Dominance and Crisis*. University of California Press, Berkeley, CA

T. Hauschild

Homelessness: United States

Homelessness has now been a fixture on the political scene in the United States and other countries for two decades. Yet the scope of the problem and the appropriate remedies for it still prompt fierce debates. This article examines the difficulties of defining homelessness and establishing useful estimates of population size and dynamics. It then provides basic descriptive information about homelessness in the United States and looks at its patterns and antecedents. The article ends by describing the growth of homeless-specific service programs and considering the probable future directions for homeless research and policy. Information about homeless populations and services for the United States is based on data

collected in October and November of 1996 by the National Survey of Homeless Assistance Providers and Clients (NSHAPC). NSHAPC provides the most comprehensive overview available for the country as a whole (Burt et al. 1999).

1. Defining Homelessness

Homelessness did not exist as a public policy issue from the end of the Great Depression until the early 1980s. Certainly there were people without homes, or with very inadequate or dangerous homes. In some poor areas of many cities known as 'skid rows,' shelters, or 'missions' run by religious groups offered a bed and sometimes a meal in exchange for the chance to save a soul. These skid row areas also contained hotels where one could rent a small room, with perhaps a hotplate and a small sink, for a day, a week, or a month depending on how much money one had. In addition, rooming or boarding houses offered inexpensive accommodations.

Before 1980, the US Census classified people as 'homeless' if they lived by themselves and not with family members. Thus, all of the single people living alone in skid row areas (and elsewhere) might have been called homeless—a definition quite different from the one in use today, when very large numbers of the general population live alone by choice. Because of definitional problems and the absence of a concept close to what today would be called homelessness, there are no good estimates of the size of the literally homeless population for the 1960s and 1970s. The best guess is that in the late 1970s formal shelters (i.e., missions) for the homeless might have been able to accommodate about 50,000 people at any one time in the entire country, and that the number of literally homeless people probably was no more than 100,000 and could have been considerably less.

Since 1987, official US policy defines people as literally homeless if they sleep in a shelter for homeless people or in places not meant for human habitation such as vehicles, abandoned building, outdoor locations, or transportation facilities or stations. However, the people included in any actual study of the homeless, and therefore the descriptions of homeless populations and estimates of their size, will vary greatly depending on study methods (Taeuber 1991). If studies go only to shelters they will miss all homeless people who do not stay in them. Even studies that include searches of nonshelter locations have a very hard time finding most unsheltered homeless people. The best methods developed to date use services such as soup kitchens and outreach programs to find the nonsheltered homeless, and succeed better than street searches. In some European countries with a right to housing, homelessness is sometimes defined quite differently, as people who come to a housing authority needing housing (which they usually receive). Ob-

viously both definitions of homelessness and methods used to find homeless people will affect estimates of population size, while public policies will affect the types of remedy available and the probable length of homeless careers.

2. Homeless Numbers and Rates

The best point prevalence estimate of the size of the homeless population in the US before the early 1980s was 100,000 or fewer. By 1984 the estimate was up to 250,000 to 300,000, and by 1987 to 500,000 to 600,000. Estimates for 1996 of homeless people *using homeless assistance programs* (not all homeless people) range from 444,000 for October to 842,000 for February (Burt et al. 2001). All of these rates include any children accompanying a homeless adult, but not runaway and homeless youth.

The best way to compare estimates over time and place is to convert them to rates, which for homeless estimates is usually a rate per 10,000 population. Homelessness rates based on the 1987 estimate of 500,000 to 600,000 homeless people were 19.4/10,000 to 24.3/10,000. Projecting the February 1996 estimate of homeless service users to the whole homeless population assuming that service users during this winter season were 80 percent of the population produces an estimated rate of 38.0/10,000. This significant increase in the estimated rate of homelessness at any given time owes a good deal to the more than 200 percent expansion of homeless shelter capacity between 1987 and 1996 (see below), all of whose residents are classified as homeless.

Annual prevalence rates are a far better vehicle that point prevalence rates for understanding the proportion of a population affected by homelessness over time, because they account for the very large number of homeless spells lasting less than one year. Estimates for New York and Philadelphia indicate that 1 percent (100/10,000) of the entire population of those cities used a homeless shelter during a year's period (Culhane and Hornburg 1997, pp. 101–34). The same level of homelessness was estimated for the entire US for 1996 by Burt et al. (2001). When considered as a proportion of the *poor* population, the rate is close to 10 percent experiencing homelessness in a year's time.

3. Major Characteristics of Homeless People in the United States

3.1 Demographic Characteristics

In 1996, information from NSHAPC indicated that about one-third of homeless services users were mem-

bers of homeless families; 23 percent were the children in these families, and 11 percent were their parents. These families comprised 15 percent of all homeless *households*; 86 percent of which were headed by women. Among family households, 38 percent were white non-Hispanic, 43 percent were black non-Hispanic, 15 percent were Hispanic, and the remainder were from other races/ethnicities. Forty-one percent had never married, and only 23 percent were currently married. Educational attainment was very low, with 53 percent having less than a high school education. These families had a mean monthly income of \$475, or less than half of the US federal poverty level for a family of three (Burt et al. 1999).

Nonfamily homeless households were 61 percent men by themselves, 15 percent women by themselves, and 10 percent men or women with at least one other adult (Burt et al. 2001). Among nonfamily homeless service users, 41 percent were white non-Hispanic, 40 percent were black non-Hispanic, 10 percent were Hispanic, and 9 percent were from other races/ethnicities. Half of these homeless adults had never married, and only 7 percent were currently married. Equal proportions (36–37 percent) had completed high school and had less than a high school education. Mean monthly income was \$348 or 51 percent of the federal poverty level for a single person (Burt et al. 1999).

3.2 Other Characteristics

In 1996, less than 10 percent of homeless adults worked in a regular job while homeless. However, 44 percent did some paid work for a month. Homeless adults were almost disconnected from the financial support of family and friends, with only 21 percent receiving money from these sources in an average month. Homeless families were somewhat better connected to public benefits, with 52 percent receiving the primary public welfare benefit of Aid to Families with Dependent Children and associated food and healthcare benefits. About 1 in 10 homeless adults received Supplemental Security Income, a program for people with serious long-term disabilities preventing them from working. Many studies find high levels of alcohol, drug, and mental health problems among homeless populations; the NSHAPC included. During the year before the study, 46, 38, and 45 percent of homeless adults reported alcohol, drug, and mental health problems, respectively. Only one in four did not report any of these problems. Homeless women were more likely to report mental health problems without either alcohol or drug problems, while homeless men were most likely to report having problems only with alcohol and not with drugs or mental health issues.

3.3 Patterns of Homelessness

Peoples' experiences of homelessness vary considerably. Many researchers have differentiated three patterns of homelessness—first-time/crisis/transitional, episodic, and chronic—although the exact classification criteria vary across studies. The first group includes individuals or families who are homeless only once or twice and usually for a relatively short period of time (e.g., less than 3, 6, or 12 months). Often they have been precariously housed and find themselves homeless after an unexpected crisis such as unemployment, divorce/separation, or eviction. They usually secure some type of housing relatively quickly and often do not become homeless again. Episodically homeless people are defined by their pattern of cycling in and out of homelessness over significant periods of time. Many spend time in various institutions such as inpatient treatment centers, detoxification programs, and jail. Chronically homeless people have relied on shelters or lived on the streets for many months or years, and usually have multiple barriers to securing stable employment and housing. They may have only a few distinct spells of homelessness, but each one lasts for a very long time.

Point-in-time studies, which overestimate the proportion of homeless people with long spells, usually find that about one-quarter of homeless adults are in the first group, and that half or more have extensive histories of either episodic or continuous homelessness. Studies that are able to observe flows into and out of the homeless condition over time give a different picture, showing only 10 to 20 percent with long-term homelessness and up to half with quite short-term or crisis patterns (Culhane and Hornburg 1997, pp. 101–34).

4. Antecedents of or Factors Associated with Homelessness

Both societal and personal factors affect homelessness. The societal factors predict *when* and *where* homeless populations will increase or decrease; the personal factors help to predict *who* will be vulnerable to these societal pressures. The two sides of the equation determining the ability to afford housing are household income and housing costs. In the US, housing costs for the lowest-cost units has increased significantly in the past 2 decades. Also, homelessness most recently became a public issue in the early 1980s, during a time of record-high unemployment (the recession of 1981–1982) that reduced household income among both low-wage workers and workers in manufacturing, many of whom lost their jobs. The record prosperity of the US during most of the 1990s has been accompanied by the development of a two-tiered labor market that has relegated many people without a college or technical education to low-paying service jobs. At the

same time, wealthier households have pushed housing costs to record-high levels. The differential between earning power and housing costs sets the societal stage for homelessness. A public safety net of housing and other programs could assure that households do not founder under such circumstances. Some European countries have such a safety net, but the US does not. Some level of homelessness is the result.

That being said, certain people are more likely than others to become homeless. Poverty is the *sine qua non* of homelessness. The explanatory value of demographic characteristics such as race/ethnicity diminish or disappear in many analyses that simultaneously include income variables plus characteristics such as physical health, childhood experiences, and variables relating to alcohol and drug use and mental health problems. Regression results from an increasing number of studies on both families and singles confirm that sexual or physical abuse by a household member, foster care, or other out-of-home placement in childhood differentiate both homeless families and homeless singles from their housed counterparts. Physical and mental health conditions in adulthood, and early and current alcohol or drug use are also usually found to increase the probability that a person will experience homelessness when other things are held constant.

5. Growth in Services and Evidence of Effectiveness

5.1 Shelter/Housing Capacity

NSHAPC offers the first opportunity in many years to update knowledge about homeless assistance programs from the national perspective. Estimates based on NSHAPC data indicate that the nation's shelter/housing capacity within the homeless assistance system grew by 220 percent between 1988 and 1996, from the 275,000 beds estimated by the US Department of Housing and Urban Development (HUD) for 1988 to more than 600,000 beds in 1996. This growth came on top of the 275 percent increase in shelter capacity between 1984 and 1988 from 100,000 to 275,000 beds. One of the most striking things about the growth of shelter/housing capacity is the extent to which the system has gone beyond emergency shelter to include large numbers of transitional housing units, and permanent housing units for disabled people who were once homeless. In fact, relatively little growth occurred in emergency shelter programs, or beds specifically, between 1988 and 1996. HUD estimated 5,400 shelter programs in 1988, with a capacity of 275,000 beds. Virtually all of these programs and their beds supplied emergency accommodation, either through emergency shelters or through voucher arrangements with hotels and motels. In 1996, NSHAPC data indicate that the number of beds available through more than

8,800 emergency shelter plus voucher distribution programs had grown to only 333,500, or a mere 21 percent over the level in 1988.

In contrast, transitional and permanent housing units exploded, going from basically none in 1988 to about 275,000 in 6,300 programs by 1996. Thus, this entirely new capacity to help people with housing needs beyond emergency shelter was essentially equal, in 1996, to the entire homeless assistance system in 1988. This growth coincided with the initial availability and eventual expansion of federal funding for new types of shelter/housing stimulated by homeless legislation.

5.2 Soup Kitchen/M Meal Capacity

No national data exist to describe soup kitchen and other prepared meal distribution programs for 1987 that are equivalent to HUD's 1988 survey of shelter capacity. However, a 1987 Urban Institute study (Burt and Cohen 1989) did provide a roughly equivalent look at soup kitchen and mobile food program capacity in cities over 100,000, which can be contrasted to NSHAPC's estimates for central cities in 1996. Big-city prepared meal capacity increased almost four-fold between 1987 and 1996, from 97,000 up to 382,100 meals a day. Nationally, NSHAPC estimates are that soup kitchens and mobile food programs expected to serve almost 570,000 meals a day in February 1996, of which one-third were served in suburban/urban fringe areas or rural areas outside of central cities.

Unlike most shelter/housing programs, users of which are presumed to be literally homeless, soup kitchens and mobile food programs commonly serve many nonhomeless people (although some may have had earlier episodes of homelessness). In addition, people may eat more than one meal a day at these programs. For both these reasons, the estimate of meal service levels in 1996 cannot be interpreted as a level of homelessness. However, it certainly can be interpreted as a level of need. Also, soup kitchen/prepared meal programs have not had even one-tenth of the federal financial incentive to increase their service level that has been available to open or expand shelter programs over the past decade. This makes the service growth in the emergency food arena between 1987 and 1996 even more remarkable than the growth of the shelter/housing segment of homeless assistance. It attests to the responsiveness of many communities' emergency capacity, as well as to the ongoing problem of hunger even in a decade when most Americans enjoyed remarkable prosperity.

5.3 Evidence of Effectiveness

Effectiveness should only be assessed against expectations that are realistic to the programs and

services examined. If the goal is to supply shelter and food, then shelters and soup kitchens are effective in meeting that goal. If the goal is to end, or prevent, homelessness, then the growth in shelter capacity in the US has probably had the opposite effect. This perverse result arises from the definition of homelessness, which includes residence in a homeless shelter. It is unclear whether the underlying level of need—the pool of people desperate enough to seek shelter from the homeless service system—grew in the 1990s. But the expanded shelter capacity gave such people a place to go that did not exist in the 1980s, and they have used it.

In 1999, a federally sponsored symposium featured papers summarizing ten years of research on the effectiveness of services for homeless people (Fosburg and Dennis 1999). Most of the findings revealed that intermediate program goals were being achieved. For instance, outreach programs were able to engage homeless people with serious disabilities, and eventually connect them to services. Case management programs were also able to connect homeless people to services. Psychiatric services were able to help mentally ill homeless people with their mental health problems. However, on average, none of these programs were able to end homelessness for their participants unless they supplied housing (Shinn and Baumohl 1999). If they supplied housing, especially with supportive services as required, programs could help even the most disabled homeless people leave the streets and maintain themselves in stable residential situations.

6. Future Directions for Policy and Research

In the US the argument has come full circle after almost 2 decades. The notion that homelessness was a problem of 'housing, housing, and housing' was made forcefully in the early 1980s, and has now proved to be quite accurate. Service programs have helped many homeless people in many ways. But to help people leave homelessness for good, or to divert people from entering into homelessness, the resources required are housing resources. These include help with landlords, back rent, evictions, and move-in expenses as diversion outlays, and actual housing as a means to end homelessness. A big question exists as to the value of supportive services to accompany the housing, once the housing itself is supplied. Research on this issue is summarized by Shinn and Baumohl (1999). Some research, mostly on homeless families, suggests that housing is the key and that services add little or nothing. Other research, more focused on chronically homeless people with significant alcohol, drug, and/or mental health problems, indicates an important role for supportive or supervisory services to accompany housing.

Future research could profitably take two directions. The first direction would focus on what supports

or services made available *just before* a homeless episode or at the moment of its precipitating event could avert the transition to literal homelessness by stabilizing the situation or simply redefining the issue. For instance, some community support services for persons with chronic mental illness offer 'respite' facilities where a person can stay for several days during a crisis in their illness. Respite stays avert complete dissolution of the person's life in the community, including loss of housing. It would be important to learn how these types of interventions can work best, for whom, and when. The second direction for research would focus on what services *in addition to housing* would make the most difference for people who are already homeless. Some communities in the US have already shifted resources from emergency shelter to affordable and supported housing, actually reducing the need for emergency shelter beds. More attention needs to be focused on developing a range of such approaches, because it should be clear that continuing to increase the capacity of the homeless service network will only serve to bring more and more people into the category of homeless.

See also: Drug Addiction; Drug Addiction: Sociological Aspects; Drug Use and Abuse; Psychosocial Aspects; Housing Economics; Mental Health Programs: Children and Adolescents; Mentally Ill: Public Attitudes; Poverty, Culture of; Social Stratification; Social Support and Health; Social Support, Psychology of; Social Welfare Policy: Comparisons; Street Children: Cultural Concerns; Street Children: Psychological Perspectives; Unemployment: Structural; Welfare; Welfare Programs, Economics of; Welfare State

Bibliography

- Burt M R, Aron L Y, Douglas T, Valente J, Lee E, Iwen B 1999 *Homelessness—Programs and the People They Serve: Summary and Technical Reports*. Interagency Council on the Homeless/Department of Housing and Urban Development, Washington, DC
- Burt M R, Aron L Y, Lee E 2001 *Helping America's Homeless: Emergency Shelter or Affordable Housing?* Urban Institute Press, Washington, DC
- Burt M R, Cohen B E 1989 *America's Homeless: Numbers, Characteristics, and Programs that Serve Them*. Urban Institute Press, Washington, DC
- Culhane D P, Hornburg S P (eds.) 1997 *Understanding Homelessness: New Policy and Research Perspectives*. Fannie Mae Foundation, Washington, DC
- Fosburg L B, Dennis D L (eds.) 1999 *Practical Lessons: The 1998 National Symposium on Homeless Research*. US Department of Housing and Urban Development and US Department of Health and Human Services, Washington, DC
- Shinn M, Baumohl J 1999 Rethinking the prevention of homelessness. In: Fosburg L B, Dennis D L (eds.) *Practical Lessons: The 1998 National Symposium on Homeless Research*.

US Department of Housing and Urban Development and US Department of Health and Human Services, Washington, DC
Taeuber C M (ed.) 1991 *Enumerating Homeless Persons: Methods and Data Needs*. US Department of Commerce, Bureau of the Census, Washington, DC

M. R. Burt

Homosexuality and Psychiatry

Psychiatry has served both to reflect and to help define Western societal attitudes toward homosexuality for much of the past two centuries. Perspectives on homosexuality—concerning its origins, its social and personal meanings, its diagnostic implications, and its clinical relevance—persistently represented one of the most contentious issues within American psychiatry during the latter part of the twentieth century. With a primary emphasis on the American experience, this article presents current definitions related to homosexuality and sexual orientation and examines the history of psychiatric theories and findings about homosexuality. The relationship between homosexuality and psychiatry in other countries and cultures has often paralleled what took place in the US but also reflects the diverse attitudes and policies regarding both sexuality and mental illness throughout the world. An example of such diversity is the recent legislation of gay and lesbian relationships by some European countries in contrast to the continued opposition to such acceptance in the United States.

1. Definitions

Increased understanding of sexuality in general has led to greater specificity in the use of terms related to sex, gender, sexuality, and sexual orientation. Terms such as gender identity and gender or sex role tended to be confused with sexual orientation in early writing about homosexuality; however, it is now recognized that these are related but discrete characteristics that may interact with but are not synonymous with homosexuality.

Sexual orientation refers to a person's erotic response tendency, homosexual, bisexual, or heterosexual, toward persons of the same or other sex as reflected in such indicators as the proportion of dreams and fantasies directed to one or the other sex, the sex of one's sexual partners, and the extent of physiological response to erotic stimuli associated with one or both sexes. Sexual orientation consists of three components, desire, behavior, and identity, which may or may not be congruent in an individual. The Kinsey scale—a 7-point continuous scale, with 0

representing exclusive heterosexuality, 6 representing exclusive homosexuality, and 3 representing equal amounts of both, or bisexuality—is the most widely used scale for describing sexual orientation. The terms homosexuality, heterosexuality, and bisexuality have not been applied consistently in theoretical, research, or popular discourse; they have been used to refer to a wide range of constructs, including categories of sexual desire, gender role attributes, forms of sexual behavior, personal and social identities, types of personalities and persons, degrees of normality and abnormality, and the presence or absence of mental illness.

The presence of erotic desire for someone of the same or other biological sex does not imply the concurrent existence of any other characteristics in that individual. Thus, for example, a woman may engage in homosexual behavior, demonstrate typical feminine gender role characteristics, be married to a man, and experience a heterosexual identity; or a man may have homosexual desire, have sex only with women, and show gender role nonconformity.

The terms gay, gay man, lesbian, and bisexual are used to refer to men or women who have developed a sexual orientation identity that is homosexual or bisexual. While gay sometimes is used as an inclusive term to refer to both men and women, reference to two distinct groups, gay men and lesbians, is more appropriate because of the significant differences between these groups in their development and experiences.

Homosexual, or lesbian and gay, and bisexual identities are the subjective experiences of being homosexual or bisexual—being significantly or exclusively attracted to someone of the same sex—with the identities being both self- and other-ascribed. Models of lesbian, gay, and bisexual identity development usually portray a series of linear progressive stages involving tasks such as coming out (the process of recognizing one's homosexual or bisexual attraction and acknowledging it to oneself and to others), involvement in lesbian, gay, and bisexual communities, establishing same-sex relationships, and integration of one's sexual identity into other aspects of the self. Not all persons who experience homoerotic desire or participate in homosexual behavior also develop a lesbian, gay, or bisexual identity.

In contrast to sexual orientation, gender identity generally is understood to refer to a persistent sense of oneself as being male, female, or ambivalent and is contrasted with sex, which refers to the biological attributes of being male or female. While gender identity describes an inner experience of being male or female, gender role and social-sex role are defined as the degree to which an individual's outer behavior or appearance can be described as masculine, feminine, or androgynous. The large majority of persons, regardless of their sexual orientation, have a gender identity and gender role consistent with their bio-

logical sex, though sometimes gay and lesbian persons may demonstrate a greater range of gender role attributes.

2. Historical Conceptualizations of Homosexuality within Psychiatry

Prior to the time homosexuality was first studied as a psychiatric condition in the second half of the nineteenth century, the meaning of homosexual behavior was very different than it is today in Western culture. Homosexual behaviors take place in all societies and homosexual acts have been both ritualized and prohibited in various cultures. However, the contemporary idea of identifying a type of person based on regular participation in homosexual behavior and on development of an identity associated with this behavior is generally believed to be a product of the medicalization of homosexuality during the nineteenth century.

2.1 Medicalization of Homosexuality

During the second half of the nineteenth century a number of individuals in Europe began to study homosexuality scientifically. The term homosexual was first coined by the Hungarian physician Kertbeny in 1869. The conceptual basis for homosexuality arose from the earlier ideological construction of two genders, male and female, with dichotomous characteristics, and the attempt to find explanations for homosexuality derivative from these male and female categories. Thus, the homosexual was conceived of as a person of one biological sex with the sexual longing of the other sex, as portrayed, for example, in the German jurist and homosexual rights advocate Karl Heinrich Ulrich's description of the male homosexual as a female soul in a male body.

Subsequently, the German psychiatrists Carl Westphal and Richard von Krafft-Ebing adopted many of Ulrich's ideas, but they considered the traits he described to be unnatural. The German physician and sexual emancipator Magnus Hirschfeld also believed that there was a genetic basis for homosexuality. While he understood homosexuality to be a malformation in development and ascribed to a number of different biological etiologies, he also asserted that it was natural. Similarly, the English sex reformer Havelock Ellis based his efforts to alter attitudes about homosexuality on the belief that homosexuality was an inborn trait. Specific biological theories about the origins of homosexuality began to be examined during this period, including a variety of constitutional factors, such as different anatomical structures in the brains and in nerves in the rectums of homosexual men, and endocrinological determinants, including atrophied and abnormal testicles in homosexual men.

These early biological theories about the origin of homosexuality, with their emphasis on heredity, endocrinological influences, and anatomical or structural differences between homosexual and heterosexual persons, presage many current efforts to find a biological basis for homosexuality. Their almost exclusive emphasis on the study of homosexuality in men also parallels later twentieth-century biological research, which largely has ignored homosexuality in women. Further, the biological arguments helped to create the notion of a homosexual person by emphasizing the existence of fundamental constitutional differences between those who are attracted to persons of the same sex and those who are attracted to persons of the other sex.

2.2 Psychoanalytic Approaches

During the latter part of the nineteenth and the first half of the twentieth century, psychoanalysis, with its emphasis on intrapsychic and interpersonal development, represented an alternative approach to the biological sciences for studying and understanding homosexuality. Freud believed that heterosexuality was the natural outcome of normal development and that homosexuality represented an arrest in development from an instinctual bisexuality to a mature heterosexuality that arises as the result of a variety of factors. However, he did not view homosexuality as a sign of degeneracy and even asserted that homosexuality could occur in persons who had no other signs of deviation and no impairment in their functioning. Later psychoanalytic theorists, including Sandor Rado, who refuted Freud's model of innate bisexuality and developed his own theory of homosexuality as a type of phobia, emphasized far more heavily the beliefs that homosexuality is the result of problems in psychosexual development and that its occurrence is a sign of severe psychopathology.

The outcomes of the growth of interest in studying homosexuality within the biological and psychoanalytic fields during this period were similar: to establish a focus of interest about homosexuality outside of the realms of morality and religion within science and medicine; to reify a category of person, the homosexual, associated with homosexual behavior; and to confirm the abnormality and the pathology inherent in homosexuality.

2.3 Normalization of Homosexuality

The publication of the two volumes on male and female sexuality by Kinsey et al. (1948, 1953) marked the beginning of the move away from the view of homosexuality as pathology and toward an understanding of homosexuality as a normal variant of

sexual desire and behavior. Kinsey's studies demonstrated the widespread presence of homosexual feelings and behavior among samples of several thousand American men and women thereby suggesting that homosexuality was not an isolated and aberrant phenomenon. Kinsey's conceptualization of sexual orientation as a continuum also introduced the idea that homosexuality and heterosexuality were not dichotomous categories, but rather occurred in varying degrees in individuals.

Ford and Beach (1951) demonstrated that homosexuality was common across cultures and occurred in almost all nonhuman primate species. The psychologist Hooker's (1957) publication of the results of her ground-breaking study comparing the projective test results from 30 homosexual men and 30 heterosexual men showed that experienced judges could not distinguish between the two groups based on their test results, a finding which established for the first time that homosexuality in men was not associated with mental illness. The effect of this significant finding was to require that future views of homosexuality as an illness could only be based on an *a priori* assumption that homosexuality itself was pathological. By the 1970s, influential psychiatrists like Judd Marmor, a psychoanalyst and president of the APA, began to publish works that moved the psychiatric profession further away from the perspective of homosexuality as a type of mental illness and toward a consideration of the perspectives offered in the newer research findings.

During this period of transition and in spite of the availability of contradictory findings from empirical studies, some psychologists, particularly behaviorists who attempted to extinguish homosexual expression through aversive techniques, and psychiatrists, primarily psychoanalysts such as Irving Bieber and Charles Socarides, persisted in publishing works that presented homosexuality as the outcome of developmental disturbance and as a sign of severe psychopathology. Bieber's work (Bieber et al. 1962), which has been criticized extensively for its methodology, was particularly influential in its portrayal of a pathogenic family type, consisting of a detached and rejecting father and a close-binding and domineering mother, that led to homosexuality in adult men.

3. Classification of Homosexuality by the American Psychiatric Association

Homosexuality was classified as a type of mental disorder in the first *Diagnostic and Statistical Manual of Mental Disorders (DSM-I)* (1952). It was designated, along with the other conditions labeled as sexual deviations, as a type of sociopathic personality disturbance, defined by the existence of certain behaviors and not by the presence of distress or dysfunction. This classification of homosexuality as a

form of mental illness was not controversial at the time *DSM-I* was published, reflecting as it did the negative societal attitudes about homosexuality and the prevailing medical views of homosexuality. In *DSM-II* (1968), homosexuality continued to be listed as a type of sexual deviation.

Beginning in 1968, gay activists began to confront the APA about its classification of homosexuality as a type of mental disorder. Following a series of dramatic encounters between gay rights advocates and psychiatrists at several annual meetings of the APA, the organization began officially to re-evaluate its position. Following further political protest and significant scientific study of the issue, a proposal from the Nomenclature Committee of the APA to eliminate homosexuality from the *DSM* was approved by the APA's Council on Research and Development, Reference Committee, and Assembly of District Branches before being accepted by the APA's Board of Trustees in December 1973. As a result, for the first time in modern history homosexuality *per se*, was not classified as a type of illness. This action had enormous social significance in shifting the predominant and official view of homosexuality from one of sickness to one of acceptable behavior. The decision to declassify homosexuality was accompanied by passage of a position statement by the APA that supported the protection of the civil rights of homosexual persons:

Whereas homosexuality in and of itself implies no impairment in judgment, stability, reliability, or vocational capabilities, therefore, be it resolved that the American Psychiatric Association deplores all public and private discrimination against homosexuals in such areas as employment, housing, public accommodations, and licensing, and declares that no burden of proof of such judgment, capacity, or reliability shall be placed on homosexuals greater than that imposed on any other persons. Further, the APA supports and urges the enactment of civil rights legislation at local, state, and federal levels that would insure homosexual citizens the same protections now guaranteed to others. Further, the APA supports and urges the repeal of all legislation making criminal offenses of sexual acts performed by consenting adults in private.

The rationale for removing homosexuality from the *DSM* was based, in part, on the clarification of the definition of a mental disorder as meeting the criteria of either experience of subjective distress or of generalized impairment in social effectiveness or functioning as a result of the condition. Neither of these criteria applied to homosexual persons who were satisfied with their sexual orientation and who did not demonstrate impaired functioning. In what appeared to be a compromise between opposing views, a new diagnostic category, 'sexual orientation disturbance,' was introduced to apply to those homosexual persons who 'are either disturbed by, in conflict with, or wish to change their sexual orientation.' When the *DSM-III* (1980) was published, sexual orientation disturbance

was replaced with a new diagnostic category, 'ego-dystonic homosexuality,' a type of general psychosexual disorder that labeled homosexuality as a mental disorder only when it was accompanied by persistent distress. Finally, after further review of the appropriateness and utility of this diagnosis, all reference to homosexuality was removed from *DSM-III-R* (1986), representing the final acknowledgment by the American psychiatric profession that homosexuality is not a type of mental illness. This change in the official view of homosexuality signified a dramatic conceptual shift in the cultural meaning and significance of homosexual behavior within American society.

4. *Effects of Declassification*

The decision to remove homosexuality from the *DSM* led to widespread effects within psychiatry. At the organizational level, groups of lesbian, gay, and bisexual psychiatrists formed and many issues relevant to the mental health of lesbian, gay, and bisexual persons could now be considered. At a clinical level, while controversy has persisted among psychiatrists about how to conceptualize homosexuality, new understandings and techniques for evaluation and treatment based on gay and lesbian affirmative approaches were developed.

These changes were accompanied and shaped by larger contextual shifts that occurred as a result of removing homosexuality from the realm of pathology. Thus, for example, once homosexuality was no longer viewed as a disease, it became a topic that could more legitimately be studied within a variety of fields outside of psychiatry and medicine with their emphasis on pathology, such as the law, the social sciences, and the humanities. The expansion of gay and lesbian identities, communities, and culture in the years since the APA's action produced a plethora of new issues that interacted with and influenced mental health, such as coming out, relationship concerns, childbearing and raising, family interactions, and spirituality. Finally, new health concerns, particularly AIDS, substance abuse, and reproductive issues, have created a whole set of additional mental health issues relevant to lesbian, gay, and bisexual persons.

Following the APA's action to depathologize homosexuality, the formerly secretive 'GayPA,' consisting of an informal group of closeted homosexual psychiatrists, began to evolve into a more open and extensive social and political network of gay and lesbian psychiatrists. At the 1978 annual meeting, a formal structure was adopted for the Gay, Lesbian, and Bisexual Caucus of the American Psychiatric Association, which was subsequently re-named the Association of Gay and Lesbian Psychiatrists (AGLP). Consisting of several hundred members, the AGLP is an affiliated group of the APA and is one of

the largest organizations of gay and lesbian physicians in the world. Many of its members served as founding members of other gay and lesbian medical groups, and it has continued to provide support to members and to raise political and clinical issues relevant to gay, lesbian, and bisexual mental health within the psychiatric profession. In addition, a Committee on Gay, Lesbian, and Bisexual Issues was established within the structure of the APA to study and make recommendations regarding mental health concerns. As a designated minority and underrepresented group within the APA, members of the Caucus of Lesbian, Gay, and Bisexual Psychiatrists elect representatives to the Assembly of District Branches, and openly lesbian and gay psychiatrists have been appointed and elected to many leadership positions within the APA.

The substantive issues discussed within the APA relevant to gay, lesbian, and bisexual concerns have involved a variety of topics, such as parenting and custody, education, reparative therapy, and AIDS. This activism and openness could not have occurred without the radical paradigm shift that took place when homosexuality was no longer considered a disease. The reconstruction of homosexuality as a normal variation of sexual desire, behavior, and identity allowed its introduction as a legitimate topic for discussion within the field of psychiatry.

Within the 1980s and 1990s psychiatrists in academic settings have been acknowledged and promoted for their work on gay and lesbian topics. Openly gay and lesbian researchers and clinicians have helped to establish a whole new field of gay and lesbian mental health. Publication of a section on homosexuality in the *Review of Psychiatry* (Oldham et al. 1993), a comprehensive *Textbook of Homosexuality and Mental Health* (Cabaj and Stein 1996), a curriculum for learning about gay, lesbian, and bisexual issues in psychiatry training programs (Stein 1994), a *Journal of Gay and Lesbian Psychotherapy*, and numerous scholarly articles and volumes on affirmative approaches to working with gay, lesbian, and bisexual patients has provided a breadth of valuable resources for helping gay men and lesbians within mental health settings. As a result of this material, many mental health facilities and practitioners are increasingly sensitive to gay and lesbian needs and informed about gay and lesbian concerns.

Problems have remained even after the revolution in perceptions of homosexuality precipitated by the diagnostic change made by the APA in 1973. Some psychiatrists and other mental health professionals continue to view homosexuality and gay, lesbian, and bisexual persons in negative ways and to perpetrate harmful practices, such as involuntary hospitalization of gay, lesbian, and bisexual youth because of their sexual orientation and presentation of unsubstantiated claims for 'curing' homosexuality. Groups within the APA concerned with gay, lesbian, and bisexual mental health concerns have worked to enact

new policies that will eradicate the vestiges of heterosexism and antisexuality. In December 1998 they were successful in having the APA approve a Position Statement on Psychiatric Treatment and Sexual Orientation which states that:

... the American Psychiatric Association opposes any psychiatric treatment, such as 'reparative' or 'conversion' therapy which is based on the assumption that homosexuality *per se* is a mental disorder or based on the *a priori* assumption that the patient should change his/her homosexual orientation.

5. Future of Homosexuality and Psychiatry

While the discourse within psychiatry regarding homosexuality has been expressed primarily in the form of scientific arguments about developmental and therapeutic conceptualizations related to sexual orientation, these arguments reflect the larger social debate within the US about attitudes and policy toward homosexuality. The psychiatric profession specified a clear governing principle—that homosexuality is not a type of mental illness—in its 1973 decision to de-pathologize homosexuality, but struggled for the next 25 years to articulate a consistent strategy for dealing with the implications of this decision. Perhaps official psychiatric policy was ahead of the larger society in its unequivocal rejection of a view of homosexuality as a type of disease. Like American society in general, however, psychiatry continued to maneuver to accommodate two totally conflicting viewpoints about sexuality, morality, and social values that were condensed in attitudes toward homosexuality. It remains to be seen if the definitive opposition to reparative therapy approaches to homosexuality expressed in the 1998 position statement represents final closure within psychiatry on this enduring debate.

The impact of diagnostic and treatment changes in relation to homosexuality within psychiatry during the past three decades has extended far beyond the mental health field. First, in 1973, by removing homosexuality itself from the medical spheres of etiology and pathology, the psychiatric profession yielded dominion over sexual orientation to those who dealt with normal variations of behavior, such as psychology, sociology, and other biological and behavioral sciences. The freeing of homosexuality from its century-long ownership by psychiatry can be seen as a major impetus to the expansion and deepening of understanding about the topic within many fields and to the burgeoning growth and expression of gay and lesbian identities and communities. Once the gate to the prison of pathology had been unlocked, those who had previously been incarcerated because of their sexual identities lost no time in claiming their newly found freedom and, at the beginning of the twenty-first century, are demanding assertively equal access to fundamental social institutions and rights such as

marriage, parenting, and legal protection based on an assumption of their normality.

See also: Ethics and Psychiatry; Gay, Lesbian, and Bisexual Youth; Gay/Lesbian Movements; Heterosexism and Homophobia; Mental Health and Normality; Sexual Orientation and the Law; Sexual Orientation: Biological Influences; Sexual Orientation: Historical and Social Construction; Sexual Risk Behaviors

Bibliography

- American Psychiatric Association 1952 *Diagnostic and Statistical Manual: Mental Disorders*. American Psychiatric Association, Washington, DC
- American Psychiatric Association 1968 *Diagnostic and Statistical Manual: Mental Disorders*, 2nd edn. American Psychiatric Association, Washington, DC
- American Psychiatric Association 1980 *Diagnostic and Statistical Manual: Mental Disorders*, 3rd edn. American Psychiatric Association, Washington, DC
- American Psychiatric Association 1986 *Diagnostic and Statistical Manual: Mental Disorders*, rev. 3rd edn. American Psychiatric Association, Washington, DC
- Bayer R 1987 *Homosexuality and American Psychiatry: The Politic of Diagnosis/Ronald Bayer; with a New Afterword on AIDs and Homosexuality*. Princeton University Press, Princeton, NJ
- Bell A P, Weinberg M S 1978 *Homosexualities: A Study of Diversity among Men and Women*. Simon & Schuster, New York
- Bell A P, Weinberg M S, Hammersmith F K 1981 *Sexual Preference: Its Development in Men and Women*. Indiana University Press, Bloomington, IN
- Bieber I, Dain H J, Dince P R, Drellich M G, Grand H G, Grundlach R H, Kremer M W, Rifkin A H, Wilbur C B, Bieber T B 1962 *Homosexuality: A Psychoanalytic Study*. Basic Books, New York
- Cabaj R P, Stein T S (eds.) 1996 *Textbook of Homosexuality and Mental Health*. American Psychiatric Press, Washington, DC
- D'Augelli A R, Patterson C J (eds.) 1995 *Lesbian, Gay, and Bisexual Identities Over the Lifespan: Psychological Perspectives*. Oxford University Press, New York
- Ford C S, Beach F A 1951 *Patterns of Sexual Behavior*. Greenwood Press, Westport, CT
- Hooker E A 1957 The adjustment of the male overt homosexual. *Journal of Projective Techniques* 21: 17–31
- Kinsey A C, Pomeroy W B, Martin C E 1948 *Sexual Behavior in the Human Male*. Saunders, Philadelphia, PA
- Kinsey A C, Pomeroy W B, Martin C E 1953 *Sexual Behavior in the Human Female*. Saunders, Philadelphia, PA
- Lewes K 1988 *The Psychoanalytic Theory of Male Homosexuality*. Simon and Schuster, New York
- Magee M, Miller D C 1997 *Lesbian Lives: Psychoanalytic Narratives Old and New*. Analytic Press, Hillsdale, NJ
- Oldham J M, Riba M B, Tasman A 1993 *Review of Psychiatry*, Vol. 12. American Psychiatric Press, Washington, DC
- Ryan C, Futterman D 1998 *Lesbian and Gay Youth: Care and Counseling*. Columbia University Press, New York
- Stein T S 1994 A curriculum for learning in psychiatric residencies about homosexuality, gay men, and lesbians. *Academic Psychiatry* 18: 2–59

Weinberg M S, Williams C C J, Pryor D W 1994 *Dual Attraction: Understanding Bisexuality*. Oxford University Press, New York

T. S. Stein

Honor and Shame

Anthropologists conducting research into Mediterranean societies first developed the notions of honor and shame to describe the particular complex of values they found there. Honor comprised both an individual's sense of self-worth and this person's reputation in the surrounding community. Shame, by contrast, arose from the failure to act according to social values and it entailed public disgrace. This article traces the development of these concepts within anthropology and considers their current applicability.

1. Honor in Anthropology

By the middle of the twentieth century the word 'honor' was well on its way to obsolescence in ordinary English language usage. The practice of dueling to redress injuries to one's honor had ended and consequently honor had ceased to be a topic of public concern. The term was mainly frozen into stock expressions such as 'word of honor' or fixed in the aspic of formalized conventions such as the Queen's Honours List, or the school honor roll. Honor was still a familiar word, yet people were not absolutely certain what it meant; it was rich in connotations, but poor in denotation.

The Oxford-trained anthropologist Julian Pitt-Rivers, scion of a venerable English family whose forebears may well have participated in the culture of honor dueling in the last century, capitalized on the evocativeness of 'honor.' He reintroduced the word in a wide-ranging essay that drew principally on data from early modern European history. In this historical period honor drifted away from its medieval sense of a monarch's gift of fiefs in return for (military) services. In the seventeenth century it came to mean a general social reputation and then in the early nineteenth century it took on a more interiorized sense of integrity or moral dignity. If publicly insulted—called a liar or cuckolded, for example—a man was within his rights to demand the satisfaction of a duel to restore his honor.

Pitt-Rivers used a polysemous amalgamation of the historical senses of honor to interrogate his own contemporary ethnographic data from a Spanish village. His essay appeared alongside studies of other Mediterranean societies in a book entitled *Honour and Shame: The Values of Mediterranean Society* (Peri-

stian 1965) that launched honor and shame in anthropology. The premise of the volume was that '[M]editerranean peoples ... are constantly called upon to use the concepts of honor and shame in order to assess their own conduct and that of their fellows' (Peristiany 1965, p. 10).

For the most part honor and shame were conceived to be opposites and furthermore to orient different gender-linked ideals. Men were most likely to claim honor aggressively through acts of violence or cunning that earned them a fearsome reputation. This reputation was maintained by an ongoing performance of swagger and gruffness that staked out an inviolable space around a man's property, livestock, and women. As Peristiany noted (1965, p. 9), if provoked, 'a Greek Cypriot, a Bedouin and a Berber may answer "I also have a moustache"'.

For their part, women protected this honor by comporting themselves with the modesty born of a sense of shame. Among the Sarakatsani shepherds of northern Greece described by Campbell (1964, p. 287), women do not run for fear that they might fall over backwards and expose their underwear, thereby shaming themselves irretrievably. Ideally a young woman should stay at home working on her dowry, well away from the public gaze so that her name will not be on people's lips.

2. Comparative Considerations

The Mediterranean focus of the *Honour and Shame* volume spurred further ethnographic research within this 'culture area' and also attempts to theorize why such a similar set of values had been arrived at in this region. Virtually all anthropologists assumed that honor and shame were relatively consistent in form throughout the Mediterranean until Michael Herzfeld (1980) pointed out, on the basis of research in two very different Greek communities, that honor amounted to 'socially appropriate behavior.' In a quiet, law-abiding village, macho swagger and aggressivity do not form the basis of honor.

The appreciation of the relativity of honor forced specialists into a theoretical stock taking (Gilmore 1987). Two definable tendencies emerged: (a) honor and shame vary in form in different communities, and indeed may be found in completely non-Mediterranean societies such as Japan where there is a great emphasis on shame; (b) although not exclusive to the Mediterranean, Mediterranean societies do, nonetheless, seem to converge on some strikingly similar cultural expressions of honor and shame. The gestural sign of the billy goat's horns made with extended index and baby finger, for example, signifies that a man has been cuckolded in a number of Mediterranean societies. The northern European resort to the image of the cuckoo bird to represent marital betrayal

indicates the imagery of a different cultural area. In Mediterranean societies goats furnish the relevant vocabulary. A man who has been betrayed by his wife wears horns in Greek (*keratás*) and in Italian (*cornuto*) while in Spain he is just a big goat (*cabrón*). Unlike rams, billy goats allow their females to be mounted by other males without protest.

3. Performance: Class and Gender

In early modern Europe honor was mainly an affair of the aristocracy; slights from social inferiors could be ignored without consequences. The majority of early Mediterraneanists worked in small rural communities where everyone was putatively a social equal. All men began with an equal portion of honor that they could increase or lose. A single system of honor and shame thus permeated the whole community but in larger, stratified towns or cities there could be several parallel systems of honor by class. Honor itself was not economically determined, but the circles within which it was contested, and the modes of contestation, were.

There can be few societies lacking altogether in notions of precedence and social failure. The particular expressive performance of prestige in Mediterranean societies was surely what attracted the attention of anthropologists. These performances might involve elaborate verbal displays such as the spontaneous exchange of insulting rhyming couplets, or the skilled telling of tales of sexual prowess or daring animal theft. The anthropologists recording these stories usually came from the middle classes of northern European or North American societies. Their fascination with difference precluded the recognition of their own participation in systems of honor (Pina-Cabral 1989, p. 402). If transgressed a middle class person might conceivably report the incident to the police and perhaps bring charges. If defamed in public, this person might well register an accusation of libel and fight the case through the legal system rather than by direct, violent retribution. This configuration is no less a system of honor and shame, but it is not *the* system of honor and shame that Mediterraneanists had in mind.

Mediterranean honor assumed its particular form as a mode of direct action in the absence of moderating state institutions. The place to find similar configurations of honor and shame in many Western societies is often among the working classes or in criminal subcultures. American inner-city gangs have developed an honor culture that is perhaps even more 'Mediterranean' than that of Mediterranean societies as can be seen in the following statement by a Los Angeles Chicano gang member:

Honor is important to me and every homey [gang member] because if you ain't got honor, you got nothing, man. I mean,

no amount of money can buy you honor, so you do everything you can to protect it. I mean, without honor, nobody will respect you, not your friends or your family (Jankowski 1991, p. 143).

That middle class anthropologists were apparently oblivious to these alternative systems of honor in their own societies exemplifies the earlier contention that parallel, discrete codes for honorable comportment may operate in stratified societies.

Feminist studies of honor and shame have rejected the idea that honor is exclusive to men, and contended that women are not to be construed primarily as the passive defenders of male honor. Although their silence has been taken as a sign of submissiveness it might actually be a form of subversive irony through which women mock the supposed gender hierarchy. Furthermore, women's performances of suffering (their own pain, or on behalf of family members) can be understood as claims to respect in the community of fellow women sufferers. Dubisch (1995) has observed that Greek women engage in competitive suffering as a means of gaining prestige in the community. A gender studies model, then, where performances and social constructions of masculinity and femininity are treated independently, is increasingly replacing the honor and shame model which linked them together in a version of normal patriarchy.

4. The Migration of Honor

Just as the initial optimism in the revelatory power of honor and shame began to wear out in anthropology, scholars in other fields discovered the formulation and adopted it with excitement. Where only a few books about honor and dueling in European history had been written before 1970, by the 1990s many titles had been published analyzing these themes in several European countries. In his *Southern Honor* (1982), Wyatt-Brown acknowledged the influence of Pitt-Rivers and speculated that the similarities between the American South and the Mediterranean arose because 'primitive conditions of life often produce elemental social values sometimes roughly parallel from one society to another.'

Like *mana* (to which Pitt-Rivers compared it), taboo, or shaman, honor belongs to middle-range anthropological theory—terms discovered ethnographically in one region and then applied to guide investigation in other societies. The Azande distinction between witchcraft (innate) and sorcery (learned), for example, has been used in the identification of witchcraft in numerous other societies. Honor and shame appear to have been discovered in Mediterranean anthropology and to have traveled out to other societies and other disciplines. Yet if we look more closely honor and shame had already traveled from

early modern European history into Mediterranean anthropology via Pitt-Rivers' essay. That these concepts now seem more attractive to Europeanist historians than they do to Mediterraneanist anthropologists may well be because they actually arise directly from these historical societies. Honor and shame were indigenous concepts in early modern European society.

The future of honor and shame in Mediterranean studies is murky precisely because they do not emerge from, or perfectly correspond to, local vocabularies. It is striking how few Mediterranean societies actually have words for 'honor' in the sense advanced by Pitt-Rivers and how infrequently equivalent terms are employed in daily life. In Greece the words *timi* 'honor' and *philótimo* 'sense of honor' are rarely heard while in Italy the operative concept is neither shame nor honor, even though the latter would seem to be readily accessible in the Italian word *onore*. Instead, exemplary moral fiber is displayed by following 'the code of silence' (*omertà*) and not disclosing details of criminal or other in-group activities to inquisitive authorities. In his early monograph describing an Andalusian village, Pitt-Rivers (1954) focused attention on 'shame' (*vergüenza*) and made no mention of 'honor.'

The conception of honor brought into anthropology by Pitt-Rivers was always slightly vague and polysemous and this was its theoretical attraction. Honor could involve criminal aggression or Christian virtue, its forms extended into gender relations and equally into the politics and economics of patronage and mafia protection. Anthropologists now find that the vagueness of honor and shame and their mismatch with comparable local concepts blunt their usefulness for ethnographic research and for comparison. Perhaps the most important legacy of the honor and shame formulation may prove to be in the area of the anthropology of self and emotion.

See also: Culture and Emotion; Emotions, Psychological Structure of; Emotions, Sociology of; Feud and Internal War: Legal Aspects; Guilt; Mafia; Shame and the Social Bond; Tribe; Violence, History of; Violence in Anthropology

Bibliography

- Campbell J K 1964 *Honour, Family and Patronage: A Study of Institutions and Moral Values in a Greek Mountain Community*. Clarendon Press, Oxford, UK
- Dubisch J 1995 *In a Different Place: Pilgrimage, Gender and Politics at a Greek Island Shrine*. Princeton University Press, Princeton, NJ
- Gilmore D D (ed.) 1987 *Honor and Shame and the Unity of the Mediterranean*. American Anthropological Association, Special Publication 22, Washington, DC
- Herzfeld M 1980 Honor and shame: problems in the comparative analysis of moral systems. *Man* 15: 339–51

- Jankowski M S 1991 *Islands in the Street: Gangs and American Urban Society*. University of California Press, Berkeley, CA
- Peristiany J G (ed.) 1965 *Honour and Shame: The Values of Mediterranean Society*. Weidenfeld and Nicolson, London
- Pina-Cabral J 1989 The Mediterranean as a category of regional comparison: a critical view. *Current Anthropology* 30: 399–406
- Pitt-Rivers J A 1954 *The People of the Sierra*. Weidenfeld and Nicolson, London
- Pitt-Rivers J A 1965 Honour and social status. In: Peristiany J G (ed.) *Honour and Shame: The Values of Mediterranean Society*. Weidenfeld and Nicolson, London
- Wyatt-Brown B 1982 *Southern Honor: Ethics and Behavior in the Old South*. Oxford University Press, New York

C. Stewart

Hope and Hopelessness

According to mythology, only hope remained in Pandora's jar after envy, spite, and revenge had been unleashed. But, how are we to interpret this ending? Was hope to be the antidote for these evils, or a counterproductive force? Although Judeo-Christian writers espoused the virtues of hope, Plato, Sophocles, Nietzsche, and others have criticized it for prolonging suffering. Even more problematic, however, are the prevalent historical views of hope and hopelessness as being vague and useless concepts. Contrary to these traditional perspectives, however, hope and hopelessness have emerged in recent decades as having clearly articulated definitions, reliable and valid instruments for their measurement, and relationships with crucial life outcomes.

1. Definitions and Measures of Hope

In the 1950s and 1960s, psychologists (e.g., Stotland 1969) and psychiatrists (e.g., Menninger 1959) described hope in terms of positive expectations for goal attainment. In the 1970s and 1980s, researchers used experiential, spiritual, cognitive, and interpersonal frameworks for their positive expectancy hope models (Farran et al. 1995).

Three approaches to hope have received the most attention. First, Stotland (1969) hypothesized that hope reflected a greater than zero expectation of attaining important goals, and that it helps in understanding how antecedent events lead to subsequent outcomes. To measure hope, people are queried about their goal attainment expectations, or observed by raters. Although Stotland allowed for the possibility of trait propensities, he focused upon hope in particular situations.

Second, Averill et al. (1990) used a semistructured questionnaire approach, and found that people view hope as being appropriate when it is: (a) a short-term response to specific environmental conditions, and (b)

tapping important, societally endorsed goals with 50 percent attainment probabilities.

Third, Snyder (2000) and colleagues, in what is called 'hope theory,' have posited that goal-directed expectancies involve two separable cognitive components—agency, or the person's perceived capability to initiate and sustain movement toward a goal, and pathways, or the perceived capability to produce effective routes to goals. Using this model, trait and state measures of hope in children and adults have been developed and validated (e.g., the 12-item dispositional Hope Scale for adults, Snyder et al. 1991).

2. Similarities and Differences of Hope Relative to Other Constructs

Because of its increasing scholarly attention since the early 1990s, hope theory will be used in order to explore the similarities and differences of hope to other theories. Viewing hope as agentic and pathways, goal-directed thinking provides an overlap with the Scheier and Carver (1985) generalized positive outcome expectancies model of optimism (indeed, this latter model is similar to the definitions of hope in the 1950s through 1980s); however, the pathways component of hope theory adds an explicit recognition of mental 'routing' capabilities in goal-directed thought.

In contrast to the Seligman (1991) model, in which learned optimism is viewed as an attributional process for distancing oneself from undesired outcomes, hope theory emphasizes a person's self-referential thoughts about reaching desired goals.

What about hope and self-esteem? It is posited (and supported in research) that perceptions of successful goal-directed activities (i.e., hope) drive a person's self-esteem. Thus, hope is a causal determinant of self-esteem.

Hope also resembles the self-efficacy model, with the pathways component being similar to what Bandura (1997) has called outcome expectancies, and agency being similar to his efficacy expectancies. A fundamental difference in the two models, however, is that the self-efficacy thoughts are given the pivotal driving role in Bandura's theory, whereas the agency and pathways goal-directed thoughts of hope theory are given equal weight as they iterate throughout the thinking process toward the attainment of goals.

3. Definitions and Measures of Hopelessness

Beck et al. (1974) defined hopelessness as a lack of enthusiasm, a motivational tendency to give up, and a dearth of positive expectancies. Using this definition, they developed the Beck Hopelessness Scale. Patterned theoretically after the Beck Hopelessness Scale, Kazdin et al. (1983) developed and validated the Hopelessness Scale for Children (ages seven and older).

Reasoning that elderly persons often perceive pessimism and futility about their futures, Fry (1984) conceptualized this perspective as hopelessness; moreover, he developed the Geriatric Hopelessness Scale to measure diminished abilities, lost interpersonal worth, and difficulty in recovering spiritual faith.

4. Hope and Hopelessness as Related to Other Variables

Dispositionally high-hope, as compared to low-hope, persons appraise their goals more in terms of successes than failures, they have higher perceived probabilities of attaining goals, and are more likely to experience positive emotions. Also, the positive appraisals of high-hope relative to low-hope people become especially advantageous when coping with stressors. Likewise, higher hope correlates positively with subsequent successful academic outcomes throughout the range of grades (elementary school through college), and better athletic performances (Snyder 2000).

Turning to the antithesis of hope, elevated hopelessness has correlated significantly with higher depression, as well as to more suicidal intentions and suicide completions in adults (Beck et al. 1990); moreover, greater hopelessness in children has related to more depression and suicidal ideation (Asarnow and Guthrie 1989).

5. Developmental Processes

5.1 Building Hope in Children

Newborns supplant raw sensations with perceptual representations. This learning of 'What's out there' leads to the selection of important objects (e.g., mother's face), thereby reflecting the emergence of desired goals. Infants also learn to recognize co-occurrence in events, and babies can anticipate events and act with intention (Schulman 1991). Also, babies learn 'what goes with what' survival skills by attending to those persons and things that fulfill their needs. Additionally, by age three, children can articulate their expectations about desired outcomes.

From ages three to six, hopeful thinking is enhanced as the child's vocabulary grows from approximately 50 to 10,000 words. By providing the means for describing goals and communicating with others, words and language help to frame hopeful thinking.

Through their reading or physical exploration, 7 to 12-year-old children expand their knowledge for use in goal pursuits; moreover, they increase their memory storehouses. These children also realize that their hopeful thoughts must be negotiated in the contexts of other persons (Flapan 1968).

Spanning the years 13 to 18, adolescents test their hopeful thoughts by gaining more understanding and experiences with interpersonal relationships.

5.2 Undermining Hope in Children

The events that diminish hope typically are more damaging when encountered earlier in the child's life, or when they continue for longer periods of time over childhood. Neglectful caregiving interferes with hopeful thinking because children in such circumstances do not have anyone educating them about goal-directed thinking. Likewise, abuse is particularly damaging because children fear the very caregivers who are supposed to instruct them in hopeful thinking. Also, parental loss by divorce or death can diminish hopeful thinking because children come to doubt the stability of relationships, as well as the fact that one can meet goals in the contexts of such relationships. Lastly, caregiving in which children's efforts are not supported lessens hope (Dweck 1999).

6. Increasing Hope

6.1 Children

In regard to goals, babies are aided by caregivers in identifying, naming, and using specific words for desired objects. The caregiver's interest and conversations about the child's goals are productive, as are reinforcements of goal-making.

To increase agentic thinking, children profit by being reminded that they are responsible for making things happen. Praise when children show determination to reach their goals solidifies their subsequent agentic thoughts. Similarly, helping children to practice positive self-referential talk (e.g., 'I can do this') enhances agency (Kanfer et al. 1975).

Pathways thinking can be promoted by showing babies how causality works (e.g., using drums). Songs, stories, and rhymes illustrate how words are sequenced, and how one thing leads to another. Teaching mental scripts about what children are to do in given situations is helpful, as are instructions about how to break routes into smaller, more easily attained steps.

For all of the aforementioned hope lessons, it should be emphasized that close attachments with caregivers are essential for fostering children's hopes.

6.2 Adults

It is important in producing hope that adults set goals based on their desires rather than what other people want for them. Additional helpful strategies include producing an array of goals, setting goals that stretch one's talents, and clearly articulating and setting concrete goal markers so as to facilitate the noticing of progress (Nezu and D'Zurilla 1981).

To enhance agentic thinking, particularly during difficult times, positive self-talk is useful. By learning to anticipate blockages, and to think of these impediments as challenges, people can enhance their determination (especially for when they encounter prob-

lems). People also gain mental lifts by remembering previously successful experiences. Furthermore, switching from truly blocked goals to more reachable ones increases agency.

To enhance one's pathways thinking, breaking a long-range goal into smaller subgoals is helpful. Other productive strategies include developing mental maps to goals, planning alternative routes, and rehearsing mentally. Friendships also provide powerful guidance for pathways-related thinking.

7. Future Directions

Research is needed into parental and school environments that are maximally conducive to hope, as well as the environments that must be avoided so as to lessen hopelessness. Also, additional hope and hopelessness research is warranted in regard to health, psychotherapy, interpersonal relationships, worker satisfaction and performance, stress and coping, and aging.

See also: Childhood Depression; Depression; Depression, Clinical Psychology of; Depression, Hopelessness, Optimism, and Health; Health Risk Appraisal and Optimistic Bias; Learned Helplessness

Bibliography

- Asarnow J R, Guthrie D 1989 Suicidal behavior, depression, and hopelessness in child psychiatric inpatients: A replication and extension. *Journal of Clinical Child Psychology* **18**: 129–36
- Averill J R, Catlin G, Chon K K 1990 *Rules of Hope*. Springer-Verlag, New York
- Bandura A 1997 *Self-efficacy: The Exercise of Control*. Freeman, New York
- Beck A T, Brown G, Berchick R J, Stewart B L, Steer R A 1990 Relationship between hopelessness and ultimate suicide: A replication with psychiatric outpatients. *American Journal of Psychiatry* **147**: 190–95
- Beck A T, Weissman A, Lester D, Trexler L 1974 The measurement of pessimism: The hopelessness scale. *Journal of Consulting and Clinical Psychology* **42**: 861–5
- Dweck C S 1999 *Self-theories: Their Role in Motivation, Personality, and Development*. Psychology Press, Philadelphia
- Farran C J, Herth A K, Popovich J M 1995 *Hope and Hopelessness: Critical Clinical Constructs*. Sage, Thousand Oaks, CA
- Flapan D 1968 *Children's Understanding of Social Interaction*. Teachers College Press, New York
- Fry P S 1984 Development of a geriatric scale of hopelessness: Implications for counseling and intervention with the depressed elderly. *Journal of Counseling Psychology* **31**: 322–31
- Kanfer F H, Karoly P, Newman A 1975 Reduction of children's fear of the dark by competence-related and situational threat-related verbal cues. *Journal of Consulting and Clinical Psychology* **43**: 251–8
- Kazdin A E, French N H, Unis A S, Esveltdt-Dawson K, Sherick R B 1983 Hopelessness, depression, and suicidal intent among psychiatrically disturbed inpatient children. *Journal of Consulting and Clinical Psychology* **51**: 504–10
- Menninger K 1959 The academic lecture on hope. *The American Journal of Psychiatry* **109**: 481–91
- Nezu A, D'Zurilla T J 1981 Effects of problem definition and formulation on decision making in the social problem-solving process. *Behavior Therapy* **12**: 100–6
- Scheier M F, Carver C S 1985 Optimism, coping, and health: Assessment and implications of generalized outcome expectancies. *Health Psychology* **4**: 219–47
- Schulman M 1991 *The Passionate Mind*. Free Press, New York
- Seligman M E P 1991 *Learned Optimism*. Knopf, New York
- Snyder C R (ed.) 2000 *The Handbook of Hope: Theory, Measures & Applications*. Academic Press, San Diego, CA
- Snyder C R, Harris C, Anderson J R, Holleran S A, Irving L M, Sigmon S T, Yoshinobu L, Gibb J, Langelle C, Harney P 1991 The will and the ways: Development and validation of an individual-differences measure of hope. *Journal of Personality and Social Psychology* **60**: 570–85
- Stotland E 1969 *The Psychology of Hope*, 1st edn. Jossey-Bass, San Francisco

C. R. Snyder

Horticultural Societies

A horticultural society is one in which people's livelihoods depend on the cultivation of small plots, commonly called gardens. These frequently support a wide range of crops and meet staple subsistence needs. The proportion of produce sold on the market varies, to finance the purchase of commodities manufactured elsewhere (e.g., clothing, processed food, hardware, etc.). The scale of society and its organization may vary widely too, from small tribal groups to large peasant populations.

1. The Term 'Horticulture'

Horticulture is distinguished from agriculture by scale of production, specialization, and commercialization. Agriculture features in industrialized societies supplying food to large urban populations. It is, in some regards, the reverse of horticulture. It involves the farming of large fields, usually under a monocrop (all wheat, potatoes, beans, etc.), whereas the gardens of horticulturists usually feature intercropping, the mixing up of several different crops, as opposed to their segregation in separate rows or plots, for which there are sound ecological reasons (e.g., pest and disease control, maximizing nutrient and water availability, increasing photosynthetic efficiency, exploiting micro-environmental variations to best effect, and so on). Agriculture features highly mechanized production techniques, whereas horticultural societies rely more on human labor and animal drought power. Farmers in agricultural systems meet few of their own subsistence requirements, specializing in the large-scale production of a few foodstuffs for sale on the market. They engage in agribusiness, intensive production making wide use of inputs to boost yields, frequently

obtained from off-farm and industrially manufactured (inorganic fertilizers, biocides, etc.). The term horticulture has a quite different meaning in urban-industrial societies, which can confuse the unwary. It refers to the intensive production, often under artificial glasshouse conditions, of high-value crops, such as early vegetables, salads, and flowers by commercial growers.

Other food production regimes from which horticultural societies are distinguished include gatherer-hunters and pastoralists, which support lower population densities (Forde 1934), industrialized urban ones having higher densities. These categories were bequeathed to us by Victorian scholars who thought of them comprising an evolutionary hierarchy, putting horticultural societies somewhere in the middle of the range between the gatherer-hunter and industrial poles. This scheme continues to inform thinking to the current day (as evidenced by entries in this encyclopedia), and underpins notions of development conceived as facilitating the passage of communities to the urban-industrial pole. Anthropologists continue to use it, although unhappy with the continued evolutionary implications, arguing that we should not judge urban-industrial life, notably liberal market democracy, as the superior socioeconomic order to which all human beings should aspire. It is an approximate classification scheme. At some ill-defined point, populations are thought to manage their natural resources sufficiently closely to move beyond gatherer-hunting to farming, this is accompanied by a more sedentary lifestyle. They qualify as horticultural societies when they emphasize crop cultivation over animal herding, although mixed arable and livestock farming is common. The label horticultural society covers a wide range of different subsistence regimes and socio-cultural orders. The term is more commonly used in North America than Europe, where instead of lumping these societies together, it is more usual to distinguish the different agronomic regimes that comprise horticultural orders.

2. Agronomic Regimes

We find varying degrees of sedentariness among people who depend largely on cultivation to supply them with their food. The variation informs a preliminary distinction between different horticultural strategies. There is also some correlation with increasing population densities.

Where we find the least permanent settlement and physically most direct technologies, we have cultivation systems that fallow the land not by rotating crops but by rotating land use. We call this type of horticulture *shifting cultivation*, because after people have grown crops, and fertility declines, and yields start to fall (these cultivators not using any fertilizer to supplement the soil's nutrients), they abandon the plot

to the forest and move elsewhere to clear a new garden on fresh land. Secondary regrowth, meanwhile, recolonizes the abandoned area, which is left fallow for tens of years to allow nature to restore its fertility; following which it may be cleared and planted again in the next generation. Shifting cultivation is labor intensive, featuring hand tools, such as axes, machetes, hoes, and pointed digging sticks. We find a sexual division of labor, orchestrated into a joint effort. Men usually do the heavy axe-related work. They fell the trees and clear the forest. They also enclose the area to be gardened with rough fencing if this is necessary to keep out animals. The subsequent burning off of the dry and dead vegetation cleared from the plot is often done by both men and women working together. This burning is an important step, reducing standing vegetation to ash, which is readily mixed with the soil, liberating locked up nutrients in a form more available to planted crops. The release of nutrients (such as nitrogen, potash, phosphorus, and so on) through burning has led to the coining of the term 'swidden' to describe this type of cultivation. (Swidden is an Old English word meaning to burn.) Following clearing and burning off, women usually take over and are responsible for breaking up the soil, planting crops, weeding, and finally, harvesting the plot. People living around the world in Asia, Africa, the Pacific, and Latin America grow a large variety of crops using this cultivation technique: for example, grains such as corn, millet, and sorghum; roots such as manioc, yams, and sweet potato; and fruits such as bananas, pawpaws, and melons.

Many Amazonian Indian populations practice shifting cultivation, horticultural crops supplying the bulk of their food regardless of the time spent gathering and hunting in the surrounding forest (Watters 1971). Among the Yanomamo, plantains and bananas are the staple crop. They select well-drained sites, relatively clear of thorny underbrush. They do not attempt to burn large tree trunks, left strewn plots in a jumble, serving as boundaries between patches, the crops planted there belong to those who cultivate them. Newly established gardens yield crops in cycles, the rapidly growing ones such as maize coming first, followed by others including manioc, taro, sweet potato, and mapuey. The Yanomamo also grow chewing tobacco, cotton for hammock yarn, drugs for shamanistic rituals, and many fruits including papaya, avocados, gourds, and hot peppers. A garden produces constantly as it matures, but a regular supply of food demands careful management. Gardens have direction, moving little by little as people gradually abandon older areas as thorny underbrush encroaches. They constantly clear adjacent land and transplant secondary plantain suckers as they appear. Sometimes villages are obliged to establish new gardens elsewhere. This is usually a political decision, communities moving when threatened by raids from enemy villages. It requires a large effort to

establish a new garden some way off, to transport needed cuttings and seeds. New gardens also produce sporadically at first, and people may face hunger, whereas established 'rolling gardens' offer predictable subsistence.

At some ill-defined point, agronomists distinguish *bush fallow* systems. These are broadly similar to shifting systems with regard to technology and cropping strategies, but with a reduced fallow interval and longer periods under cultivation. Instead of one or two years under cultivation followed by two or more decades under fallow forest, these horticulturists may farm land for 5–10 years with a fallow interval of similar duration under bush or savanna vegetation. Stock may graze on the fallow, cattle, goats, sheep, and so on. Settlement is increasingly stationary, and land holdings are usually more clearly defined with agreed field boundaries. These fallow systems sometimes evolve from shifting regimes with increasing population densities. They may demand management of soil fertility through the application of manure or fertilizer, otherwise yields will decline over time. Degradation of the natural resource base may occur with increasing population pressure and demand to grow cash crops.

Again, at some ill-defined point, we distinguish *ley* systems. These are ones in which farmers systematically use the land to pasture livestock during the non-arable phase. At one end are unregulated leys where people agist their animals on unmanipulated natural regrowth of grasses and herbs. They have communal grazing rights, and practice little pasture management. At the other end are regulated leys where farmers manage the grazing vegetation, frequently putting down grasses and legumes as fodder. They maintain individual pastures which they may fence off. They may graze meadows directly or use them to produce hay or silage. At this point we usually think they have evolved into agricultural regimes.

Bush fallows and unregulated leys are common in interlacustrine East Africa (Allan 1967). Lugbaraland, for example, comprises fields and homesteads scattered across open upland country. Homesteads comprise compounds, clusters of family huts and granaries, under which are ghost shrines and magical plants. Fields surround them. Each homestead cultivates three types of fields. Men do the heavy work of opening the land, women weed and harvest. Both sow, depending on the crop, men handling more cash crops. They intercrop in all. Postharvest care and use of crops is the responsibility of women. The 'riverine fields', irrigated by channels from streams, are highly fertile and need only short fallows, they support sweet potato, maize, sugar cane, and bananas. The 'home fields', fertilized with ashes and manure, support nutrient-demanding crops such as beer-brewing white sorghums. The 'outside fields' support staple eleusine, together with sorghum, simsim, and pigeon

pea, cultivated for two or three years, after which they are left under bush fallow to restore fertility. People allow unregulated grazing of cattle and goats on fallow. They recognize different fallows: from recent and regaining fertility; a few years old, the presence of certain plants showing restored fertility; cleared land left for a few months before hoeing; through to uncultivated land. Each woman in a homestead should have areas in all fields, of equal fertility. If not, quarrelling is likely. The fields of a homestead are likely to be scattered, due to differences in soil and water. This sets a limit on the maximum size of farming groups. A household short of certain farmland may move and take up residence with kinsmen elsewhere, usually a wife's or mother's brother.

It is generally agreed that with the invention of the plough we approach the realm of *agriculture*. The plough allows people to harness either the strength and energy of draught animals or engine-powered machines to assist in their preparation of land for the planting of crops. The result is that one person, together with their team of oxen or whatever, can accomplish considerably more work alone, releasing others to follow different occupations. We have the emergence of increasingly specialized labor, a move from so-called organic to mechanical solidarity. We find the practice of crop rotation and occasional fallowing to maintain soil fertility with the plough and associated agricultural machinery, together with the increasing use of amendments such as manure, fertilizer, biocides, etc.. In regions that suffer water deficits, people may have irrigation technologies. In some places they may have plantations of perennial commercial tree crops. The upshot is permanent farming, with clearly demarcated fields. While it is inappropriate to go into the growth of the first civilizations among the early cultivators of the fertile crescent in the Middle East, of Mesoamerica and of the Far East, we should note some of the consequences of these developments, founded on a settled agricultural way of life (Sauer 1969). They allowed for greater population densities, which, combined with the relative permanence of settlement that results with agriculture, led to the emergence of wider-scale political groups than those based on the family or tribe. These developments opened the way over millennia to the evolution of industrial-urban societies, sometimes with associated peasantries.

Before turning from the agronomic aspects of horticultural societies to consider the social orders associated with them, it needs to be stressed that while the foregoing regimes have been presented with an evolutionary dimension, it would be wrong to consider them as discrete steps leading from shifting cultivation to mechanized agriculture. This typology is no more than a low-level generalization by which we can usefully classify cultures. When we look closely we find that many societies do not fit comfortably into any one

category, but span two or more, as the previous East Uganda example illustrates. In the Baliem valley of New Guinea too, where sweet potato is the staple and people herd pigs, we find three sorts of cultivations. The largest are the permanent plots on the valley floor, some of which the Dani recultivate immediately upon harvest and others they leave fallow for months, even years. These gardens feature elaborate ditch systems which drain away/store water, delineate and protect cultivations from pigs, and serve as mulching basins supplying organic rich mud for slopping over beds to ensure continued fertility. Other gardens men clear in a classic shifting mode on the mountain slopes, burning off plots and planting for a year or two. The third and smallest are kitchen gardens maintained adjacent to houses, supporting a wide range of valuable crops such as tobacco, sugar cane, and bananas, continually fertilized with household waste. Households commonly have several different widely separated fields at different stages.

3. *Sociocultural Orders*

We now turn briefly to look at the gross nature of the social orders found associated with horticulturists. It is difficult to distinguish between horticultural regimes according to political economy (how people distribute resources, etc.). They occur in one of three broad categories: egalitarian with balanced reciprocity, chiefdoms with redistribution systems, or peasantries in nations with markets. There is some correlation between permanence of settlement, cultivation regime, and polity, with shifting cultivators frequently living in egalitarian orders and ley farmers in chiefdoms.

The *egalitarian societies* featuring balanced reciprocity are characterized by groupings commonly called tribes. These are relatively large groups of people (that is, in comparison to hunter-gatherer bands), who occupy demarcated territories, to which resources they lay specific claim and rights. Characteristically these societies recognize all individuals as equals and place no persons above others, with authority and power to rule over them. They are what we call stateless or acephalous, and have evolved various strategies to maintain social order without formal government, judiciary, and so on; employing, for instance, descent and age in Africa, and ceremonial exchange in Melanesia. The productive group is not the tribe, but usually the family, with some help from extra kin in occasional large cooperative ventures. Under this household mode of production farmers consume most of their produce, although they distribute some to others. The formal distribution of food and goods is likely to be significant, taking place predominantly in reciprocal transactions. Relatives and friends who give things to one another keep a careful account. They give, not with the expectation of receiving an immediate return gift (or payment), but

with the expectation of an equal return sometime later. There is often a competitive element in these transactions, those doing best winning higher status. They sometimes constitute grand institutions and are attended with some ceremony. Exchange is a facet of social and political relations and may exist independently of material wants and needs. This takes us away from subsistence and horticulture towards socio-political issues.

In Melanesia, ceremonial exchanges of balanced reciprocity are central to the maintenance of social order. New Guinea highlanders live in an egalitarian society. Each family supplies its own food and provides its own shelter. It occupies a compound often comprising several related families and friends, with men's and women's houses, and pig sty. Sometimes friends and neighbors make up a cooperative work party to garden or build a house. But there are no leaders. The Dani combine for conviviality, each person doing the same thing, the output comprising the sum total of their efforts (it is more fun mudding gardens together, for example, even though this is a one-man job). Feuding is common, clans fighting with bows and arrows and spears. Social order is promoted through exchange, primarily of pigs, which are only killed on ceremonial occasions. The giving and receiving of pigs constantly refurbishes the social network, people are bound to one another through their lifetimes by the exchange of hundreds, even thousands, of pigs. Clan affiliation determines access to land, all have equal rights. These are rights of usufruct, not land ownership. Men do the heavy, spasmodic garden and construction work, and are more involved in public activities such as politics, warfare, and rituals. Women undertake routine cultivation work and domestic responsibilities, herding pigs, and tending family needs. There is little call for distribution of goods with this minimal sexual division of labor. There is some trade with people elsewhere for a few imported goods.

The *chiefdoms* with redistribution systems are hierarchically organized societies, where members are ranked one above the other. In the top rank are the rulers, commonly called chiefs. These persons have authority vested in them, often sanctioned by some supernatural connection, and govern those living in their chiefdom, kingdom, tribe, or whatever. Everyday rule is usually delegated to lower rank persons, who are locally responsible for day-to-day government, administration of justice, and so on. These chiefdoms are often feudal in nature, with these local leaders owing allegiance to the chief and supplying a military force if required by him. Societies organized in this way exploit raw materials on a larger scale. They do this, for example, by territorial conquests (as epitomized by the Zulu under their king Shaka) and through extensive trading networks (for example, the Aztecs controlled mining activities as far away as northern Mexico). These societies can also organize

larger productive groups through the authority of their rulers, who by so doing not only assist people but also provide some sort of rationale for the existence of their power (for example, the farming terraces, irrigation works, storehouses for produce, and so on, constructed by the Inca). Redistributive systems predominate in these societies, where those of lower rank give food and goods to those of higher rank in the form of tribute. The chief will rearrange everything he receives and then redistribute it to his subjects (usually via local-level officers in his government). He may add to it any surplus goods which his probably superior productive position supplies (i.e., from his large, communally-worked gardens).

The chiefdoms of South Africa are good examples of this political order. The horticulture practiced is broadly similar to East Africa, the main crops being maize and millet, together with ground nuts, gourds, sugar cane, and pumpkins. Farming is seasonal, dependent on the summer rains. One source of rulers' political authority is the belief that they can control the advent of the rains. People may have a precarious existence, food supplies fluctuating annually between plenty and scarcity bordering on famine. The Swazi, for example, talk of lean summer months as the time 'to swallow the pickings of the teeth.' Horticulture is not a prestigious occupation and is left primarily to women, whose main tool is the iron hoe. Except, that is, for ploughing, which involves highly valued oxen handled by men, and the cultivation of cash crops, notably tobacco and cotton. Households cultivate large fields using ox-drawn ploughs. The power rulers wield also relates to their right to allocate land to their subjects, who depend on it for their livelihood. The Swazi say that their political overlords 'serve' them the land, securing their rights by allegiance and usage. Women are allocated garden areas by their husbands on their marriage. Homestead plot size depends primarily on supply of resident labor. People use grazing lands communally. They value pastoralism highly, having the so-called African 'cattle complex': these animals have great symbolic value, being exchanged to ratify marriage, sacrificed to propitiate ancestors, and paid in tribute to rulers. They call the king the 'Bull' of the nation, and the royal herd far exceeds any other, supplying animals redistributed to subjects at national events and sacrifices. In addition to sex and age, pedigree also structures the division of labor. Men and women cooperate in agriculture, man's work is heavier and more spasmodic, woman's routine and more continuous. Aristocrats do not engage in manual labor to the same extent as their subjects, they are responsible for successfully organizing their efforts, supervising work parties and arranging for specialists to doctor land and seed. The rulers have several 'gardens of kingship' in different localities, of which local headmen arrange communal cultivation. Homesteads rely on kinsmen and neighbors to assist at bottlenecks, customarily organized

into two competitive work parties they spur each other on in songs, rhythmically coordinating their work through music. The size of communal work parties reflects status: they enhance prestige, being rewarded with beer and food at the host's home.

The *peasantries* associated with markets are also incorporated into hierarchically organized societies. They are usually encapsulated within nation states where they occupy marginal and low positions, relatively powerless and commonly exploited. These are horticultural communities in peripheral, under-privileged situations, frequently taken advantage of, even abused by the dominant society. They usually feature extensive poverty. The larger political formations which rule significant aspects of their lives may supply them with certain services. Peasants depend in some regards on economic and political institutions beyond their local control. They may maintain informal reciprocal exchange networks within their communities, but outside they are obliged to participate in impersonal commodity markets. While they supply their own subsistence needs (similar to tribespeople, from whom many of the Third World's peasants have descended in recent times), they are obliged also to cultivate a surplus to sell on outside markets to earn needed cash. In addition to using this money to purchase goods manufactured elsewhere, they need some cash to pay the taxes levied on them by the colonial/national governments to which they are subject. They may also need to purchase outside inputs, such as inorganic fertilizers and biocides, to maintain their yields, especially if their farming system is under increasing demographic pressure, perhaps resulting in increasing environmental degradation. This may contribute to growing poverty. What peasants receive financially in return for their surplus horticultural produce is on the whole unfair, the elite higher up the social hierarchy, controlling access to imperfect markets and the arrangements for the distribution of produce, make a disproportionate profit to maintain their privileged position and life-style.

The Mayan Indians of Mexico have occupied a peasant status increasingly since the Spanish conquest (Wilken 1987). They sell crops surplus to their subsistence needs, largely maize and beans, and trade salt from the Ixtapa works at fiesta day markets. In turn they purchase goods such as coffee, sugar, fish, liquor, and bread from markets. They have been involved in marketing for centuries, historical evidence showing that they traded with Aztecs in prehispanic times. Maize is their staple, believed to have an 'inner soul' in the ear, together with beans, squashes, chillies, chayotes, and fruits such as peaches. They keep chickens and pigs, sheep for wool woven into *chamaras*, blankets and shawls, and mules for carrying heavy loads. The Mayans cultivate three types of land. Small, individually owned garden plots located near highland homes, their ancestral patrimony. Larger *ejido* plots

acquired during the 1940s agrarian reform program, allocated by local committees to households. And plots in the lowlands rented from wealthy *Ladino* landowners by patrilineal groups, which they divide equally among themselves. They use centuries' old swidden techniques, working with axes, machetes, bill hooks, and hoes. After cutting and burning vegetation, they plant using pointed digging sticks. In old fields they cut and rake maize stubble into piles for burning, and then hoe sheep manure into the turned topsoil. They prepare fields for planting at the onset of the rainy season. Ritual attends planting, a critical task in which a shaman offers candles, incense, liquor, and prayers to the Earth Owner, who controls wind and clouds. People cultivate fields for varying periods of time, depending on elevation and soil quality, before allowing brush to grow again. They store threshed maize in granaries or indoor grain bins prior to sale.

Again this sociocultural typology, like the previous agronomic one, is only gross, the elements constituting approximate cultural sets which have proved useful in organizing our ideas over many years. It continues to structure our thoughts in a preliminary fashion. There is some broad correlation between peoples' horticultural arrangements, that is how they produce food and other items necessary for life, and how they are socially organized, both to effect production and to achieve distribution of its fruits. But this is not to imply, as some writers argue, that mode of production determines the nature of social arrangements.

4. Future Directions

Several questions remain to be answered about horticultural societies. Many aspects are still poorly understood, and insight sometimes blinkered by orthodoxies. The Baliem Valley, for example, supports one of the highest population densities in New Guinea (~160 persons per km²), with no evidence that it is pushing against the environmental limits. How can the region's intensive horticulture system support such a high population density without any outside fertilizer amendments? The accepted shifting cultivation model which pertains to this part of the world predicts evidence of environmental strain, of the kind commonly associated with overpopulation. Yet the sophisticated horticultural regime does not seem to exhaust the soil (Sillitoe 1996).

A major contributor to our understanding of horticultural societies has been 'farming systems research' conducted under the aegis of international development agencies (Shaner et al. 1982). This involves multidisciplinary terms, using survey techniques largely, to detail the ecological, technical, and socioeconomic aspects of horticultural regimes and identify constraints on production amenable to scientific research. It was initially associated with 'modernization,' research intended to advance the development

of horticultural societies into agricultural ones, to enmesh them in commercial farming and specialized production for the market place. It was thought that this would alleviate poverty, leading to increased production of a food surplus necessary to support increasingly urban populations specializing in industrial, administrative, and service sectors. To this end, agricultural research institutes worldwide embarked on programs to breed new 'high-yielding varieties' of crops and develop associated technologies to increase production. The scientifically resoundingly successful 'green revolution' diminished the threat of famine. Nonetheless, it increasingly became evident that while the 'revolution' had benefited resource-rich and wealthy landowners, it had largely passed poor farmers by, increasing wealth differentials. It also worryingly increased dependency on outside manufactured inputs (inorganic fertilizers, fossil-fueled machinery, chemical pesticides, and herbicides) and gave rise to concern about long-term sustainability with evidence of increasing environmental pollution, susceptibility to mutated pathogens, insecure supply of needed inputs, and so on. The move towards cash cropping also put farmers at the mercy of market fluctuations, sometimes causing their families considerable hardship, further increasing resistance to modernization and adoption of commercial production. We become aware that 'modernization' not only scarcely helped poor households farming in marginal and fragile, resource-poor environments, but may have made some wealthier ones vulnerable too.

The realization that not only were some small-scale subsistence-focused horticultural societies proving very resistant to change as envisioned by modernization, but that other communities were increasingly exposed to serious perturbations beyond their control in attempting to accommodate its interventions, prompted a rethink of development priorities in the 1980s. It was increasingly acknowledged that poor people need to be involved in the planning and implementation of any development interventions intended to assist them, with the consequences of what international bankers euphemistically called economic restructuring taking effect (a dramatic instance of outside market forces impacting negatively on farmers' fragile commercial enterprises, as their nations' burgeoning debts were rescheduled). The 'participation' approach to development came of age. This promises to advance understanding and the interests of horticultural societies in the future, particularly with the concomitant emergence of 'indigenous knowledge' research, which represents the coming together of applied farming systems research and academic disciplines such as anthropology, geography, and ecology which have long esteemed understanding of horticultural societies (Warren et al. 1995). The assumption is that any development assistance needs to be sympathetic to local sociocultural and environmental contexts and to accommodate

peoples' aspirations. Understanding these implies their cooperation. It also furthers meaningful appreciation of the complex relations that inform natural resources management, as understood by local farmers, those who contend daily with them—the collapse of many interventions has been attributed to a failure to comprehend the complexity of their livelihood systems. These changes in research and development agendas bode well for furthering understanding in the future, and promoting the interests of horticultural societies *along their own lines*.

See also: Agricultural Change Theory

Bibliography

- Allan W 1967 *The African Husbandman*. Oliver and Boyd, London
- Altieri M A 1995 *Agroecology: The Science of Sustainable Agriculture*, 2nd edn. Westview Press, Boulder, CO
- Bayliss-Smith T P 1982 *The Ecology of Agricultural Systems*. Cambridge University Press, Cambridge, UK
- Carlstein T 1982 *Time Resources, Society and Ecology*. Allen & Unwin, London
- Forde C D 1934 *Habitat, Economy and Society: A Geographical Introduction to Ethnology*. Methuen, London
- Gliessman S R (ed.) 1990 *Agroecology: Researching the Ecological Basis for Sustainable Agriculture*. Springer Verlag, New York
- Grigg D B 1974 *The Agricultural Systems of the World*. Cambridge University Press, London
- Norman M J T 1979 *Annual Cropping Systems in the Tropics*. University Presses of Florida, Gainesville, FL
- Ruthenberg H 1976 *Farming Systems in the Tropics*, 2nd edn. Clarendon Press, Oxford, UK
- Sauer C O 1969 *Agricultural Origins and Dispersals*, 2nd edn. MIT Press, Cambridge, MA
- Shaner W W, Phillipp P F, Schmehl W R 1982 *Farming Systems Research and Development*. Westview Press, Boulder, CO
- Sillitoe P 1996 *A Place Against Time: Land and Environment in the Papua New Guinea Highlands*. Harwood Academic, Amsterdam
- Warren D M, Slikkerveer L J, Brokensha D 1995 *The Cultural Dimensions of Development: Indigenous Knowledge Systems*. Intermediate Technology Publications, London
- Watters R F 1971 *Shifting Cultivation in Latin America*. Food and Agriculture Organization, Rome
- Wilken G C 1987 *Good Farmers: Traditional Agricultural Resource Management in Mexico and Central America*. University of California Press, Berkeley, CA

P. Sillitoe

Hospitals and Health: Psychosocial Aspects

During the second half of the twentieth century, the dominant mode of health care delivery in the United States has evolved from individual, patient–physician

relationships to care provided by organizations (Starr 1982). At the same time, social scientists have moved from examining hospitals as closed systems, which operate independently of their environment, to seeing them as open systems, both affected by and responding to external forces (Scott 1998). This perspective assumes that structural forms and organizational behavior in the healthcare sector are influenced by factors such as regulation, change in professional power, competition among providers, and mutualistic behavior among different types of providers. The context in which hospitals operate has been described as having two essential components: the technical environment and the institutional environment. The technical aspects of an organization's task environment involve the production of products and services and the market forces that reward the efficient and effective production of those products and services. Institutional environments encompass the rules and requirements to which organizations must conform if they are to receive the social and political support and legitimization needed to survive.

Organizations in the health care sector have traditionally been subject to both strong institutional and strong technical pressures. These forces have operated in separate but parallel fashion to shape the practices and structures of hospitals (Alexander and Scott 1984). For example, hospitals traditionally maintained very loose control over physician staff as a reflection of strong professional requirements for autonomy. At the same time, they were accountable for the 'technical' quality of care through legal and quality assurance controls. Recently, market forces have played a more visible role in shaping national policy debate concerning healthcare access, quality, and costs. In view of this, some have argued that the influence of traditional institutional structures and practices, such as regulation, professional norms, and local control, have weakened and are now subordinated to the technical demands of the market. However, there is sustained evidence that historical institutional forces have not been eliminated, but instead are interacting with the new 'market' ethos to constrain and shape the structures and practices of hospitals, including those originating in the corporate sector. These interactions have given rise to a unique set of changes and practices that differ from both those common to the corporate sector and those common to hospitals in more institutionally dominant eras.

1. Institutional Environment of Hospitals

Institutional environments feature elaborate rules and requirements to which individual organizations must conform if they are to receive support and legitimacy. Attention is directed away from control and coordination of technical processes and towards conformity to externally defined requirements or regulations. The more institutionalized the environment of an

organization, the more its structures and procedures are governed by external controls. This control requires specific arrangements in the organization that are independent of contributions to technical performance (Scott 1995). For much of the twentieth century and up until today, hospitals and other healthcare organizations have been subject to strong institutional pressures, including a broad array of governmental regulations and requirements. Further, numerous professional specifications have governed what types of personnel might be hired, how tasks are distributed among them, and what procedures must be followed in performing these tasks.

The healthcare field in the US has long been characterized by institutional pressures that are both strong and heterogeneous (Stevens 1989). All branches of the federal government make decisions centrally that affect the entire healthcare sector. Local and state governments also exercise authority to regulate health care. This fragmented structure of authority makes it possible for various actors to pass laws to meet particular needs, and regulations are therefore uncoordinated, at best, and inconsistent or conflicting, at worst. Further, a long tradition of volunteerism in American hospitals has complemented the inculcation of professional norms into the institutional structure of these organizations. The notions of local control, community accountability, and philanthropic support of hospitals have served to keep decision-making from becoming overly centralized. At the same time, this has discouraged the formation of a cohesive, integrated healthcare system.

2. Technical Environment of Hospitals

Notwithstanding the institutional demands on hospitals, the technical requirements of modern medicine are considerable and often require tight internal controls and careful coordination if performance is to be effective. Hospitals have also increasingly been pressured to improve efficiency in the use of resources. Retrospective cost reimbursement was singled out as the primary explanation for both rising costs and inappropriate utilization. Critics also identified professional specialization and hospital-based care as major contributors to rising costs. Growing specialization and fragmentation in delivery rendered services increasingly impersonal, while the predominance of inpatient services threatened accessibility and provision of comprehensive, family-oriented care.

Historically, political attention had focused on expanding access to health care, as exemplified by the passage of legislation in the 1960s extending health insurance coverage to many of the elderly (Medicare) and the poor (Medicaid). The 1980s, however, marked the beginning of a new era that shifted the focus to controlling costs (Scott et al. 2000). Several years of

double-digit increases in prices threatened the economic interests of actors who paid for health care; in response they initiated policies and practices to control costs. In 1984, the federal government's Medicare program became the first major payer to attempt to control hospital costs by paying a fixed fee for each service, rather than the 'usual and reasonable' fees that hospitals routinely charged. This key change in payment policy had significant financial and market effects for hospitals. The resulting reductions in payments to hospitals and other organizational providers cut their budgets and contributed to cash flow problems. Hospitals and other providers tried to compensate by increasing their patient and service volume. In turn, this created more competition for patients.

A series of remedies have been proposed to deal with the cost escalation crisis. The solutions imposed have often clashed with the historical institutional functions of hospitals, namely, providing care to the community, running on a voluntary basis and engaging in philanthropic enterprise. Direct regulatory approaches to containing costs in the 1970s have shifted to greater emphasis on legal, budgetary, and market responses to addressing these issues. Taken as a whole, these remedies fall under the general rubric of a competitive approach to controlling costs, even though all involved substantial government oversight and occasionally intervention. Such approaches are consistent with a general societal shift toward privatization and market solutions to larger social problems. While recently court rulings have relaxed some regulations stifling collaboration among doctors and organizational providers, new regulations have reduced price-fixing and other forms of anticompetitive behavior. Together these rulings have provided both inducements and constraints on organizational behavior, such as severe limitations on hospital acquisitions and mergers.

Budgetary remedies have been applied at both the federal and state levels. At the federal level, the government has budgeted a fixed reimbursement for each inpatient admission based on diagnosis. States have also enacted rate-setting programs that prospectively set reimbursement levels for hospitals and other providers. These budgetary interventions have forced healthcare providers to attend to efficiency and have raised the level of competition among these providers. Finally, market remedies have included the introduction of so-called 'managed care' strategies, designed to reduce utilization and thereby costs (see *Managed Care*). Examples of these strategies have included competitive bidding and selective contracting in government programs and the use of preferred providers and risk-based financial incentives by private insurance firms. Both state and federal governments have begun to publish information on the performance of healthcare providers, aimed at redressing the traditional asymmetry of information between providers and consumers of health care.

As expected, these interventions have lowered healthcare costs and decreased utilization (Robinson 1996). But there have also been other, unanticipated effects. At the field level, the 1990s have witnessed an increasing number of entries and exits of new and existing organizational forms into the healthcare system. Although most hospitals and other organizational providers that have closed have tended to be primarily inner-city and rural institutions and have done so for primarily technical/rational reasons, many others have remained operating despite financial losses and technical inefficiencies. The survival of so many marginal institutions in the face of increasing pressures from a market-based system clearly reflects the importance of local/community support for these institutions, and the salience of legitimacy in organizational survival (Rohrer 1989). This illustrates well the complex combination of technical and institutional forces that shape organizational life in the healthcare sector.

3. Hospital Responses to Hybrid Environments

A fundamental question becomes, how do organizations operating within such hybrid environments accommodate different environmental demands that are often diametrically opposed to each other? The convergence of traditional institutional and new technical forces impinging upon hospitals has created new organizational, strategic, and survival challenges for healthcare providers. Neo-institutional theory suggests that organizations functioning in sectors that are highly developed both institutionally and technically will develop more complex and elaborate administrative systems and will experience higher levels of internal conflict (Scott 1995). Historically, we know this to be true for many organizations in the healthcare sector. Hospitals, for example, have typically operated within a dual hierarchy of control, represented on the one hand by hospital administration and on the other by hospital medical staff and physicians. These dual lines of control operate in parallel with each other but rarely have intersected in such a way as to disrupt the relative spheres of influence over issues related to control, resource allocation and professional prerogatives.

Conflict has been expressed in hospitals largely through increasing parity between managerial and professional logic and ideologies. The emphasis of managerial ideology on rational systems and cost-effective care has increasingly conflicted with the professional orientation toward individual doctor/patient relations and the micro delivery of specific medical services (Goodrick et al. 1997). Conflicting claims to the right way to deliver health care, the presence of alternative rationales for organizing work (professional vs. managerial), and the inherent conflicts between technical and institutional pressures in

the healthcare sector have eroded the legitimacy of hospitals. This has opened the way for a number of new organizational forms, practices, and strategies, particularly those that have symbolically represented the new ethos of privatization and corporate rationality. For example, health organizations are increasingly lobbying government and engaging in marketing campaigns in order to influence legislation and public opinion. Rather than being passive recipients of institutional changes in the broader environment, they seek actively to shape societal beliefs toward health care and the way health care is organized.

4. Organizational Heterogeneity

Health care delivery has been dominated traditionally by hierarchical organizations such as hospitals and healthcare systems. However, as competitive pressures and the demands of managing care under limited resources have increased over the past decade, a wide variety of new organizational forms have emerged (see *Health Care Delivery Services*). These range from managed care organizations, physician-hospital arrangements, and physician practice management companies, to diagnosis-specific service contracts with specialized providers, such as home health care and rehabilitation firms.

As traditional institutional norms and practices have eroded and have been replaced by a stronger corporate rationality, we have observed greater heterogeneity among organizations in the healthcare sector. This new ethos suggests there is no one best way to organize in order to achieve technical efficiency and positive performance in the marketplace. More variation in organizational forms is exhibited as hospitals become less concerned with mapping the characteristics of the institutional environment onto their own structures, and more concerned with achieving technical efficiencies and buffering their technical cores from environmental disturbances.

A clear response to the new market forces in health care has been the adoption of structures and practices from the corporate sector (Burns 1990). These structures include multidivisional organizational designs, parent holding-company arrangements, contract management of facilities, and vertical integration across a number of different service lines. While these practices clearly have gained popularity during the 1990s, there is limited evidence that they actually improve the quality or lower the cost of care. Neo-institutional theorists have argued that whereas early adopters of these corporate structures may have done so for technical/rational reasons, later adopters may have viewed them as the prevailing institutional practices under the new market-based health care environment.

5. Hospital Governance

Historically, most nonprofit hospitals were overseen by autonomous governing boards composed largely of representatives of the communities the hospitals served. This type of governance arrangement reflected a longstanding institutional belief that the control of nonprofit hospitals should be vested in local communities. However, during the 1980s and 1990s, two developments have altered this traditional governance model of nonprofit hospitals, both of which are part of a larger trend toward the so-called 'corporatization' of health care. One development is the consolidation of over 40 percent of US hospitals into complex health-care systems. In such systems, the locus of control for member hospitals is shifted to varying degrees outside local communities to system-level governing boards and corporate offices (Toomey and Toomey 1993). Another development is corporate restructuring of hospitals where hospital assets and decision-making authority are transferred to multiple corporate entities that may include a parent/holding company and various subsidiaries. Such restructuring often entails the creation of multiple governing bodies that can attenuate the decision-making authority of the provider's community-based board.

Under a traditional model of governance, nonprofit hospitals resemble consumer cooperatives in that community representatives are in a position to determine hospital policies. As such, a traditional form of governance is arguably a constraint on hospital price inflation, since community representatives have an incentive to resist price increases that they ultimately will incur in the form of higher insurance premiums. Also, form of governance may impact the accessibility of hospital services since a community-based board may be inclined to invest resources to expand the accessibility of hospital services to its own community (e.g., through the provision of uncompensated care). Policymakers have expressed concern that system membership and corporate restructuring are diminishing the role of communities in the control of nonprofit hospitals, and government agencies that have oversight responsibilities for nonprofit hospitals are currently examining alternatives (see *Health Care Organizations: For-profit and Non-profit*).

6. Physician–Organization Structures

As managed care and competitive pressures have compelled hospitals to consolidate and provide services along a broad continuum of care, the need to integrate formerly loosely coupled elements of these organizations has increased. Perhaps the most problematic aspect of such integration is the alignment of incentives and strategic goals between delivery organizations and physicians. Hospitals seek to engage in

strategic relations with physicians in order to: (a) protect against the risks of agency from physicians who have traditionally borne little risk for the care of patients, (b) reduce transaction costs by developing economies of scale and integrating physicians into the structure of the delivery organization, (c) achieve market power both up and down stream over suppliers and customers respectively, and (d) develop capabilities that differentiate systems from competitors in the areas of innovation, technology, and quality. While these market-driven reasons are clearly critical to hospitals and healthcare systems, the alignment of economic and strategic interests between organizations and physicians is difficult given the differences in cultural context, socialization, orientation to patient care, and economic incentives.

As managed care organizations and health care regulators have recognized the need to align the economic incentives of providers with the goals of cost containment and high quality, they have changed methods of provider compensation from simple discounting to risk sharing (e.g., capitation, *per diem*, or per case reimbursement for hospitals). Providers under these systems assume actuarial and utilization risk and reward. In principle, risk sharing promotes appropriate behavior, since high utilization and costs will result in deficits for either the physician group or its physicians. In practice, the physician group's form of compensation, incentives, and governance moderate these effects. For example, the incentives inherent in risk sharing can be lost if individual physician compensation is not strongly linked to holding costs below the group's fixed budget. Alternatively, if physicians feel they have no say in the formulation of incentives in compensation systems, they may be less willing to abide by them. These examples suggest that the incentives created by risk sharing are effective only under certain conditions. Several authors have recognized that the effects of such payment mechanisms may partially depend on organizational factors such as type of governance, physician leadership, information dissemination, sanctions, and organizational culture (Kralewski et al. 2000, Alexander et al. 2000).

Physicians and organizational healthcare providers recognize the need to work together to acquire managed care contracts and to manage risk. To facilitate such contracting, avoid institutionalized constraints inherent in hospital-medical staff relations, and develop the administrative infrastructure necessary for managing service delivery, many healthcare systems have established market-based organizational arrangements that link the delivery organization with one or more physician groups. Market-based arrangements are different from internal hospital structures in that they tend to be formally structured, contractual in character, corporate in organization, and include physicians outside the formal membership of the medical staff.

Given fundamental differences between methods of organizing by physicians and delivery organizations, however, it is highly questionable whether structural solutions to achieving greater integration between physicians and delivery organizations can be completely effective. There is only limited evidence to suggest that structural solutions work and considerably more evidence that self-selection by physicians into closely aligned structural arrangements may account for more of the explained variance in alignment than structure itself. Hospitals' attempts to integrate and align with physicians have not always rationalized the delivery of health care, but instead in many cases have made it more complex. Physicians and their interests vary widely, and hospitals have discovered that no one structural vehicle best accommodates the interests of all physician subgroups. Consequently, it is not uncommon for multiple market-based arrangements to exist within a single organization. This makes the coordination of care and alignment of incentives at least as difficult, if not more so, as it has been under the traditional medical staff model.

7. Organization of Work in Hospitals

Largely as a result of increasing competitive pressures and external emphases on cost containment, the organization of work within hospitals has shifted. Two trends are particularly in evidence: (a) the use of teams to perform managerial and clinical work, as opposed to performing such work individually, and (b) the use of contracts to obtain labor externally rather than to hire employees.

Because of demands by managed care firms and integrated healthcare systems to provide comprehensive care to patients and/or enrollees in health plans, there is growing recognition that patients need comprehensive care in order to address their multiple problems effectively. Given the broad spectrum of knowledge necessary for comprehensive care, there is growing belief that teams of practitioners with complementary roles and skills are best suited to providing high quality, cost-effective patient care (Fried and Rundall 1994). Teamwork in the clinical round is also perceived as an antidote to fragmentation in services, a traditional problem in healthcare delivery. For example, repeated patient assessments by various healthcare providers add unnecessary costs to patient care, when one combined assessment would suffice. Advocates of team-oriented care cite studies that consistently show that collaboration among physicians, nurses, and other healthcare personnel enhances the quality of care and patient satisfaction (Nikolaus et al. 1999, Le et al. 1998).

Team-based care stands in marked contrast to the traditional hierarchical work relations in most hospitals, which accord the physician a role of absolute

autonomy and control over patient-care processes. The growing emphasis on prevention, wellness, and cost-effectiveness in care delivery make this hierarchical structure outmoded. Flatter structures and more collegial relations among providers appear to be the developing norm. For example, many nurses advocate a collaborative relationship with physicians where nurses have primary responsibility for health maintenance, prevention, and diagnosis and treatment of routine cases, while physicians retain responsibility for the diagnosis and treatment of complex cases and the prescription of medications.

Contracting for labor services as an alternative to employment stems from the need for more flexibility in an increasingly competitive environment. Hospitals need to be able to make rapid changes in their workforce in response to demands from a volatile market. Rather than staffing to accommodate periods of peak demand with attendant down time for slack periods, hospitals have increasingly adopted a contingent approach to staffing, responding to the particular needs of the organization at a given time (Katzman 1999, Rizzuto et al. 1999). Contracting typically occurs in housekeeping, business, and clinical services. Proponents argue that contracting frees top managers to concentrate on core services, brings specialized expertise to the organization, improves flexibility through the ability to terminate contracts, reduces supply costs through superior purchasing power, and promotes the recruitment of specialists who may be in short supply. Critics argue, however, that the contract-management firm and the contracting organization may fail to clarify performance goals or that contracted staff may have difficulty working with other departments and full-time personnel. Perhaps most importantly, some believe that contracting for a wide variety of services without a clear strategy risks damaging organizational coherence and control (D'Aunno et al. 1996).

8. Quality of Care in Hospitals

Another illustration of the organizational response to technical and institutional pressures lies in the area of healthcare quality. On the one hand, decline in public support and trust for hospitals has reduced their impetus to assure quality and discipline members of their organizations for lapses in quality. Responsibility for quality has begun to shift to the payers of health care and their agents (e.g., managed care organizations), as well as to the courts which adjudicate an increasing number of malpractice suits against hospitals and physicians. These trends reflect a growing tendency to think of health care as a commodity rather than an entrusted service provided by professionals and professional organizations. On the other hand, there is greater emphasis placed on quality

improvement systems within organizations as a counterbalance to external controls over quality and as a response to market pressures to increase the efficiency and effectiveness with which care is delivered. Again, this duality of external control vs. internal control is a juxtaposition based in the competing demands placed on hospitals by institutional and technical forces.

Hospitals, in particular, are being challenged by increasingly competitive markets and regulatory changes to produce high quality care without raising costs. Given the multiple pressures and problems associated with healthcare quality, hospitals face the additional challenge of developing policies, procedures, and incentives to improve quality within a domain that has traditionally been the province of highly autonomous professionals. Quality improvement (QI) practices have emerged in recent years partially as an alternative to quality assurance and as a reaction to the shortcomings of outcome-based control. With mounting pressure to maintain or improve quality in the face of reduced resources, hospitals have begun to look to process improvements as a mechanism to provide high quality care while containing healthcare costs. While the adoption of QI appears to be an internally generated response to changing market conditions, for the most part, accrediting bodies such as the Joint Commission on Accreditation of Healthcare Organizations (JCAHO) have also encouraged hospitals to adopt QI.

Given the traditional emphasis on provider-patient relations, individualized approaches to improving quality have received the vast majority of attention to date. Although the knowledge and practices of individual physicians are important for high quality care, increasingly clinicians practice within groups or systems of care. As such, the quality of care delivered depends critically on the ability of these groups or systems to prevent errors, coordinate care among settings and practitioners, and ensure that relevant, accurate information is available when needed.

One of the most promising recent developments is the effort to institutionalize quality improvement in hospitals as a coordinated, organization-wide approach to care strategies (Berwick et al. 1990, Gaucher and Coffey 1993, Carman et al. 1996). These practices, collectively, are designed to systematically examine and improve processes of care and care support in organizational settings. Typically, these quality improvement efforts are strongly rooted in evidence-based procedures and rely extensively on data collected from hospitals about care processes and patient outcomes.

9. Hospital Behavior in Other Nations

The focus of this article has been to associate change in hospital structure and behavior to change in the broader market and institutional contexts in which

these organizations operate. The United States has been prominently featured in this discussion because of the somewhat unusual combination of strong technical and institutional pressures to which hospitals must respond. Other countries, however, may diverge markedly in the degree to which either institutional or market forces play a role in shaping hospital behavior. For example, in countries where healthcare delivery is highly centralized and funded exclusively by the government, we would expect to find much more structural homogeneity among hospitals, in marked contrast to the increasing diversity and differentiation among US hospitals. In other nations, where privatization and multiple sources of healthcare funding predominate, hospitals are more likely to resemble in their structure and behavior other commercial vendors of services to the market.

Recent trends in OECD countries would seem to indicate that they are facing technical pressures for increased efficiency similar to those in the United States (Normand 1993). Significant reductions both in the number of hospital beds per 1,000 and in average length of hospital stay have resulted, particularly in Scandinavian countries. In the countries of the former Soviet Union, on the other hand, average length of stay has not dropped, despite mounting financial strain due to economic instability. One potential explanation for this is lingering institutional norms regarding appropriate clinical practice (Hensher et al. 1999). This may illustrate how specific institutional beliefs about health and health care in a given nation may shape hospital behavior in ways that differ substantively from the ways in which institutional forces shape hospital behavior in the US. For example, in Taiwan, the deeply seated belief that family is the primary care giver is reflected in the hospital practices of incorporating family members in patient education, extended visiting hours, and even boarding with the patient during the hospital stay. Cross-national differences in hospital structure and behavior are thus likely to be a reflection of both the structure and content of the broader environment from which hospitals obtain patients, critical resources, and legitimacy.

See also: Doctor-Patient Interaction: Psychosocial Aspects; Gender and Health Care; Health Care Delivery Services; Health Care: Legal Issues; Health Care Markets: Theory and Practice; Health Care, Rationing of; Health Care Systems: Comparative; Health Care Technology; Health Professionals, Allied; Medical Profession, The

Bibliography

Alexander J A, Scott W R 1984 The impact of regulation on the administrative structure of hospitals. *Hospital and Health Service Administration* **29**: 71-85

- Alexander J A, Vaughn T, Burns L 2000 Organizational, strategic and marketing determinants of operating policies of physician organization arrangements. *Health Services Management Research* **13**: 1–15
- Berwick D M, Godfrey A B, Reosner J 1990 *Curing Health Care: New Strategies for Quality Improvement*. Jossey-Bass, San Francisco, CA
- Burns L R 1990 The transformation of the American hospital: From community institution toward business enterprise. *Comparative Social Research* **12**: 77–112
- Carman, J M, Shortell S M, Foster R W, Hughes E F, Boerstler H, O'Brien J L, O'Conner E J 1996 Keys for successful implementation of total quality management in hospitals. *Health Care Management Review* **21**(1): 48–60
- D'Aunno T, Alexander J A, Laughlin C 1996 Business as usual? Changes in health care's workforce and organization of work. *Hospital and Health Service Administration* **41**: 3–18
- Fennell M, Alexander J 1993 Perspectives on organizational change in the US medical care sector. *Annual Review of Sociology* **19**: 89–112
- Fried B, Rundall T 1994 Managing groups and teams. In: Shortell S M, Kaluzny A D (eds.) *Health Care Management: Organization Design and Behavior*, 3rd edn. Delmar Publishers, Albany, NY
- Gamm L D 1992 Health care markets as interorganizational fields: A conceptual perspective. *Health Services Management Research* **5**(1): 44–53
- Gaucher E J, Coffey R J 1993 *Total Quality in Healthcare: From Theory to Practice*. Jossey-Bass, San Francisco, CA
- Goodrick E, Meindl J R, Flood A B 1997 Business as usual: The adoption of managerial ideology by U.S. hospitals. *Research in the Sociology of Health Care* **14**: 27–50
- Hensher M, Edwards N, Stokes R 1999 International trends in the provision and utilization of hospital care. *British Medical Journal* **319**: 845–8
- Katzman C N 1999 Outsourcing keeps growing. Contract management survey finds times are good for firms serving healthcare industry. *Modern Healthcare* **29**(35): 42, 44, 46–50
- Kralewski J E, Rich E C, Feldman R, Dowd B E, Bernhardt T, Johnson C, Gold W 2000 The effects of medical group practice and physician payment methods on costs of care. *Health Services Research* **35**(3): 591–613
- Le C T, Winter T D, Boyd K J, Ackerson L, Hurley L B 1998 Experience with a managed care approach to HIV infection: Effectiveness of an interdisciplinary team. *American Journal of Managed Care* **4**(5): 647–57
- Nikolaus T, Specht-Leible N, Bach M, Oster P, Schlierf G 1999 A randomized trial of comprehensive geriatric assessment and home intervention in the care of hospitalized patients. *Age and Ageing* **28**(6): 543–50
- Normand C 1993 Comparisons among national health care systems in the European marketplace. *Hospital Formulary* **28**(Supp.1): 6–11
- Rizzuto G T, Yallapragada R R, Roe C W, Authement J P 1999 Contract labor use in healthcare. *Hospital Topics* **77**(4): 15–19
- Robinson J C 1996 Decline in hospital utilization and cost inflation under managed care in California. *Journal of the American Medical Association* **276**(13): 1060–4
- Rohrer J E 1989 Closing rural hospitals: Reducing 'institutional bias' or denial of access? *Journal of Public Health Policy* **10**(3): 353–8
- Scott W R 1995 *Institutions and Organizations*. Sage, Thousand Oaks, CA
- Scott W R 1998 *Organizations: Rational, Natural and Open Systems*, 4th edn. Prentice-Hall, Upper Saddle River, NJ
- Scott W R, Ruef M, Mendel P J, Caronna C A (eds.) 2000 *Institutional Change and Healthcare Organizations: From Professional Dominance to Managed Care*. University of Chicago Press, Chicago
- Starr P 1982 *The Social Transformation of American Medicine*. Basic Books, New York
- Stevens R 1989 *In Sickness and in Wealth: American Hospitals in the Twentieth Century*. Basic Books, New York
- Toomey R K, Toomey R E 1993 The role of governing boards in multihospital systems. *Health Care Management Review* **18**(1): 21–30

J. A. Alexander and M. E. Comfort

Hotelling, Harold (1895–1973)

Harold Hotelling (1895–1973) was one of a handful of statistical researchers who in the first half of the twentieth century developed the solid intellectual foundations of the then newly developing discipline of statistics. In addition, he was also an important contributor to several areas of mathematical economics. His academic career includes: professorships at Stanford University, Columbia University, and the University of North Carolina at Chapel Hill, and his professional leadership involved major statistical and economic societies. He is perhaps best known for his statistical work in multivariate analysis, introducing the concepts of principal components, canonical correlations and Hotelling's- T^2 . In mathematical economics, he contributed to natural resource economics, spatial economic competition, and introduced the welfare equilibrium proposition.

1. Early Life

Born on September 29, 1895 in Fulda, Minnesota, Harold Hotelling, whose parents were Clair Hotelling and Lucy Rawson, was the oldest of six children. Having moved with his family to Seattle at age nine, Hotelling studied journalism at the University of Washington, taking time from his studies to work on a weekly paper. In 1919, he received his BA in Journalism, graduating Phi Beta Kappa. Encouraged, in part, by the mathematician and well-known historian, Eric Temple Bell, he went on to graduate study in mathematics at the University of Washington, receiving an MA in 1921. It was during this time that he married Floy Tracy in 1920. He received his Ph.D. in Mathematics in 1924 from Princeton University, doing dissertation research on dynamic topology with Oswald Veblen, although his original intent was to study mathematical economics. His interest in economics took root while growing up and in his belief in

‘the importance ... of a study of economics [in] that [it] might possibly impress some people to the point of making some changes in the institutions with which they had become familiar [and] which I [Hotelling] was already prepared to condemn vigorously.’ Alternatively, he ‘couldn’t escape the lure of mathematics’ and notes that he ‘hoped and [it] appeared that the mathematics would be extremely prolific of new economic theorems and ideas if followed out’ (Hotelling 1963).

2. Beginning at Stanford University

Invited to join the Food Research Institute at Stanford University, Hotelling was a research associate there from 1924 to 1927, following which he was an associate professor in Stanford’s Department of Mathematics (1927–31). An early indication of the breadth of his intellectual interests is illustrated by the fact that in 1927 he taught courses on mathematical statistics, differential geometry, and analysis situs (a forerunner of topology), and in that year published papers on the application of the latter to statistics, and on differential equations subject to error. While working in the Food Research Institute, he was introduced to the work of R. A. Fisher (five years his elder), and through their mutual correspondence they became friends. In fact, Hotelling wrote the book review in the *Journal of the American Statistical Association* (Hotelling 1927) of the first edition of Fisher’s *Statistical Methods for Research Workers* (and also did subsequent reviews there of the Second through Seventh Edns.). After the planned visit of Fisher to Stanford in the summer of 1928 was abandoned, Hotelling visited Fisher at the Rothamsted Experimental Station during the second half of 1929. Their connections continued and at the outbreak of World War II, Hotelling, along with three others, wrote to Fisher volunteering to care for Fishers’ six daughters in the United States during the war. However, a last minute cancellation kept the Fisher children in Britain (Box 1978, p. 377).

While at Stanford, Hotelling wrote a number of path-breaking papers in statistics and in mathematical economics. The most notable among these statistical results was his generalization of Student’s t -ratio to handle multivariate response data (Hotelling 1931). This generalization remains known as Hotelling T^2 . His paper with Holbrook Working (Working and Hotelling 1929) on modeling and linear prediction first demonstrated that the standard error of the estimate of a linear regression depended on the independent variable’s value, thereby giving rise to the well-known Working–Hotelling curved simultaneous confidence regions for a linear regression. His research contributions to mathematical economics were equally notable. His 1929 paper provided a game-theoretic solution given by two differential equations to a problem of price and location competition between

two entities in a spatial setting. His 1931 paper on the economics of exhaustible resources is recognized as the seminal paper in the field of natural resource economics. In that paper, Hotelling demonstrates that in an equilibrium setting, prices of exhaustible natural resources will tend to rise over time at a percentage rate equaling the interest rate.

3. Columbia University and the Statistical Research Group

Hotelling moved to Columbia University in 1931 as a Professor of Economics, having been recruited to further the development of mathematical economics, and also to initiate mathematical statistics. Throughout his career at Columbia and later at the University of North Carolina at Chapel Hill, Hotelling was superlative at both attracting outstanding researchers and mentoring young researchers destined for stardom. The Nobel-prize winning economist, Milton Friedman, was among the earliest mentored by Hotelling, spending 1933–4 at Columbia learning mathematical economics from him. Later mentoring Kenneth Arrow, another Nobel-prize winning economist, Hotelling was also instrumental in attracting Abraham Wald, W. Allen Wallis, and Jacob Wolfowitz to the statistics facility at Columbia. As a teacher, Hotelling was perceived to be organized and comprehensive, interweaving theory with application. Importantly, an appreciation of research was imparted to his students, and he was able at an interpersonal level to instill in his students their own confidence to be successful at research (Madow 1960).

Hotelling was one of three original voting fellows of the Institute of Mathematical Statistics on September 12, 1935, and was President of that Institute in 1941. He began his association with the Cowles Commission, participating in its first Summer conference in 1935, and served as President of the Econometric Society during 1936–7. In 1939–40, he was a Visiting Lecturer at the Indian Statistical Institute and was President of the Indian Statistical Congress (1940).

Along with Wallis and Wolfowitz, Hotelling was a charter member of the Statistical Research Group (SRG), at Columbia, for which Wallis was appointed Director of Research. The SRG, founded on July 1, 1942, and in existence until September 30, 1945, was an extraordinary research group of statisticians whose goals were to provide research support and analyses for the armed forces in furthering the quality and accuracy of their war efforts. The origins of sequential analysis stem from Wald’s work for the SRG, and among Hotelling’s contributions were techniques for control charts for multivariate data (Kruskal 1980). (A detailed history of the SRG is provided by Wallis 1980).

Early in Hotelling’s career at Columbia, his wife, Floy, died in 1932, leaving behind two small children.

In 1934, Hotelling married Susanna Edmundson and together they had six children, five sons and a daughter (who died in her infancy).

Hotelling, during his prewar years at Columbia, continued to produce ground-breaking research in statistics and economics. His 1933 and 1936 papers introduced and lay the groundwork for principal component analysis as a multivariate statistical technique for simplifying the structure of large numbers of correlated measures (Hotelling 1933, 1936a). In 1936, he extended notions of multiple correlation to canonical correlations in order to measure and study the relationships between two dependent sets of correlated responses (Hotelling 1936b). In the field of economics, his 1935 paper provided a consistent theory uniting demand and utility by means of a constrained optimization problem and his presidential address to the Econometric Society in 1938 (Hotelling 1938) introduced a welfare equilibrium proposition which leads in certain societal settings to the argument that goods be priced at their marginal costs. In 1940, his innovative and insightful paper on the teaching of statistics in universities was published (Hotelling 1940). In this prescient paper, Hotelling observes the inevitability of the increased demand for statisticians, and argues that the teaching of statistics in universities would be best done solely by statisticians in separate departments of statistics (a quite novel idea in 1940!). With regard to this article, Jerzy Neyman (1960) later wrote ‘in my opinion the establishment of the excellent departments of statistics in such universities as Chicago, Columbia, Harvard, North Carolina, and Stanford is due to a considerable extent to the publication of Hotelling’s article.’

4. Career at University of North Carolina

After the war in 1946, Hotelling was recruited by Frank Graham, President of the University of North Carolina at Chapel Hill to start the Department of Mathematical Statistics there upon his arrival. In addition to being appointed professor of mathematical statistics and professor of economics, he was named Associate Director of the Institute of Statistics (directed by Gertrude Cox), which was constituted with Chapel Hill’s department and the Department of Experimental Statistics at North Carolina State, as well as related groups in psychometrics, social sciences and biostatistics. It is notable that when Hotelling was planning to go to Chapel Hill, he proposed that Wald join him there. In order to entice Wald to remain at Columbia, a Department of Mathematical Statistics was promptly proposed and officially approved in December 1946 (Hoxie 1955). As at Columbia, Hotelling attracted outstanding statisticians to the Chapel Hill faculty including R. C. Bose, Wassily Hoeffding, P. L. Hsu (who had been at Columbia with Hotelling), William Madow, Herbert Robbins, S. N. Roy, and Walter Smith. Hotelling relinquished the chairman-

ship to George Nicholson in 1952, and devoted the remainder of his career there to his research and his students.

His research at North Carolina involved, on the whole, elaborations and in-depth further study of some of his earlier research. Noteworthy among these is his work in 1947 and 1951 on hypothesis tests related to multivariate analysis of variance, which lead to the testing criterion referred to as the Lawley–Hotelling trace. He formally retired in 1966, but continued to be active in the department. He was elected to the National Academy of Sciences in 1970, and received a number of other academic honors during this period. Subsequent to a serious stroke in mid-1972, Harold Hotelling died in Chapel Hill, North Carolina on December 26, 1973.

5. Hotelling’s Contribution to Multivariate Statistical Analysis

One of the origins of statistical hypothesis testing is a test that the mean μ of a normal distribution has a specified value μ_0 . The resultant Student’s t -test, obtained in 1908, remains as one of the most used statistical procedures. The t -statistic is

$$t = \sqrt{N}(\bar{x} - \mu_0)/s$$

where $\bar{x}$ and s , respectively, represent the mean and standard deviation of a sample of size N . Hotelling recognized that many experiments will have correlated measures $x_1, \dots, x_p$, which arise from a multivariate normal distribution, and he provided the Hotelling’s T^2 test to handle the means of multiple measurements (Hotelling 1931). For a sample of size N let $\bar{x} = (\bar{x}_1, \dots, \bar{x}_p)$ denote the sample mean vector and $S = (s_{ij})$ represent the sample covariance matrix.

The T^2 -statistic is

$$T^2 = N(\bar{x} - \mu_0)' S^{-1}(\bar{x} - \mu_0)$$

where $\mu_0 = (\mu_{10}, \dots, \mu_{p0})$ is the specified mean vector in the hypothesis. When the hypothesis is true, $(N-p)T^2/p(N-1)$ has an F -distribution with p and $N-p$ degrees of freedom.

Analogous to Student’s t -test, the T^2 -statistic can also be used to test for the equality of the mean vectors of two normal distributions that have the same unknown covariance matrix but unequal means, μ and ν . Now the statistic becomes

$$T^2 = \frac{N_1 N_2}{N_1 + N_2} (\bar{x} - \bar{y})' S^{-1}(\bar{x} - \bar{y})$$

where $\bar{x}$ and $\bar{y}$ are mean vectors from samples of size N_1 and N_2 , respectively, and S represents the pooled sample covariance matrix. Although there are many parameters in this model, under the null hypothesis the

distribution of $(N_1 + N_2 - p - 1)T^2/p$ has an F -distribution with p and $N_1 + N_2 - p - 1$ d.f.

The T^2 -statistic additionally arises as a linear discriminant, as noted by Fisher, and is the sample analog of the Mahalanobis squared distance.

Quite remarkably, Hotelling noted that for the two-sample problem, the T^2 statistic is invariant under the transformation $\bar{x} \rightarrow \bar{x}A$, $\bar{y} \rightarrow \bar{y}A$, where A is any nonsingular matrix. This anticipated a general theory of statistical invariance developed much later. In particular, the T^2 -test has a number of optimal properties among invariant tests: (a) the T^2 -statistic is the only invariant function of the sufficient statistics $\bar{x}$ and S ; (b) as a test statistic it is the uniformly most powerful test among all invariant tests; and (c) the noncentral distribution depends on the scalar parameter $(\mu - \nu)\Sigma^{-1}(\mu - \nu)'$, which is the only invariant of the parameters.

Large data sets either in exploratory data analysis, data mining, or model building are accompanied by a large number of variables. The reduction of numbers of variables is often a first step in trying to uncover patterns in the data. Because multivariate measures $x_1, \dots, x_p$ are related, there is the need to disentangle the correlational structure. To accomplish this, Hotelling (1933, 1936a) introduced the p principal components $z_1, \dots, z_p$, which are linear combinations of the underlying measures, namely $z_k = \sum_{i=1}^p a_{ki}x_i$, $k = 1, \dots, p$, where the coefficients are normalized by $\sum_{i=1}^p a_{ki}^2 = 1$. The first set of coefficients $a_{11}, \dots, a_{1p}$ is chosen so as to maximize the variance of z_1 . The second set of coefficients $a_{21}, \dots, a_{2p}$ is chosen to maximize the variance of z_2 subject to the constraint that z_1 and z_2 are uncorrelated. At each step the variance is maximized, but with the additional constraints that each z_k be uncorrelated with $z_1, \dots, z_{k-1}$. Their ordered variances $\text{Var}(z_1) \geq \dots \geq \text{Var}(z_p)$ turn out to be the characteristic roots of the population covariance matrix.

In effect, principal components provide transformed coordinates that are uncorrelated, thereby disentangling the measures. Because the principal components are ordered according to their variances, this permits the researcher to focus on a subset of principal components that contributes substantially to the total variance. In multiple regression, the predictor variables often are closely related, and the problem of multicollinearity results. Principal components is one important procedure that introduces orthogonality and reduces dimensionality, thus removing multicollinearity.

To build a hierarchy of correlational measures one can start with the ordinary Pearson product moment correlation between x and y . With multiple measures $x_1, \dots, x_p$ and y one generates the multiple correlation of y and the linear combination $y^* \equiv \sum a_i x_i$, which is the maximum correlation between y and y^* . Suppose now there are multiple measures $x_1, \dots, x_p$ and $y_1, \dots, y_q$. To analyze the relation between these two

sets of variates, Hotelling (1936b) defined canonical correlations between u_k and v_k , which are the correlations of linear combinations $u_k = \sum a_{kj}x_j$ and $v_k = \sum b_{kj}y_j$ where $k = 1, 2, \dots, K \equiv \min(p, q)$. The coefficients are normalized so that the variances of u_k and v_k are 1, and are chosen so as to maximize the correlations, and such that $u_1, \dots, u_k$ are mutually uncorrelated and $v_1, \dots, v_k$ are mutually uncorrelated $k = 1, \dots, K$. The new variables u_k and v_k , $k = 1, \dots, k$ are called the canonical variables; furthermore, it turns out that u_i and $v_j (i \neq j)$ are uncorrelated. The discovery of the corresponding matrix factorization is quite remarkable. A special case of this factorization leads to the singular value decomposition, which is a very important representation in numerical linear algebra.

As a result of this decomposition, one can transform the original variables to new coordinates $(u_1, v_1), \dots, (u_K, v_K)$, that exhibit a simple structure: (a) the pairs are uncorrelated; (b) the pairs are ordered decreasingly in accordance to their correlations (which are non-negative); that is, (u_1, v_1) are maximally correlated; and (c) the variances of the u_k 's and v_k 's are normalized to be one. In a 1935 paper, Hotelling had earlier focused on the first pair (u_1, v_1) , and he called u_1 the most predictable criterion and v_1 the best predictor.

Principal components and canonical correlations can be viewed in terms of roots of a determinantal equation and estimated from their sample analogues. It was this context that opened an entire area of statistical research on the distribution of the roots of random determinantal equations. In 1939, a number of papers appeared simultaneously in the statistical literature in which the central distribution of the roots were obtained. From then until the present, there is a continuing flow of theory concerning the noncentral distributions of roots of sample determinantal equations.

Hotelling's training as a mathematician provided him with a warehouse of tools that he applied to statistical problems. Among his first papers was a presentation in 1926 to the American Mathematical Society on the use of topological principles in the derivation of distributions. Many of his later results were obtained with geometrical arguments. He was able to exploit asymptotic theory to obtain normal approximations and limiting results. Hotelling, in 1930, was the first to attempt to state precise conditions for the consistency and asymptotic normality of the maximum likelihood estimator. With Lester Frankel in 1938, he obtained expansions for the distribution of Student's t -statistic, thereby leading to a useful normal approximation.

With his interest in relations between variables, Hotelling recognized that the required normality assumptions may be unrealistic. This led him to study the nonparametric version of the product moment correlation. Together with Margaret Pabst, Hotelling

studied the Spearman rank correlation coefficient and they showed that for large samples it approximately follows a normal distribution. In 1953, Hotelling continued his analysis of correlations and obtained expansions of Fisher's z -transformation of Pearson's product moment correlation. This work remains among the definitive results in this area.

Hotelling wrote on matrix calculations and developed methods that still stand. His work on weighing designs led to a number of combinatorial generalizations. Although Hotelling had a mathematical focus for his research, he was not averse to working on applied problems. The range of his applied papers include results concerning the physical state of protoplasm, duration of pregnancy, causes of birth rate fluctuations, anthropomorphic measures of Southwest Indians, and effects of pregnancy on dental caries.

Harold Hotelling's originality and imagination is seen in both the genius of his research that produced innovative methods that stimulated whole areas of modern statistics and in his prescient leadership in the discipline of statistics. Further details concerning Hotelling, including his complete bibliography, can be found in Olkin *et al.* (1960), in Pfouts (1960), and in three articles appearing as 'Three papers in honor of Harold Hotelling at 65' in *The American Statistician* (14), published in 1960 (pp. 15–25).

See also: Fisher, Ronald A (1890–1962); Multivariate Analysis: Overview; Statistics, History of

Bibliography

- Box J F 1978 *R. A. Fisher: The Life of a Scientist*. Wiley, New York
- Hotelling H 1927 Review of 'Statistical methods for research workers.' *Journal of the American Statistical Association* **22**: 411–12
- Hotelling H 1931 The generalization of Student's ratio. *Annals of Mathematical Statistics* **2**: 360–78
- Hotelling H 1933 Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24**: 417–41 and 498–520
- Hotelling H 1936a Simplified calculation of principal components. *Psychometrika* **1**: 27–35
- Hotelling H 1936b Relations between two sets of variates. *Biometrika* **27**: 321–77
- Hotelling H 1938 The general welfare in relation to problems of taxation and of railway and utility rates. (Presidential address to the Econometric Society). *Econometrica* **6**: 242–69
- Hotelling H 1940 The teaching of statistics. *Annals of Mathematical Statistics* **11**: 457–70
- Hotelling H 1963 'Remarks by Harold Hotelling at a Seminar in which Honorary Degree Recipients were asked to give brief intellectual biographies.' As transcribed from a tape recording — not corrected or edited. University of Rochester. Private Communication
- Hoxie R G 1955 The Department of Mathematical Statistics. In: *A History of the Faculty of Political Science*,

Columbia University. Columbia University Press, New York, pp. 250–5

- Kruskal W H 1980 Comment: First interaction with Harold Hotelling; testing the Norden bombsight. *Journal of the American Statistical Association* **75**: 331–33
- Madow W G 1960 Harold Hotelling as a teacher. *American Statistician* **14**: 15–17
- Neyman J 1960 Harold Hotelling: A leader in mathematical statistics. In: Olkin I, Ghurye S G, Hoeffding W, Madow W G, Mann H B (eds.) *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press, Stanford, CA, pp. 6–10
- Olkin I, Ghurye S G, Hoeffding W, Madow W G, Mann H B 1960 *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press, Stanford, CA
- Pfouts R W 1960 *Essays in Economics and Econometrics: A Volume in Honor of Harold Hotelling*. University of North Carolina Press, Chapel Hill, NC
- Wallis W A 1980 The Statistical Research Group, 1942–45. *Journal of the American Statistical Association* **75**: 320–30
- Working H, Hotelling H 1929 Applications of the theory of error to the interpretation of trends. *Journal of the American Statistical Association* **24**(March supp.): 73–85

I. Olkin and A. R. Sampson

Household Archaeology

1. Introduction: The Emergence of Household Archaeology

Household archaeology emerged in the late 1970s and early 1980s as a late development of social archaeology on both sides of the Atlantic. The aim of household archaeology is to create a context in which a humanized reconstruction of the past may be nurtured, through the study of intrasettlement relations. Such an investigation at the microscale involves the intensive recording of archaeological observations and detailed excavation of an area broad enough to include architectural and other features of the built environment. The criticism of such an effort is that it produces an isolated picture of one part of a social group—perhaps only one household—that does not provide a statistically valid extrapolation to the rest of a settlement, or from one settlement to the regional scale.

Household archaeology developed both as an extension of and a reaction, by both processual ('new') and Marxist-oriented archaeologists to the social evolutionary features of social archaeology. Social archaeology provided data and inferences for explanations of change and variation of past cultural systems, focusing on intersettlement relations, especially in terms of the interaction of political leaders on a

regional basis within the parameters of ecological features such as distance and resource distribution. Marxist-oriented studies added the dimensions of ideology and power relations in production and distribution.

A criticism of the macroscale models of social change has been that, in identifying the longer-term evolutionary trajectory of, for example, resource control, hierarchical social relations, and population growth, and in establishing the general social form of a group, the context of shorter term changes and variability is given a lower priority in explanation. In this way a dehumanization of the past has been created in order to 'write history (or prehistory) large.'

Household archaeology developed in the 1980s and 1990s as the result of a growing interest in intra-settlement relations, but it is a sub-discipline that is neither unified, organized, nor recognized within archaeology. The aims and agendas of the archaeologists who carry out studies that are termed 'household archaeology' are as variable as are the theoretical strands of archaeology itself.

2. *Household Archaeology: Archaeology at a Small Scale*

When it was initially developed in the late 1970s the study of households aimed at filling in the most detailed level of settlement pattern and activity analysis, which could then be extrapolated to make more general statements about demographic trends, specialized production, class structures and so on (Flannery and Winter 1976). Later, in the important volumes edited by Wilk and Rathje and Wilk and Ashmore, the importance of households as 'fundamental elements of human society' was emphasized (Ashmore and Wilk 1988, p. 1, Wilk and Rathje 1982). At this time it was suggested that, since households are the level at which social groups articulate directly with economic and ecological processes, their study would offer a chance for archaeologists to examine social adaptation (in a functionalist sense) with direct reference to the empirical details of the archaeological record. It was assumed this would be done within the methodological framework of scientific logical positivism. In other words, household analysis would allow archaeologists to 'bridge the existing 'mid-level theory gap' in archaeology' (Wilk and Rathje 1982, p. 617). Thus household archaeology developed alongside the empirical theory-building in archaeology, known also as middle range research, that lies at the heart of the history of processual archaeology at the end of the twentieth century.

It is clear that household archaeology in the 1980s implied examining and analyzing social change at a microscale as a complement to analyses of social and economic change at a broader scale of regions and

continents. Household archaeology served to provide a richness of detail, but was not an end in itself. The ultimate target was to understand social evolutionary change.

3. *Sources of Inspiration for Household Archaeology*

The explicit inspiration for household archaeology comprised the studies of households and multiscalar histories written by historians (including historians of the so-called Annales school) (Braudel 1972), and by anthropologists, especially those interested in historical demography and historical anthropology (Arnould and Netting 1982, Laslett and Wall 1972, Netting et al. 1984), (Hammel and Laslett 1974, p. 76). It is the latter who coined the term 'co-resident domestic group,' as a way to bring together the enormous diversity of the minimal social group—family, household, houseful and to use a term that was not subject to 'cultural bias' and would have value for 'objective scientific analysis.' The "'co-resident domestic group'" consists of those who share the same physical space for the purpose of eating, sleeping and taking rest and leisure, growing up, childrearing and procreating' (Hammel and Laslett 1974, p. 76).

Archaeologists envied the historians and social anthropologists their knowledge about the kinship relations and the demographic composition of the members of a domestic group that allowed them to explore the 'co-resident domestic group' in terms of variability in their size and kinship composition, and the cooperative activities of their members. Traditional scientific archaeological methodology puts limitations on the investigation of the social composition of or relations within the domestic group. In view of the dangers inherent in using architectural features or any other archaeological data to identify a 'household' on a prehistoric site, Wilk and Rathje pointed out that it was more productive for archaeologists to investigate questions such as what co-residents of a domestic group (households) do, rather than what their social form is (who lives there and how they are related) (Wilk and Rathje 1982). Thus, for example, it was argued that to understand transformations in the social relations of production in Neolithic Southeast Europe, it was not essential to be able to identify 'households' in the historical and anthropological sense of the term (Tringham 1990, p. 603). It was sufficient to be able to investigate changes in co-residence and the cooperative activities of kin-based co-resident domestic groups.

It is not surprising that the impetus for household archaeology came from archaeologists who had focused on the excavation and interpretation of well-preserved architectural remains. Architecture—more broadly, the built environment—provides the essential

context of co-resident domestic groups. Archaeologists working in prehistoric Mesoamerica led the field (Ashmore and Wilk 1988, Wilk and Rathje 1982). Simultaneous and parallel developments were occurring in the Southwest US with archaeologists working with spectacular Pueblo architecture, such as Grasshopper Pueblo (Ciolek-Torrello 1986), and in Europe and the Near East with archaeologists working in Neolithic architectural contexts (Bogucki 1988, Tringham 1990, Tringham 1991). The Chacmool conference on Household and Community provided a welcome synthesis of the diverse routes to household archaeology in place by 1989 (MacEachern et al. 1989).

4. *Modeling Household Activities in Archaeology*

In modeling the economic and social role of co-resident domestic groups, participants of 'household archaeology' have followed social anthropologists in their models of the domestic mode of production (Chayanov 1966 (reprinted 1986), Netting et al. 1984, Sahlins 1972). These models characterized self-sustaining agriculturists and horticulturists who live in permanent villages and carry out relatively small-scale production. Such studies have emphasized the importance of the function of households as the organizing units of cooperative labor, production, consumption and sharing of resources, the transmission of property and rights from one generation to another, and the creation and maintenance of ties and alliances with other units through marriage and other forms of exchange. These same essential features of social reproduction became the objects of investigation of household archaeology (Ashmore and Wilk 1988, Wilk and Rathje 1982).

The study of 'what households do' has been enriched by the Marxist focus on ideology, power and social inequality. In the domestic mode of production the pattern of domination and inequality between and within different households of a settlement and between different settlements, although temporary and flexible, is thought to comprise a crucial variable in stimulating and enabling larger-scale transformations such as the intensification of production, the increase of the labor supply, and the increase in cultural complexity (Meillassoux 1978).

Inequalities certainly exist, along with innumerable variations, between households and are the result of differences in the cycles of demographic and economic evolution through which households pass—cycles in which they expand, accumulating wealth, power, property, and members, and then contract. Thus, at any one time the households of a settlement will demonstrate differences in membership, the composition of the labor force, and the structure of dominance within the co-resident group and within

the settlement as a whole. They will demonstrate differences in relations with groups outside the village, and, most importantly, differences in access to the products and processes of production. This concept of small-scale social inequality is the basis of a number of archaeological models that use the emerging importance of households as the unit of social reproduction to explain larger social and economic transformations (Rathje and McGuire 1982, Tringham 1990).

These differential patterns of social and economic relations reflecting social inequalities, however, have such a low archaeological visibility that these societies have traditionally been regarded as 'egalitarian.' The difficulty in recognizing inequality in 'egalitarian' societies is due to the fact that it is ephemeral and not cumulative. An additional significant factor, however, is that inequality in such societies is not based on such criteria as the ownership of land and other material property, which are easily definable or easily recognizable in the archaeological data. Control of labor—not land, capital or property ownership—is the crucial variable on which inequality is based in precapitalist self-sustaining societies. Thus, those who control the circulation of labor and its products, control the social reproduction of society. At the scale of the domestic unit or household, this may be an individual within the household, or it may be 'senior' households who control the labor of 'junior' households. It has been argued that inequality at this scale creates tensions in numerous spheres and is a crucial variable in the process of socioeconomic transformation that can be observed in the archaeological record (Tringham 1990, p. 608).

5. *Archaeological Evidence of Household Inequality*

The primary requirement for the examination of the social context of production and economic life at an archaeological site (in other words, of household archaeology) is that a number of domestic architectural units be exposed by excavation. This will enable a comparative view of co-resident domestic groups across the settlement at any one time. This information is enhanced by the provision of such a picture through time. Only when a large number of domestic structures is exposed is it possible to tell whether, for example, all the structures in a given excavated occupation horizon represent residences or parts of residences; whether the differentiated units of co-residence can also be differentiated as units of cooperative production; whether all the domestic structures were occupied at the same time; whether they all became part of the archaeological record at the same time or whether they were destroyed or abandoned at different times, and, finally whether they appear to be in a parallel stage of their use-lives at the

time of destruction or abandonment or whether some seem to have been more recently established than others. The study of archaeological architecture in terms of the use-life of the buildings has greatly enhanced these efforts for the purposes of household archaeology (Tringham 1994). In this way the context and nature of the construction, maintenance, occupation, modification, and abandonment/destruction, and replacement of a residence all become objects of knowledge. Traditionally, they would have comprised elements that were subsumed under the formal features of the building. The investigation of a building through different stages of its use-life, brings to light many important aspects that would otherwise have been ignored; for example, modifications such as the addition or removal of rooms during the cycle of growth and decline of a domestic unit; efforts to prolong the use-life of a building by repairs and renewals; symbolic emphasis placed on particular use-life stages such as re-plastering, the destruction of old houses (for example by fire), or founding of new houses (Tringham 1994).

Inequalities within and between domestic units are manifested archaeologically by the spatial differentiation of the production process (that is, in the production, consumption, and distribution of food and nonfood resources). In each case the pattern of economic activity (and inequality) should have a distinctive appearance, demonstrating differential access to the materials, means, process, and products of production. The expectation is that different households will show differential access not only to the products of exchange with other groups but also to the raw materials, techniques, and equipment for making use of resources, such as food items, minerals, and ceramics. Such a differentiation of economic activity is frequently, but not necessarily, accompanied by specific styles of material elaboration, such as ceramic decoration.

6. Trends in Household Archaeology

For many years household archaeology remained marginal to the mainstream of processual (new) archaeology. In his review of Ashmore and Wilk's edited volume on *Household Archaeology in Mesoamerica*, David Freidel noted that for the interest in and impact of 'household archaeology' to be sustained, it would need to be incorporated into a more ambitious theoretical program in archaeology (Freidel 1989). Its incorporation into Marxist models of power and social inequality certainly helped in this respect. The theoretical basis that has contributed most to sustaining interest in household archaeology, however, has been that of feminist anthropology and theories of social practice (Bourdieu 1977, Moore 1988). A common theme of these diverse post-structural theories has been their development beyond

neo-Marxian models of power and labor control to models of how ideologies of gender, class, labor, production, identity, etc., are constructed, maintained and transformed. They focus on domestic arenas and the contexts of daily practice as being the most significant places where these negotiations between social agents take place. During the later 1990s these models have also come to characterize the many faces of post-processual archaeology, one of which is an archaeology of households that has transformed into an archaeology of gendered places (Tringham 1994).

In the archaeology of gendered places, the archaeological record is not only analyzed and observed at a microscale, but also interpreted at a microscale. It challenges the search for the universal category of 'kin-based co-residential domestic group' and embraces the richness of the diversity of the social context of domestic action that has been forefronted in social anthropology for many years (Moore 1988). It lifts the household domain out of its previous position of 'assumed general knowledge' and as marginal to the 'great events' of history. It provides a richly-based, multiscalar theoretical framework to consider the role of the labor of players in prehistory: women, men and children (Tringham 1991).

The theoretical basis of feminist anthropology and social practice theory obviously creates some serious epistemological challenges to the scientific logical positivist methodology that characterizes mainstream processual (new) archaeology. It is as difficult to identify 'gender' in archaeological architecture as it is to find 'households' and 'families.' Responses to this challenge have been articulated in many publications (Hodder 1999, Wylie 1992). They are based in a celebration of the ambiguity of archaeological data, a critical awareness of the scientific process of inference, and a richly researched utilization of sources for interpretation—ethnography, history, art and literature. Above all in the post-processual epistemological framework, archaeology is identified as an interpretive discipline rather than a hypothetico-deductive search for true (pre)history. Thus in order to think about the prehistoric construction and practice of gender ideologies and dominance relations within or between households, a researcher may interpret the archaeological data and their meaning in many ways. It is not required to identify gender or 'domestic unit' in the archaeological record.

The database for an archaeology of gendered places remains especially but not exclusively in the record of the built environment. Along with the nature of household archaeology itself, however, the excavation and interpretation of archaeological architecture has changed. According to post-processual concepts, the built environment can be thought of as both a passive container that has a use-life and reflects social action and as an 'active' arena or medium for and symbolic expression of social actions. Thus social actions, practices, negotiations, and dominance ideologies of

whoever passes through a building are constrained and channeled by the materiality of that building. (Tringham 1994). In practice this has led to the necessity of retrieving a highly detailed microstratigraphic record of the traces of human activity, in which the sequence of every depositional event is recorded and interpreted. It is time-consuming, expensive, but immensely rewarding (Hodder 1999, Tringham 1994).

Along similar lines, household archaeology in the late 1990s has embraced the ambiguities recognized by practice theory anthropologists about kinship categories, including 'family' and 'household.' In their view, it is the meaning and cultural value ascribed to a relationship that are significant rather than some set of biological ties (Gillespie 2000). The enactment of kinship often involves the negotiation of identity and the identification of relationships through particular places on the landscape—buildings and sets of material culture (Joyce 2000).

At the beginning of the twenty-first century, 'what households do' is still a focus of household archaeology, but the 'doing' is far more complex than physical action. Moreover, a 'co-resident domestic group' is no longer a faceless economic category, but comprises the complex life-histories of many people as well as buildings and places. Finally, household analysis is no longer a step along the ultimate goal of understanding social evolutionary change. It plays an essential role in the creation of a multiscalar history that considers such questions as how—in practice—was a sense of place created; how was social memory about this group maintained; what were the myths about garbage; what would the walls of a house say if only they could speak (Tringham 2000).

See also: Ancestors, Anthropology of; Anthropology, History of; Ethnoarchaeology; Family and Kinship, History of; Gender, Feminism, and Sexuality in Archaeological Studies; Historical Demography; Horticultural Societies; Kinship in Anthropology; Marxist Archaeology

Bibliography

- Arnould E J, Netting R M 1982 Households: Changing form and function. *Current Anthropology* **23**: 571–6
- Ashmore W, Wilk R R 1988 House and household in the Mesoamerican past: An introduction. In: Wilk R R, Ashmore W (eds.) *Household and Community in the Mesoamerican Past*. University of New Mexico Press, Albuquerque, NM, pp. 1–28
- Bogucki P 1988 *Forest Farmers and Stockherders: Early Agriculture and its Consequences in North-Central Europe*. New Studies in Archaeology. Cambridge University Press, Cambridge, UK
- Bourdieu P 1977 *Outline of a Theory of Practice*. [trans. Nice R]. Cambridge University Press, Cambridge, UK
- Braudel F 1972 *The Mediterranean and the Mediterranean World in the Age of Philip II*. [trans. Reynolds S]. Harper and Row, New York
- Chayanov A V 1966 (reprinted 1986) *The Theory of Peasant Economy*. University of Wisconsin Press, IL
- Ciolek-Torrello R 1986 Room function and household at Grasshopper Pueblo. In: Benson C, Upham S (eds.) *Mogollon Variability*. 107–19. New Mexico State University, The University Museum. Occasional Papers 15, Las Cruces, NM
- Flannery K V 1976 Analyzing household activities. In: Flannery K V (eds.) *The Early Mesoamerican Village*. Academic Press, New York
- Freidel D 1989 Household and community in the Mesoamerican past. *Science* 19 May: 863–4
- Gillespie S 2000 Beyond kinship: an introduction. In: Gillespie S D, Joyce R A (eds.) *Beyond Kinship: Social and Material Reproduction in House Societies*. University of Pennsylvania Press Philadelphia, PA, pp. 1–21
- Hammel E A, Laslett P 1974 Comparing household structure over time and between cultures. *Comparative Studies in Society and History* **16**: 73–109
- Hodder I 1999 *The Archaeological Process: An Introduction*. Blackwell, Oxford, UK
- Joyce R 2000 Heirlooms and houses: Materiality and social memory. In: Gillespie S, Joyce R (eds.) *Beyond Kinship: Social and Material Reproduction in House Societies*. University of Pennsylvania Press, Philadelphia, PA, pp. 189–212
- Laslett P, Wall R (eds.) 1972 *Household and Family in Past Time*. Cambridge University Press, Cambridge, UK
- MacEachern S, Archer D J W, Garvin R D (eds.) 1989 *Households and Communities*. The 21st Annual Conference: Chacmool Calgary, AB
- Meillassoux C 1978 'The economy' in agricultural self-sustaining societies: A preliminary analysis. In: Seddon D (ed.) [trans. Lackner H]. *The Relations of Production: Marxist Approaches in Economic Anthropology*. Frank Cass, London, pp. 127–57
- Moore H L 1988 *Feminism and Anthropology*. University of Minnesota Press, Minneapolis, MN
- Netting R, McC, Wilk R R, Arnould E J (eds.) 1984 *Households: Comparative and Historical Studies of the Domestic Group*. University of California Press, Berkeley, CA
- Rathje W L, McGuire R H 1982 Rich Men ... Poor Men. *American Behavioral Scientist* **25**(6): 705–16
- Sahlins M 1972 *Stone Age Economics*. Aldine Chicago
- Tringham R 1990 Conclusion: Selevac in the wider context of European prehistory. In: Tringham R, Krstic D (eds.) *Selevac: a Neolithic Village in Yugoslavia*. Monumenta Archaeologica No. 15. UCLA Institute of Archaeology Press, Los Angeles, pp. 567–616
- Tringham R 1991 Households with faces: The challenge of gender in prehistoric architectural remains. In: Gero J M, Conkey M W (eds.) *Engendering Archaeology: Women and Prehistory*. Basil Blackwell, Oxford, UK, pp. 93–131
- Tringham R 1994 Engendered places in prehistory. *Gender, Place, and Culture* **1**: 169–203
- Tringham R 2000 The continuous house: A view from the deep past. In: Gillespie S, Joyce R (eds.) *Beyond Kinship: Social and Material Reproduction in House Societies*. University of Pennsylvania Press, Philadelphia, PA
- Wilk R R, Rathje W L 1982 Household archaeology. *American Behavioral Scientist* **25**(6): 617–40
- Wylie A 1992 The interplay of evidential constraints and political interests: recent archaeological research on gender. *American Antiquity* **57**: 15–35

R. Tringham

Household in Anthropology

Anthropologists use the term household loosely to refer to a residential group whose members cooperate in some activities of production, consumption, or child rearing. They have distinguished the household from the family by defining the former as a residential unit and the latter as a kinship unit. Close scrutiny of this distinction reveals key assumptions that have shaped the way anthropologists and other social scientists have conceptualized and studied not only households, but *families*, *kinship*, and *domestic groups*.

1. Problems in Defining the Household

The grounds for analytically separating families and households lie in the broader claim that kinship and locality are two distinct principles of organization (Keesing 1958). This claim, in turn, derives from the discipline's focus in the nineteenth and early twentieth centuries on the principles organizing social groups and relations in nonstate societies. Kinship, which was assumed to be universally defined on the basis of genealogical linkages rooted in biological reproduction, was considered the most important among these principles. As a key mode of recruiting members to (and limiting membership in) politically important groups such lineages and clans, kinship (genealogical propinquity) was distinguished from *locality* (spatial propinquity), which defined residential units such as hamlets and villages. As anthropologists expanded their range of ethnographic research to include peasant communities, they drew on the distinction between kinship and locality to differentiate family from household. The family as a corporate kin group was distinguished from the household as a collection of kin and sometimes nonkin who share a common residence. The corporate character of the peasant family was viewed as rooted in its jural rights to property, usually land, which its members held in common. Family members supposedly retained these rights regardless of whether they were members of the same household. Hence, family and household, it was argued, are both 'logically distinct' and 'empirically different' (Bender 1967, p. 493).

Actual usage of the term household in the anthropological literature, however, calls into question both the logical distinction and the empirical difference between families and households. For one, residential units composed of individuals who are not genealogically linked, such as orphanages, college dormitories, men's houses (in societies where gender segregation among adults is practiced), convents, and army barracks, usually are not considered households. At the same time, however, families whose members move between residential units that are located on different continents are referred to as 'transnational households' (Basch et al. 1994). In other words,

anthropologists attach the term household only to groups of people sharing both a common living space and activities that are identified as *domestic*. *Domestic activities* include those activities related to food production and consumption or to reproduction, including childbearing and child rearing. As these are the very same activities that are considered the functional basis of the 'family,' in practice, households are identified using both 'kinship' and 'residence' criteria. Thus, the claim for the logical distinction and empirical difference between family and household is not affirmed by scholarly practice.

Because 'domestic' activities are sometimes carried out by groups of people who do not reside together, some anthropologists (Bender 1967) have suggested that it would be better to employ alternate terms in our ethnographic descriptions. Among the alternatives that have been suggested are 'co-residential group,' 'budget unit,' and 'domestic group.' Less important than the terms employed, however, is the need for ethnographers to describe precisely the composition, organization, activities, and functions of these units. In addition, the cultural meanings people attach to these units need to be explicated. In the absence of this knowledge, anthropologist and others are likely to project their own cultural meanings and expectations on them.

The difficulties of identifying households are illustrated by innumerable cases (see Segalen 1984) which demonstrate that neither the boundaries nor the functions of household are easily discerned. Questions have been raised as to how to treat residential groups that move through a seasonal cycle of dispersal and concentration, how to handle the movement of people between dwelling units, whether to define as a single household the huts or houses that share a common yard or those that share a common stove, and whether to include servants, apprentices, boarders, lodgers, and nonkin in the household. More recently, studies of transnational migration (Rouse 1991, Basch et al. 1994) have raised the question as to whether residential units located in different nations and even on different continents constitute a single household. Studies of gay and lesbian kinship (Weston 1991), on the other hand, have challenged the assumption that heterosexual marriage and parenthood are the universal, natural core of families and households.

2. Problems in the Classification of Households Types

Much of the literature on households has been concerned with explaining variations among households that are distributed over time and space. In describing this variation, most scholars have focused on the character of the genealogical links between members of the household as its defining feature. Terms such as 'stem family household,' 'nuclear family

household,' and 'patrilateral extended household' are based on the configuration of genealogically-defined kin types included in the household—in other words, its *household composition*. In addition, the most genealogically 'close' members of the household are assumed to form its structural core. For example, if a household contains two adult brothers, their wives and children, it is labeled a fraternal-joint household. The assumption here is that the relationship between the brothers constitutes the structural core that shapes the relations of all the members of the household. Accordingly, studies of household change over time track changes in the genealogical composition of the household.

The classification of households on the basis of their genealogical composition rests on the shaky assumption that people who are similarly linked genealogically will have the same or greatly similar social relations. In other words, it is assumed that the labeling of a household as a 'stem-family household' captures its organizational structure and function. But a stem-family household found at one time and place does not necessarily share the same organizational dynamics as a stem-family household at another time and place. A household, or any social unit, is much more than the sum of its genealogical links. It is the negotiated outcome of the culturally meaningful actions and relations of its members. The actions and relations of household members with identical genealogically defined links can differ radically. Even households with the same genealogical composition in the same community may have significantly different structures of authority and divisions of labor (Yanagisako 1984).

3. What Explains Variation among Households?

A key question that has fueled comparative studies of households by anthropologists, historians, demographers, and other social scientists is what explains variations among them. Two related questions have been asked of variations observed in household composition. First, why do a significant proportion of the households in a society differ from what people report to be the ideal type of household? Second, irrespective of these ideals, what explains the distribution of household types within and between societies? Three types of variables have been offered to answer both these questions: demographic variables, economic variables, and social stratification variables.

3.1 Demographic Variables

Demographic variables affect the size and composition of households, thus acting as constraints on the attainment of the culturally ideal household type. Age at marriage, life expectancy, and fertility shape the composition of households at any given time in a population. The economy of the household is also shaped by demographic processes. Among peasants

for whom the household is the unit of production, changes in the demographic structure of the household as it moves through the developmental cycle entail changes in the ratio of consumers to workers (Chayanov 1966).

Fortes's (1958) model of the developmental cycle of domestic groups draws on these demographic processes to explain the varieties of domestic groups found in a society. Fortes hypothesized that the variety of household types found in a community or society at any point in time are merely different phases in the developmental cycle of a 'single general form.' As all households pass through different stages of the developmental cycle at different times, a census taken at a particular point in time will catch only a few in the ideal, complete phase. Hence, an ideal household type such as the patrilineal stem family household of Japan (which includes a married couple, their adult son and his wife and children) is better regarded as 'a transitory state in the development of the household, rather than as a form' (Hammel 1975, p. 142).

Fortes's model has been useful in helping us understand the impact of events such as birth, marriage, and death on the composition of households. It cannot account for all the observed variation in households, however, because other cultural and social forces (see below) are also involved. Fortes himself later (1978, p. 18) conceded that a model of a uniform developmental cycle could not explain why the actual histories of different families entail different developmental sequences. The model may also blind us to historical shifts wherein more recently formed households embark on a different trajectory than older ones. In addition, while demographic factors undoubtedly shape households, it is a mistake to treat them as exogenous, biological constraints on the composition and economy of households. To do so confuses social replacement with biological reproduction and overlooks the strategies people actively employ to control household size and composition, including their 'kinship' practices of marriage, fosterage, adoption, timing and spacing of births, and timing of family division. The collection of longitudinal data on households, whether by utilizing historical sources or through retrospective oral histories, is necessary for understanding both continuity and change. Cohort analysis is an effective tool for helping ascertain similarities and differences in the household histories of members of successive cohorts, defined on the basis of year of birth, marriage, or other historically significant events experienced by the members of a community.

3.2 Economic Variables

Economic variables are also adduced to explain cross-societal variation in household composition and organization. In some cases a mix of 'economic' and

'ecological' factors, including labor needs of production, defensive needs, care of children, taxation, and conscription practices of the state, are identified as the functional basis for the prevalence of certain household types. In others, the labor requirements of a particular productive activity are singled out as the primary determinant of household type. In their cross-societal analysis of 60 randomly selected societies, Pasternak et al. (1976) conclude that labor requirements are the most powerful determinant of the formation of extended family households. They hypothesize that in societies where work outside the home makes it difficult for a mother to tend to children and other domestic tasks or when the outside activities of the father make it difficult for him to perform his subsistence work, extended family households will prevail. Although they are able to demonstrate a correlation between activity requirements and presence of extended family households, they lack diachronic data to support the hypothesized causal relation between their independent variable (activity requirements) and their dependent variable (presence of extended family households). Consequently, they must resort to arguing that it is more plausible that activity requirements cause extended family households rather than vice versa. However, it is even more plausible that household composition and productive activities do not lie in a simple, unidirectional causal relation but rather shape each other.

Attempts like that of Pasternak et al. to identify the determinants of household composition and structure through statistical analysis of cross-societal comparisons have declined significantly over the past 30 years as the inherent problems and limitations of such studies have been increasingly recognized. Among these problems is that of discerning whether the units that have been selected for comparison are appropriate. As all societies contain different frequencies of a range of household types, it is misleading to classify entire societies into those with 'extended family households' vs. those with 'independent family households.' A better understanding of the relation between household composition and 'economic' factors may be gained by investigating the ways in which diverse types of households in a society articulate with one another, rather than by reducing the entire society to a single household type.

Labor requirements are often used to explain the frequency with which a particular household type appears in a society or section of society. Netting (1974) cites several ecological studies as demonstrating that household composition among horticulturalists and agriculturalists varies with the type and amount of labor required for effective crop production. His own study attributes the differences between the large, extended families of shifting cultivators and the nuclear family households of intensive agriculturalists in the same Nigerian plateau environment to different labor requirements (Netting 1965).

A shaky assumption underlies both cross-societal and intrasocietal studies that seek the determinants of household variation in productive labor requirements. In all of them, it is assumed that once having identified the subsistence activity or primary productive activity of the household, researchers can ascertain the labor requirements of that activity and proceed to show how the household efficiently meets them. In doing so, however, they overlook the fact that it is not simply technological requirements that necessitate a household type, but the entire manner in which production is socially organized. In other words, a productive activity can be accomplished by a variety of household types depending on how that production is socially organized, including the division of labor by sex and age, the use of hired labor, cooperation between households in productive tasks, and the exchange of labor between households. In assuming that the household is a self-contained labor unit, researchers fail to sufficiently investigate the existence of other forms of labor organization. But the social units engaged in production may not be the same at different phases of the productive cycle. For example, different social units may work together in the planting, weeding, and harvesting of crops. To attribute household composition to labor requirements begs the question, because it assumes the prior existence of an entire social organization of production. In order to overcome this narrow focus on technical requirements of production, the interplay of a wider range of cultural, political, and economic forces shaping productive relations within and between households needs to be considered.

3.3 Social Stratification

Differential access to resources, in particular differences in ownership and control of the means of production, also shape household variation in both peasant communities and among urban, industrial populations. A past tendency to neglect inequality in studies of households and other domestic groups can be attributed to the assumption that the communities studied by ethnographers were basically homogeneous. When anthropologists have studied peasant communities, for example, the implicit contrast with the landowners—who were generally not studied—led them to overlook significant differences among the peasants. Consequently, there has been a tendency to dismiss stratification as a significant factor in the analysis of household variation among peasants. Yet peasant communities are nowhere homogeneous, but are always internally differentiated. Differences in wealth among peasants are also overlooked because they are often obscured by idioms of kinship that bind the landed to the landless and the land-rich to the land-poor. Significant wealth differences may exist, however, between the smaller co-residential units of a single kin-based compound or between brothers who head

the main and branch households of a family. In other words, ideas and ideologies of family and kinship may obscure stratification among kin.

4. *Changes in Households over Time*

Anthropological studies of household and family change over time have focused generally on the effects of industrialization, urbanization, and the increasing integration of communities into capitalist market relations. Many of these studies challenge or argue for the refinement of the prevailing sociological theory that family, household, and kinship decline in importance as a society 'modernizes' and becomes increasingly urbanized and industrialized. They demonstrate that in some cases urbanization and industrialization have strengthened the extended family rather than weakened it and that new functions are assumed by extended families and extended family households. For example, urban migration may contribute to the maintenance of extended family households. While these studies provide an important correction to the theory of modernization and kinship decline, they are marred by the inconsistent use of ambiguous terms, including 'extended family household.' The argument over whether extended families have declined with urban-industrialization is in part the result of differences in the way in which researchers define and conceptualize household type. Quite often it is assumed that people who live in extended family households pursue common strategies and cooperative actions under the direction of a central authority—the head of the household. This not only gives undue weight to centralized authority as the defining feature of the extended family, but it also overlooks the heterogeneity in the goals and strategies of the members of a household.

5. *Recent Critical Approaches to Household*

The adaptationist, utilitarian approach to households that prevailed in much of the anthropological literature up until the 1980s has been challenged by new theoretical approaches that have had an important impact on the discipline in general. Among these perhaps the most significant has been feminist theories of gender.

In the 1970s, one of the first questions feminist anthropologists raised in rethinking the anthropological analysis of women's status and roles was whether ethnographers had overlooked women's goals, strategies, and forms of action, especially in the 'domestic sphere.' M. Wolf's (1972) study of Taiwanese joint family households, for example, revealed that family and household division was brought about as much by women's attempts to advance their own goals and those of their 'uterine family' as it was by conflicts among brothers. Collier (1974) argued that this process occurs commonly in

societies where men gain power by holding together large bodies of co-resident kin, but where women (especially younger women) gain power by breaking up such units. By attending to the informal power women often wield in societies where men are granted formal authority, feminist anthropologists alerted us to the tensions and conflicts and, therefore, the complex processes shaping households and other domestic groups. Likewise, investigations of women's relations with kin and others outside their household demonstrated that it is not just men who link mother-child units to larger institutional structures. The involvement of women in informal exchange networks between families and households had been overlooked by androcentric studies that had equated male relations with the 'public sphere.' Once studies showed that the female and male members of households—as well as the young and the old, those born in the household and those who had married into it—often do not share the same goals and strategies, it was no longer possible to retain the naive concept of a unified 'household strategy.' Finally, by calling into question the presumed universality and immutability of gender identities and the gender division of labor, feminist anthropologists denaturalized gender systems and called for the study of its historical and cultural production. They revealed the gender assumptions embedded in concepts of the 'public' and 'private' spheres (Reiter 1975), in definitions of the 'domestic sphere' and 'domestic groups,' including family and household (Yanagisako 1979) and in theories of kinship (Yanagisako and Collier 1987).

6. *Rethinking Households*

Functionalist analyses of 'household adaptation' reified the household, treating it as a social actor endowed with human consciousness and motives. Households were described as having goals, fashioning strategies, and adapting to the changing pressures of a modernizing society. Such language demonstrates the ease with which an analytic construct can be made into a social actor. The imputation of consciousness and will to households is rooted in the assumption that those who live together share a collective motive and will—in other words, a unified household strategy. It obscures the complex social processes through which people negotiate the relations that produce households, despite different and sometimes conflicting goals, strategies, and ideas about what it means to 'live together.'

Neither the composition nor the dynamic structure of households can be understood without a thorough understanding of the cultural meanings and sentiments that shape the social actions and relations of the people who live in them. This does not mean that households cannot be compared across societies, but rather that the comparative study of households must include an explication of the culturally specific and

heterogeneous meanings and motives shaping peoples' residential preferences and relations. Such an expanded strategy of research and analysis will enable anthropologists to move beyond crude comparisons of distributions of genealogically defined household types within and across societies to understanding households as the negotiated outcomes of culturally meaningful action.

See also: Collective Beliefs: Sociological Explanation; Community, Social Contexts of; Community/Society: History of the Concept; Community Studies: Anthropological; Everyday Life, Anthropology of; Family and Kinship, History of; Family, Anthropology of; Family as Institution; Home in Anthropology; Household Production; Kinship in Anthropology; Marriage; Matrilocality; Place in Geography; Residential Segregation: Geographic Aspects; Social Stratification

Bibliography

- Basch L, Schiller N G, Blanc C S 1994 *Nations Unbound: Transnational Projects, Postcolonial Predicaments, and Deterritorialized Nation-States*. Gordon and Breach, Langhorne, PA
- Bender D R 1967 A refinement of the concept of household: Families, co-residence, and domestic functions. *American Anthropologist* 69: 493–504
- Chayanov A V 1966 *The Theory of Peasant Economy*. R D Irwin, Homewood, IL
- Collier J F 1974 Women in politics. In: Rosaldo M Z, Lamphere L (eds.) *Women, Culture, and Society*. Stanford University Press, Stanford, CA
- Fortes M 1958 Introduction. In: Goody J R (ed.) *The Developmental Cycle of Domestic Groups*. Cambridge University, Cambridge, UK
- Fortes M 1978 An anthropologist's apprenticeship. *Annual Review of Anthropology* 7: 1–30
- Hammel E A 1975 Reflections on the zadruga. *Ethnology* 11: 141–51
- Keesing F M 1958 *Cultural Anthropology: the Science of Custom*. Rinehart, New York
- Netting R M 1965 Household organization and intensive agriculture: The Kofyar case. *Africa* 35: 422–9
- Netting R M 1974 Agrarian ecology. *Annual Review of Anthropology* 3: 21–56
- Pasternak B, Ember C R, Ember M 1976 On the conditions favoring extended family households. *Journal of Anthropological Research* 32: 109–23
- Reiter R R 1975 Men and women in the south of France: Public and private domains. In: Reiter R R (ed.) *Towards an Anthropology of Women*. Monthly Review Press, New York
- Rouse R 1991 Mexican migration and the social space of postmodernity. *Diaspora* 1: 8–23
- Segalen M 1984 Nuclear is not independent: Organization of the household in the Pays Bigouden Sud in the nineteenth and twentieth centuries. In: Netting R McC, Wilk R R, Arnold E J (eds.) *Households: Comparative and Historical Studies of the Domestic Group*. University of California Press, Berkeley and Los Angeles
- Weston K 1991 *Families We Choose: Lesbians, Gays, Kinship*. Columbia University Press, New York

Wolf M 1972 *Women and the Family in Rural Taiwan*. Stanford University Press, Stanford, CA

Yanagisako S J 1979 Family and household: The analysis of domestic groups. *Annual Review of Anthropology* 8: 161–205

Yanagisako S J 1984 Explicating residence: A cultural analysis of changing households among Japanese-Americans. In: Netting R McC, Wilk R R, Arnold E J (eds.) *Households: Comparative and Historical Studies of the Domestic Group*. University of California Press, Berkeley and Los Angeles

Yanagisako S J, Collier J F 1987 Toward a unified analysis of gender and kinship. In: Collier J F, Yanagisako S J (eds.) *Gender and Kinship: Essays Toward a Unified Analysis*. Stanford University Press, Stanford, CA

S. J. Yanagisako

Household Production

Household production is the production of goods and services by the members of a household for their own consumption, using their own capital and their own unpaid labor. Goods and services produced by households for their own use include accommodation, meals, clean clothes, and childcare.

The process of household production involves the transformation of purchased intermediate commodities (e.g., supermarket groceries and power-utility electricity) into *final consumption commodities* (meals and clean clothes). Households use their own capital (kitchen equipment, tables and chairs, kitchen and dining room space) and their own labor (hours spent in shopping, cooking, laundry, and ironing). The total economic value added by households in household production has been aptly named Gross Household Product (GHP) (Eisner 1989, Ironmonger 1996).

1. The Use of Labor and Capital in Household Production

In the language of economics, labor and capital are the two factors of production. However, labor, the time and effort provided by household members, is really the use of human capital; the second factor of production, capital, is the use of physical or tangible nonhuman capital (the services provided by land, dwellings, vehicles, and equipment). Economics conceptualises the production process as the use of capital (human and other) together with energy, to transform raw materials and 'unfinished' commodities (intermediate inputs) into finished goods and services ready for use by people (final consumption).

As countries industrialize, a large part of household production of food, clothing, furniture, and housing is transferred to business organizations and then purchased by households. Nevertheless, even in a world

apparently dominated by the market, a large amount of household production is necessary. Households need to add further value to put purchased commodities in the possession of ultimate consumers in the right place and at the right time.

2. Household Production in Contrast to Market and Subsistence Production

2.1 Market Production

While the market economy produces goods and services not produced by households (television sets and battleships), in many cases the market and the household are in direct competition, (restaurant meals vs. home prepared meals, and taxis vs. self-driven owned cars). Restaurants and taxis usually require immediate payment, if not by cash, then by credit card. In households, users do not pay each other by cash or credit card for meals served or transport provided.

Four separate modes of production are possible depending whether the household or the market provides the capital or the labor. Household production occurs when the household provides both its own capital and its own labor. However, the market can provide either one or both of these two factors of production. For example, a housekeeper could be employed to prepare meals, clean the house, and do the laundry. This paid housekeeper's time would be the use of 'market labor' in conjunction with 'household capital'—the kitchen, dwelling, and household equipment. Again, a household could rent a dwelling and a vehicle (market capital) but do its own cleaning, cooking, laundry, and driving (household labor).

In many countries, the use of paid domestic servants has almost disappeared; servants have been replaced by the household's own labor combined with more and better household capital and equipment. Many recent developments, such as supermarkets and automatic banking machines, involve a mode of production with households providing the required unpaid labor in lieu of the paid labor of shop assistants or bank tellers.

2.2 Subsistence Production

In developing countries many millions of households use their own capital and unpaid labor by way of fishing, collecting wood and water, growing vegetables and other food, building shelter, and making clothing. Thus, subsistence production fits the definition of household production when the goods produced are used within the household that produced them.

However, the wood, water, food, fiber, and clothing outputs from subsistence production could be sold on

the market to other households or to business enterprises. When such commodities are in fact sold, the value added in their production conceptually becomes part of market production.

3. Household Production and the Household Economy

The household economy describes the collective economic activities of households. Often the household economy is called the household sector as distinct from the business, government, and foreign sectors. However, the household sector is large enough to deserve the term household economy. The rest of the economy can then be called the market economy. Thus, the transactions between the household and the market are perhaps more akin to international trade between two economies than transactions between different industrial sectors of a single economy. The two major types of inter-economy trade are the sale of labor time by the household and the sale of household goods and services by the market.

3.1 Households as Producers

With few exceptions, economic textbooks focus on households as consumers and fail to discuss households as producers using their own labor and capital.

One of the earliest writers about the process of household production was Charlotte Perkins Gilman (*Women and Economics*, 1898) who questioned the traditional gender division of labor. She proposed moving even more household production to the market. This would yield gains from greater specialization and economies of scale and enable women to choose their work on the basis of inclination and talent.

Margaret Reid (*Economics of Household Production*, 1934) played a significant role in the development of household economics as a discipline, particularly for curricula in some North American universities. She held that, although the household is our most important economic institution, the interest of economists was concentrated on 'that part of the economic system which is organized on a price basis.' Reid is regarded as the first writer to specify the often-used third person criterion to distinguish between productive and nonproductive (consumption) activities. Her test was: 'If an activity is of such character that it might be delegated to a paid worker, then that activity shall be deemed productive' (Reid 1934). Thus, preparing a meal is productive work but eating it is nonproductive consumption or leisure. Of course eating and other consumption activities are productive in the sense that they are necessary for the continuation and re-invigoration of life. Another problem with this criterion is that student educational activities are

classed as unproductive since you would not learn to play the piano by paying someone to take the lessons for you.

A further problem is that production almost always requires a contribution from capital as well as from labor. The criterion needs to be extended to include services that might be obtained from rented capital as well as from a paid worker. Thus, accommodation services provided by a household's own dwelling are production; so are the transport services provided by a household's own vehicles and the entertainment services provided by a household's own sound and vision equipment. This extends the range of household production considered by Reid in 1934 and mostly since then by other writers. The expanded market alternative criterion would be 'An activity shall be deemed productive if it is of such a character that it might be obtained by hiring a worker or by renting capital equipment from the market.'

Gilman and Reid had little impact on mainstream economics. Economic theory continued to portray households as places only of consumption and leisure, with production of goods and services occurring only in business or public enterprises.

3.2 *The New Household Economics*

In the mid-1960s a major theoretical development took place, known as the 'new household economics' (Becker 1981, Ironmonger 1972, Lancaster 1971). In this theory the household is regarded as a productive sector with household activities modeled as a series of industries.

In this new approach, households produce commodities that are designed to satisfy separate wants such as thirst, hunger, warmth, and shelter. The characteristics, or want-satisfying qualities, of the commodities used and produced can be regarded as defining the production and consumption technology of households. With changes in incomes and prices, households still alter expenditures as in the earlier theory. However, in the new theory, households adjust their behaviour as they discover new commodities and their usefulness in household production processes.

The activities approach derived from the theory of the new household economics readily combines with the earlier input-output approach of Leontief (1941) to establish a series of household input-output tables as the framework for modeling household production.

4. *Measuring the Household Economy*

The measurement of the household economy emerged as a focal point for many researchers once the household was recognized as a major centre of production, not just consumption.

4.1 *Importance of Measuring the Household Economy*

In the 1970s a number of writers drew attention to the macroeconomic magnitude of the household economy. Boulding (1972) conceptualised households not only as the main driving force for the market economy, with household purchases covering about 60 percent of GNP, but also as the most important agent in the grants economy. This is the economy of one-way transfers, grants, or gifts, given mainly within households from those earning money incomes to other members, children, spouses, and dependants not earning a money income. Dependants receive clean, serviced accommodation, cooked meals, and clean clothes; the transfers are of transformed commodities purchased from the market which embody value added by household capital and labor.

In a perceptive book, *The Household Economy*, the American writer Scott Burns (1977) observed that 'the hours of work done outside the money economy rival those done inside and will soon surpass them.' Burns saw the household as the strongest and most important economic institution—healthy, stable, and growing. Alvin Toffler (1980) took a similar optimistic view of the household economy in his classic *The Third Wave* where he describes the many ways in which tasks are being driven back from the market to the household.

The pressure to transfer paid labor costs from the market to the unpaid labor costs of the household leads to the development of self-service petrol stations, automatic bank tellers, and Internet shopping. It also leads governments to support unpaid household-based care of sick, disabled, and elderly people instead of professional care in hospitals and nursing homes. Taxation of paid labor helps drive these technology and policy changes.

4.2 *Deficiencies in the National Statistics of Work and Production*

During the latter half of the twentieth century an almost unrecognized statistical revolution took place. At its heart was the System of National Accounts (SNA) that provides the framework for estimating GDP. Governments now spend millions of dollars collecting and publishing regular quarterly statistics of GDP and of the numbers employed in this production. These data permit comparisons of economic performance through time and between countries.

Statistics of GDP and employment are not only the common discourse of economists but have been elevated universally as major tools of economic and social policy including those operating at the international level through the International Labour Office, the International Monetary Fund, and the World Bank.

In the 1980s a number of writers exposed the deficiencies in statistics of work and production. Among these are Luisella Goldschmidt-Clermont (1982) and Ann Chadeau (1985). Perhaps the most strongly advanced reason for measuring household production is to make women's work visible as exemplified by Marilyn Waring's books *Counting for Nothing* (1988) and *Three Masquerades* (1996). Throughout the world, women provide most of the labor needed in household and subsistence production. All of this work is unpaid. On the other hand, in spite of increasing participation of women in paid jobs, most paid work is still done by men.

With few exceptions, the national statistics of work and production continue to ignore the unpaid labor and economic output contributed by women (and men) through household production. At least two-thirds of the work and economic production of women, half of the world's adult population, is excluded. Discussion of public issues such as gender equality, labor market policies, wages, and income policies, to name a few, is statistically misinformed. The point is made frequently that the increase in women's participation in paid work leads to overstatement of the increase in measured economic activity, because the reduction in household production is not counted (Nordhaus and Tobin 1973, Weinrobe 1974).

The SNA definitions used to measure production cover market transactions only. They exclude household production and make no allowance for the destruction of natural resources.

These omissions have been much criticized by the women's and environmental movements. Consequently, the UN Statistical Commission, in the 1993 revision of the SNA, has recommended that national statistical offices prepare accounts for economic activities outside the presently defined production boundary. 'Satellite' accounts separate from, but consistent with, the main SNA accounts of the market economy can be prepared to show the use of natural resources or the extent of economic production by households.

Demands for the full recognition of women's economic production culminated in the Platform for Action adopted in September 1995 at the Fourth World Conference on Women in Beijing. It enjoined 'national, regional and international statistical services and relevant governmental and United Nations agencies' to develop a more comprehensive knowledge of all forms of work and employment by:

Improving data collection on the unremunerated work which is already included in the United Nations System of National Accounts' and developing methods for 'assessing the values, in quantitative terms, of unremunerated work that is outside national accounts, such as caring for dependants and preparing food, for possible reflection in satellite or other official accounts (Platform For Action, Fourth World Conference on Women, Beijing, 1995).

As statistics currently used to measure work and valuable production are incomplete and misleading, new measures are required to produce more accurate statistical pictures.

4.3 *Methods of Estimating the Household Economy*

Some of the earliest estimates of the value of household services were those made for the United States (by Mitchell for 1919 and Kuznets for 1929) and for Sweden (by Lindahl, Dahlgren, and Korb for 1929 and earlier years); they simply multiplied the total number of households in rural and urban areas by the corresponding annual cost of hiring a domestic servant (Hawrylyshyn 1976).

The statistical offices of Denmark, Norway, and Sweden included household production in their national income estimates prior to World War II. Norway's estimates, started in 1912, were quite elaborate, valuing the unpaid work of married women on average wage income for unmarried women and widows in various age groups. They were discontinued in 1950 when the United Nations recommendations were adopted (Aslaksen and Koren 1996).

The Scandinavian view that the real economic production of goods and services included all goods and services produced by households, whether for their own use or for others, was unable to prevail in the debates about the initial standards for national income and product accounting. The market-based approach to GDP estimates continues. However, the way is clear through satellite national household accounts for separate estimates to be made of the value of household production, Gross Household Product (GHP).

The initial information base for such estimates is provided by surveys of the use of time on various household activities (Vanek 1974). International comparability of statistics of the uses of time was given a great boost by the cross-national time budget study in 12 countries conducted in the 1960s under the sponsorship of UNESCO and the International Social Science Council (Szalai 1972). Since then national statistical offices in many OECD countries have followed the Szalai methodology collecting diary-based national surveys of time use from representative samples of households.

The results show that the present statistics of employment greatly understate the volume of economic work needed. Surveys across 12 OECD countries covering various years from 1985 to 1992 show that the mean time spent in paid work was 24h per adult per week. The mean average unpaid work in the household economy was 26h per adult per week (Goldschmidt-Clermont and Pagnossin-Aligisakis 1995, Ironmonger 1995).

The representations of work provided by official statistics of employment are thus at variance with those from official statistics of work from time use surveys. In the main, official statisticians still report an

Table 1
Household productive activities, Australia, 1975–76 \$million

INPUT	Intermediate inputs		Factors of production		Total
	Materials	Energy	Labor	Capital	
OUTPUT					
Meal preparation	6219	144	7922	1140	15425
Cleaning and laundry	271	32	8140	813	9256
Childcare	1208	187	4717	588	6700
Shopping	745	224	4969	453	6390
Repairs and maintenance	483	7	1825	283	2599
Other domestic work	362	7	1743	172	2283
Education and community	392	8	2088	122	2609
Gardening	46	—	1496	182	1725
TOTAL	9726	609	32900	3753	46986

Source: Ironmonger D S (1989).

outmoded view that unpaid domestic work is somehow not ‘work’—only those working in the monetized market sector are counted as being ‘economically active.’

Three methods of imputing a value to the time used in household production have been used, all taking a wage per hour from the market economy. The first is the ‘opportunity cost’ wage that a person could be paid for doing an extra hour of work in a market job rather than an hour of unpaid household work. The cost of an hour of household work is the forgone opportunity to earn in the market. This method is usually rejected since it yields different values depending on who performed the task. The second method uses the wages of specialist paid workers who come to the household, e.g., a cleaner, a cook, a nanny, a gardener, to value the same tasks performed by household members. This ‘specialist replacement cost’ method can be criticized because these workers work more efficiently than a usual household member can, taking less time to perform the same task. Finally, a ‘generalist replacement cost’ method of valuation uses the wage rate for a generalist worker or housekeeper. This is regarded as more appropriate since the working conditions and range of activities are similar to those of household members.

Time-use surveys provide only one source of information about household production since they omit the contribution from nonhuman capital (the land, dwellings, and equipment owned by households). A more complete national accounting approach to measuring and modeling the household economy is required to cover all factors of household production, all intermediate inputs, and all the principal outputs.

5. National Household Accounts and Modeling the Household Economy

5.1 Extended National Accounts

A number of scholars have suggested that national

accounts needed extending to cover missing non-market household production. Essentially, they viewed household production as a missing part of the estimates, not a separate economy. Robert Eisner and co-workers at Northwestern University conducted the most extensive work on these lines in the 1980s. Their research culminated in the book *The Total Incomes System of Accounts* (1989). Eisner’s estimates for the United States make clear the greatly expanded role, in total economic product, of households. United States GHP in 1981 was put at \$1709 billion, 37.5 percent of the extended GNP of \$4560 billion.

5.2 Household Input–Output Tables and Satellite Household Accounts

Household production is best considered as the production of goods and services by a separate economy, complementary to and competitive with the market economy. Every country can be analyzed as two linked economies. The household sells labor to the market; it uses the money income to purchase intermediate commodities from the market which it transforms into items of final consumption through the use of its own unpaid labor and its own capital goods.

The household economy can be studied as a set of six industries providing accommodation, meals, clean clothes, transport, recreation, and care in competition with parallel market industries. Activities such as shopping and cleaning are simply ancillary activities to the principal final outputs of the household economy.

The input–output approach to measuring and modeling the industries of the market economy associated with the work of Leontief (1941) has been applied to measuring and modeling the industries of the household economy. This was started in Australia for Australian households (Ironmonger 1989). Input–output tables showing the uses of labor, capital, energy, and materials in each of the main and ancillary industries of the household economy have now been

prepared for Australia, Canada, Finland, Norway, and the United States.

How the household economy uses time, materials, energy, and capital can be seen from the Table 1 adapted from the first published household input–output table. This is for the Australian household economy. The household input–output tables are exactly the satellite accounts that have been recommended in the 1993 revision of the SNA and in the resolution of the Platform for Action of the 1995 Beijing Women's Conference. The world statistical agencies need to establish standardized methods of estimating GHP.

6. Household Production and a World of Binary Economies

The major scientific achievement of this field has been the measurement of the magnitude of household production through surveys of the uses of time. Household production is now recognized as an alternative economy to the market; in many countries the household economy absorbs more labor and at least one-third the physical capital used in the market economy.

In future, national statistical organizations will produce regular estimates of GHP. Data on outputs of household production—accommodation, meals, clean clothes, and the care of children and adults—will complement data on inputs of unpaid labor and the use of household capital.

Proper recognition of the household economy will have arrived when national household accounts are published each quarter alongside national accounts for the market economy. These data will enable greater scientific research on the organization of household production, the interactions with the market economy, the role of households in building human capital, on the effects of household technology, alternative social and economic policies on gender divisions of labor, and on family welfare.

See also: Boserup, Ester (1910–99); Childbearing, Externalities of; Economic Development and Women; Families and Households, Formal Demography of; Family and Consumer Sciences: United States; Family and Gender; Family and Kinship, History of; Family as Institution; Family Planning Programs; Feminist Perspectives; Family Theory: Feminist–Economist Critique; Feminist Economics; Gender and Feminist Studies in Economics; Gender and Technology; Gender, Economics of; Gender History; Household Archaeology; Household in Anthropology; Land Rights and Gender; Motherhood: Economic Aspects; Poverty and Gender in Affluent Nations; Poverty and Gender in Developing Nations; Rural Industrialization in Developing Nations and Gender; Time-use and Gender

Bibliography

- Aslaksen I, Koren C 1996 Unpaid household work and the distribution of extended income: the Norwegian experience. *Feminist Economics* 2(3): 65–80
- Becker G S 1981 *A Treatise on the Family*. Harvard University Press, Cambridge, MA
- Boulding K 1972 The household as Achilles' Heel. *Journal of Consumer Affairs* 6(2): 110–19
- Chadeau A 1985 Measuring household activities: some international comparisons. *Review of Income and Wealth* 3: 237–53
- Eisner R 1989 *The Total Incomes System of Accounts*. University of Chicago Press, Chicago
- Goldschmidt-Clermont L 1982 *Unpaid Work in the Household: A Review of Economic Evaluation Methods*. International Labour Office, Geneva, Switzerland
- Goldschmidt-Clermont L, Pagnossin-Aligisakis E 1995 Measures of unrecorded economic activities in fourteen countries. Human Development Report Office Occasional Papers 20, United Nations Development Programme, New York
- Hawrylyshyn O 1976 The value of household services: a survey of empirical estimates. *Review of Income and Wealth* 22: 101–31
- Ironmonger D S 1972 *New Commodities and Consumer Behaviour*. Cambridge University Press, Cambridge, UK
- Ironmonger D S 1989 Research on the household economy. In: Ironmonger D (ed.) *Households Work*. Allen & Unwin, Sydney, Australia
- Ironmonger D S 1995 Modeling the household economy. In: Dutta M (ed.) *Economics, Econometrics and The LINK: Essays in Honor of Lawrence R Klein*. Elsevier, Amsterdam
- Ironmonger D S 1996 Counting outputs, capital inputs and caring labor: estimating gross household product. *Feminist Economics* 2(3): 37–64
- Lancaster K 1971 *Consumer Demand: A New Approach*. Columbia University Press, New York
- Leontief W W 1941 *The Structure of the American Economy, 1919–1939*. Oxford University Press, Oxford, UK
- Nordhaus W D, Tobin J 1973 Is growth obsolete? In: Moss M (ed.) *The Measurement of Economic and Social Performance*. National Bureau of Economic Research, New York
- Reid M G 1934 *Economics of Household Productions*. Wiley, New York
- Szalai A 1972 *The Uses of Time: Daily Activities and Suburban Populations in Twelve Countries*. Mouton, The Hague, The Netherlands
- Vanek J 1974 Time spent in housework. *Scientific American* 231: 116–20
- Weinrobe M 1974 Household production and national production: an improvement of the record. *Review of Income and Wealth* 20: 89–102

D. Ironmonger

Housing Economics

Housing has a set of characteristics which together make it different from all other commodities.

(a) *Necessity*: Housing satisfies a basic human need, shelter.

(b) *Importance*: Most households spend more on housing than on any other class of commodity.

(c) *Spatial fixity*: Most housing units are rooted to a particular location. Living in a particular housing unit, a household therefore consumes not only the characteristics of the housing unit *per se*, but also the characteristics of its *location*—its neighbors and neighborhood, its accessibility to other locations, its taxes and public services, and its locational amenities and disamenities.

(d) *Durability*: Housing is the most durable of major commodities. If built soundly and maintained well, a housing unit will last virtually indefinitely.

(e) *Heterogeneity and complexity*: To describe a housing unit reasonably accurately requires enumerating very many characteristics (complexity). Also, housing units exhibit considerable heterogeneity in characteristics, and households considerable heterogeneity in tastes.

(f) *Housing as an asset*: Because housing is durable and the expenditure share on housing high, the housing stock constitutes a substantial proportion of a nation's wealth and of most households' portfolios. The demand for housing therefore combines consumption and asset motives.

(g) *Housing and the macroeconomy*: Since housing is durable, modest fluctuations in housing demand translate into large fluctuations in housing investment. Due to the importance and volatility of housing investment and the magnitude of the wealth effects generated by movements in housing values, the performance of the housing market has a major impact on the macroeconomy.

The interactions of these characteristics cause the economics of housing to be significantly different from the economics of other commodities. Section 2 will examine the supply of housing and Sect. 3 the demand. Section 4 will discuss how supply and demand interact to determine equilibrium in the housing market. Section 5 will look at the housing market from the perspective of real estate economics. Section 6 will consider the role of government with respect to housing.

1. Housing Supply

Before discussing how housing is supplied, a quantifiable definition of 'housing' is needed. The appropriate definition depends on context. In some contexts, highly aggregated definitions that ignore housing heterogeneity—the number of units or total floor area—are most useful. In other contexts, each housing unit is usefully viewed as a bundle of many characteristics. In yet other contexts, intermediate levels of aggregation are most helpful. The definition employed strongly influences the way the market is conceptualized and housing supply and demand described.

Most housing economics is conducted at a high level of aggregation. A nation's monetary authority is interested in the value of the nation's housing stock and the level of housing investment, not in where housing units are located within cities or how many units have two bedrooms. Similarly, the government's housing ministry is principally concerned with housing affordability, crowding, sanitary facilities, and structural soundness, and not in how many units have fireplaces.

At the highest level of aggregation, a housing unit is described simply by the flow of housing services it provides, which can be measured by its rent. Holding its characteristics fixed, an increase in the unit's rent corresponds to a rise in the price of housing services. A housing supply curve may then be defined relating the units of housing services supplied to the price of housing services. Because housing is durable, the elasticity of the housing services supply curve depends on the time horizon. Within, say, a month, the housing supply curve is almost vertical; little can be done beyond maintenance to increase the quantity of housing services supplied. As the time horizon increases, the housing services supply curve becomes increasingly elastic; within six months, a rise in the price of housing services will induce some rehabilitation; within a year, some new housing will be constructed; etc.

In many situations, the analyst is concerned with the interaction between particular housing submarkets. For example, the government might be interested in the effects of a subsidy on owner-occupied housing. To analyze this policy, the housing market would be divided into owner-occupied and rental submarkets. The two submarkets would interact on both the supply (the subsidy would encourage the conversion of housing from rental to owner occupancy) and demand (the subsidy-induced fall in the consumer price of owner-occupied housing services would cause a decrease in rental housing demand) sides. Other divisions of the housing market into submarkets are appropriate in other contexts. If the government is concerned with the living standards of the poor, the housing market is usefully divided into quality submarkets; if it is concerned with downtown revitalization, a division into downtown and suburban submarkets is suitable; etc.

The discussion thus far has treated the supply for housing services—a flow. Lurking in the background is the supply of housing stock. The two are related through the user cost of capital, which is the ratio of the annualized cost of employing a unit of capital to its supply price, and equals the interest rate plus the rate of depreciation plus the maintenance cost rate plus the effective tax rate on capital less the rate of capital gain.

In the short run, rent adjusts to clear the market for housing services. This translates, via the user cost of capital, into the equilibrium asset price or market value of housing. If this exceeds construction cost, new

construction is triggered, which increases the supply of housing stock and of housing services, causing rent to fall. This process continues until a long-run equilibrium is attained where the market-clearing rent equals construction cost times the user cost of capital.

The above conceptualization of the housing market is especially well-suited for aggregate analysis. The remainder of the section reviews several alternative conceptualizations, each of which has its particular domain of application.

The focus of the Mills–Muth model (see Mills (1972) and Muth (1961, 1969), and Brueckner (1986) for an integrative treatment) is on urban spatial structure—how housing density (the ratio of floor to land area), housing rent, and land rent vary with location. Ignore the spatial fixity and durability of structures, so that static analysis can be applied. A landowner adds capital to his land until marginal cost equals marginal revenue. Under competition land rent equals the corresponding producer's surplus. This analysis relates land rent and housing density to housing rent at a particular location. Now turn to the market. Housing rents adjust so that in equilibrium at each location the quantity of housing supplied equals the quantity demanded, with more attractive and accessible locations commanding higher rents. Then from the relationship between housing rent and housing density, the equilibrium urban spatial structure is obtained.

The Mills–Muth model ignores the durability of structures. A related class of models, reviewed in Brueckner (2000), examines the spatial structure of a growing city, accounting for durability. A developer who owns vacant land then faces the decision of *when* to build as well as at what density. The relationship between predevelopment land value and postdevelopment housing value and density can be straightforwardly derived. Normally vacant land is developed from the center outwards. Subsequently, there are successive waves of redevelopment. Each wave entails construction at higher density with the time between waves depending on the city's growth rate. This process is evident in some cities but in others has been curtailed by redevelopment restrictions.

The Mills–Muth model and dynamic durable housing models treat location, housing density, and unit size. Housing units also differ in quality. If two housing units of the same size at the same location command different rents, the higher-rent unit is presumably of higher quality. Since the poor typically live in low-quality housing, quality differentiation is particularly important to consider when evaluating policies aimed at improving low-income housing. Models of housing quality differentiation examine the filtering process (Grigsby (1963)). An urban location goes through a life cycle. Housing is of high quality when constructed but depreciates with age, and as it does so is occupied by successively poorer households. Eventually, rehabilitation/renovation or redevelopment of the site at higher density becomes profitable, and the cycle

repeats. Since the typical neighborhood was initially developed in a short time span, the filtering process gives rise to neighborhood cycles too.

More recent filtering models (Sweeney (1974), Arnott et al. (1983), and Rothenberg et al. (1991)) view the housing market as a hierarchy or continuum of quality-differentiated submarkets. At any instant, the housing market is described by the housing stock at the various qualities. Rent as a function of quality adjusts to clear the market. Construction occurs at those qualities where housing value equals or exceeds construction costs. Maintenance is carried to the point where the last dollar spent increases housing value by one dollar, and a housing unit's price is the present discounted value of rent less maintenance expenditures. These models can be enriched to treat land, as well as housing differentiation by unit size, location, and density (Anas and Arnott (1991)).

Filtering models provide a strong conceptual basis for policy analysis and are intuitively appealing, accommodating all the standard housing conversion processes—construction, maintenance, rehabilitation, upward and downward conversion (combining and subdividing housing units), abandonment, and demolition and reconstruction. Despite their popularity among academic housing economists, filtering models have not yet been widely adopted by the housing policy community.

Working with detailed descriptions of housing units is normally distracting, but in measurement is essential. The minister of housing may be uninterested in whether the poor have fireplaces (as long as they receive heat), but will be interested in whether the average quality of low-income housing has increased. This requires obtaining a quality index, an aggregate of many housing characteristics—including fireplaces. Investors, housing policy analysts, and macroeconomists are all concerned with the time trend of housing prices, which requires developing price indices. And for property tax purposes, housing values need to be estimated accurately, which necessitates detailed description. Hedonic analysis utilizes such detailed description. The market values (or rents) of housing units are regressed against a long list of their characteristics (including public services, neighborhood and neighbor characteristics, accessibility measures, and environmental variables). The regression coefficients are interpreted as implicit market prices for characteristics (Rosen (1974)). Hedonic analysis has proved very useful not only in housing valuation and housing price and quality index construction, but also in inferring the market's valuation of various neighborhood characteristics, public services, air quality, etc. But the implicit conceptualization it employs—separate markets for each of hundreds of housing characteristics—is unwieldy for most housing policy analysis.

Industry studies focus on market organization—industry concentration and the form of strategic interaction between firms. The conventional wisdom

that the housing market is competitive or monopolistically competitive, while inadequately researched, appears sound.

2. Housing Demand

The previous section indicated that how housing supply is described depends on how housing is conceptualized. The same is true of housing demand. Most housing demand studies employ the housing services conceptualization. Accordingly, there is a vast empirical literature estimating the income and price elasticities of demand for housing services. Excellent reviews are provided in Goodman (1996) and Olsen (1986). Muth, whose classic *Cities and Housing* (1969) provides the foundation of modern housing analysis, has argued that since the housing expenditure share has shown remarkable stability over time and location—about 25 percent of after-tax income—the long-run price and income elasticities of housing must be close to unity. Most empirical studies, however, estimate significantly lower price and income elasticities. How can the results of these empirical studies be reconciled with Muth's observation?

The literature has concentrated on four classes of explanations.

(a) With high moving costs (including psychic costs), a household does not instantaneously change its housing when its income or the price of housing changes; rather, housing decisions are long-term. This explains why experiments investigating how households alter their housing consumption under temporary rent subsidies have found a generally price-inelastic response. Relatedly, the demand for housing should depend more on permanent than on transitory income.

(b) That many moves are triggered by demographic change (e.g., marriage, births, retirement, and divorce) makes it more difficult to identify the economic determinants of housing choice.

(c) Tenure choice with borrowing constraints (e.g., Jones (1993), Ortalo-Magné and Rady (1999)). The typical household starts its lifecycle by renting, often under-consuming housing to save up for a down-payment. Having saved up enough, it moves into a starter home, aiming to trade up when it can afford to. The timing of these moves depends on the realization of the household's uncertain income prospects and demographic changes and on the evolution of housing prices, interest rates, and down-payment requirements.

(d) The asset or portfolio demand for housing. When deciding how expensive a home to buy, a household considers how much housing it would like to include in its wealth portfolio. A household's demand for owner-occupied housing achieves a balance between its consumption and portfolio demands. Households adjust the discrepancy between these

demands through tenure choice, customization and furnishing (Henderson and Ioannides (1983)). A household whose asset demand exceeds its consumption demand may purchase a large home but spend little on decoration, or may purchase additional housing units and rent them out. A household whose asset demand falls short of its consumption demand may choose either to own a cheaper house and decorate it nicely or to rent.

Estimating price and income elasticities accounting for all these complications, as well as the tax treatment of housing, is indeed a daunting task. Despite these difficulties, considerable progress has been made in the empirical analysis of housing demand, reflecting advances in data collection, computing power, applied econometrics, and housing economic theory, and the use of controlled experimentation. Regrettably, however, little work has been done on estimating filtering models (Rothenberg et al. (1991) is a notable exception), which has impeded their use in policy analysis.

3. Housing Allocation and Markets

Throughout history, most consumer goods have been allocated through markets. The twentieth century witnessed a sharp rise in the size and scope of government. The communist countries adopted central planning, which entailed government construction, operation, and direct allocation of housing. Elsewhere too, government intervention in the housing sector increased. The pendulum has started to swing in the opposite direction, but by historical standards government intervention in the housing sector remains high. Even in the United States, with the most free-market of western economies, much low-income housing continues to be allocated outside the market, and the allocation of market housing to be deliberately influenced by a myriad of government policies.

Why have governments everywhere doubted the ability of markets to allocate housing? Textbooks in microeconomic theory give a standard list of (classic) market failures or inefficiencies—public goods, natural monopoly, and technological externalities. None is important in the context of housing. The primary motive for extensive government intervention vis-à-vis housing has rather been to improve equity. Some believe that the income distribution generated by the free market is just but most do not. How then is equity to be achieved within a generally market-oriented economy? There are two broad approaches. One is income redistribution, general egalitarianism; the other is effective redistribution through the taxation/subsidization or nonmarket allocation of specific commodities, specific egalitarianism. Most economists favor general egalitarianism because it respects consumer sovereignty. But most governments evidently favor specific egalitarianism. They—and the person

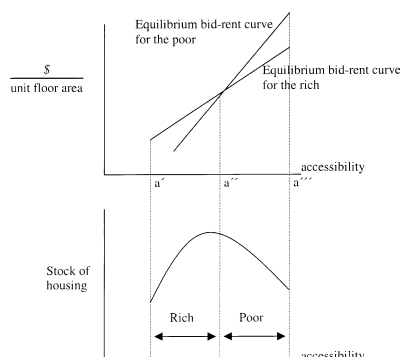


Figure 1
Equilibrium bid-rent curves

on the street—feel that the consumption of some goods (basic housing, clothing, food, education, and medical care) is more meritorious than others (alcohol, cigarettes, drugs, and luxuries). Many governments have favored the nonmarket allocation of housing over subsidization, because it permits more refined targeting and is administratively simpler. There has been remarkably little modern economic analysis of nonmarket housing allocation (Anas and Cho (1988) is a notable exception). The rest of the section will therefore consider market allocation under competition.

The relevant economic theory was developed by Alonso (1964). He asked: How does the market determine the pattern of household location within a city? Suppose that there are two household groups, rich and poor, that each household derives utility from only other goods and housing floor area, and that locations differ only in accessibility. In equilibrium, all rich households achieve the same utility level, U^R . Associated with this utility level is the equilibrium housing bid-rent curve for the rich, which indicates the housing rent (per unit floor area) payable by the rich at different locations consistent with this utility level. Similarly, in equilibrium, all poor households achieve the same utility level, U^P , associated with which is their equilibrium housing bid-rent curve.

Two basic principles apply: (a) A housing unit goes to whomever bids the most for it; and (b) rent equals the bid rent of whomever bids the most for it. The top panel of Fig. 1 plots the equilibrium bid-rent curves for the two groups as a function of accessibility, a . Applying the two principles implies that housing between accessibility levels a' and a'' goes to the rich and that between a'' and a''' to the poor, and that housing rent as a function of accessibility is given by the 'upper envelope' of the equilibrium bid-rent curves. The lower panel of the figure shows the stock of housing available at different accessibility levels. The equilibrium utility levels are determined by the conditions that, for each group, the quantity of housing

demanded equals the quantity of housing supplied. The slope of a group's housing bid-rent curve is the premium that group is willing to pay for improved accessibility. As drawn, the poor outbid the rich at more accessible locations because they are willing to pay a higher premium per unit floor area for improved accessibility. This result has been used in explaining the post-World-War-II suburbanization of wealthier US households.

The accessibility-rent tradeoff is central in determining the pattern of household location. The above analysis can be extended to treat other factors, such as the relative value different groups attach to amenities, disamenities, public services, and neighborhood composition. Each group has an equilibrium bid-rent surface over all locations within the metropolitan area, and lives where it outbids all other groups. Groups may differ not only in income but also in demographic status, race, job location, etc. Extending the analysis to account for taste heterogeneity within groups yields the probabilities that a particular location is occupied by a household in a particular group. The same competitive allocation mechanism operates with respect to other housing attributes. A household group occupies housing of a particular quality, for example, if it outbids all other household groups.

The asset value of a housing unit is the discounted present value of expected future rents minus expected costs, adjusted for uncertainty. The volume of construction depends on profitability, as measured by asset price minus supply price; similarly, the volume of rehabilitation depends on the gain in asset price through rehabilitation minus the cost; and maintenance expenditure depends on how steeply housing value rises with quality.

The above discussion indicates how a perfectly competitive housing market works. In actual housing markets, housing units and households exhibit multi-dimensional heterogeneity. To find a unit that suits its tastes, a household undertakes costly search over vacant units. The friction of search confers market power on landlords, who are consequently price-setters rather than price-takers. With entry and exit close to free, actual housing markets are therefore more accurately described as monopolistically competitive, with excess capacity taking the form of vacant units. Furthermore, housing markets are matching markets, with market tightness (the ratio of searching households to vacant units) playing an important role in the market adjustment process (Anglin and Arnott (1999)). Down-payment requirements on mortgages and credit restrictions on developers also significantly affect the market's operation.

Whether the gain in accuracy from treating the housing market as monopolistically rather than perfectly competitive justifies the increased analytical complexity is an empirical issue and depends on context. Normally little is sacrificed in using perfectly competitive models to analyze the positive effects of

specific housing policies. But imperfectly competitive models should be used to provide microfoundations for housing macroeconomics since nonprice adjustment is important, and to evaluate welfare changes from housing policies.

In the frictionless world of perfect competition, and without taxes, a household would be indifferent between renting and owning. That tenure choice is in fact important indicates that it should be considered from the perspective of imperfect competition. Though often treated in the context of housing demand, tenure choice is an aspect of market equilibrium since it concerns the contractual form through which housing services are exchanged. Rental and single-family owner occupancy are the most common contractual forms; others include condominium ownership (under which common areas are jointly owned) and cooperative ownership (under which a resident is essentially a shareholder in the building).

4. Real Estate Economics

Real estate economics is superbly surveyed in DiPasquale and Wheaton (1996). How does residential real estate economics differ from traditional housing economics? Both are concerned with the housing sector, but real estate economics regards it from an investment perspective, housing economics from a policy perspective. At the beginning of the twenty-first century much of the best research on the economics of housing is being done under the rubric of real estate economics.

One important development has been the extension of real estate investment theory to treat uncertainty satisfactorily. Because real estate investment decisions are made under uncertainty and are largely irreversible, they are usefully viewed as the exercise of real options (Titman (1985), Capozza and Li (1994)), to which the theories of options valuation and irreversible investment can be applied.

Most housing economics is microeconomics applied to the housing sector. But housing macroeconomics is important too. Housing macroeconomics is concerned with the relationships between housing macroeconomic variables (stock, starts, completions, vacancy rates, average rents, etc.), and with how fluctuations in economy-wide variables affect the housing sector, and vice versa. Most large-scale macroeconomic forecasting models of the 1960s contained a housing module. After a period of inactivity, housing macroeconomics has reappeared on the research agenda—in real estate economics, as the study of residential real estate cycles. Current work (e.g. Stein (1995)), and DiPasquale and Wheaton (1996)) suggests that the amplitude of fluctuations in residential real estate values and investment is sensitive to the severity of borrowing constraints and the form of expectations.

Real estate economists have also explored the interlinkages between the real and financial sides of

residential real estate markets, and the design of optimal contracts, considering incentive, selection, and risk-sharing effects, between banks and borrowers, developers and contractors, landlord and tenants, and between the sellers and buyers of houses and real estate agents.

5. Housing Policy

Housing policy debates tend to pit economists against planners. The divisions between the two groups are social philosophical. Since economists value individual freedom to choose, they hesitate more than planners in imposing their vision of the good life on others. Also, economists generally emphasize efficiency and believe it is best achieved via markets, with government enforcing property rights and intervening minimally to mitigate market failure, while planners tend to be suspicious of markets and to object to the inequities and apparent disorder they generate. For all these reasons, planners almost invariably favor more government intervention than do economists.

The planner's view dominated housing policy in western Europe from the 1920s through to the 1980s, especially in the Scandinavian countries where housing markets were largely supplanted by government allocation. The objective of decent and affordable housing for all was largely achieved, but at the cost of considerable inefficiency. During the last decade, these countries have been dismantling the Byzantine allocative and regulatory structures which had evolved, and moving to increasingly free housing markets. Thus, the economist's view is currently in the ascendancy.

Economists' belief in the efficiency of markets is at least well-reasoned. Two central results in microeconomic theory are the First Theorem of Welfare Economics, which states that under perfect competition a market equilibrium is efficient, and the Second Theorem, which states that (under convexity conditions) any efficient allocation can be achieved as a competitive equilibrium with appropriate lump-sum redistribution. If, therefore, an economy satisfies the Theorems' conditions, efficiency and equity can be simultaneously achieved if the government only enforces property rights and redistributes income appropriately, leaving everything else for markets to do, and almost all government intervention in the housing sector is harmful. Whether real world economies come close to satisfying the Theorems' conditions is clearly crucial in determining the appropriate role of government in housing provision, and therefore needs to be considered very carefully.

In the 1960s most neoclassical economists believed that, since buyers and sellers are many and entry and exit free, housing markets are almost perfectly competitive. Most housing economists in the late twentieth century would disagree. Housing markets have many

characteristics which violate the assumptions of perfect competition: housing is a heterogeneous good; transactions costs are important; asymmetric information is endemic; and insurance and futures markets in housing are highly incomplete. As well, because the government cannot observe an individual's 'need' but must instead infer it from a signal (e.g., income) which she can manipulate, all redistributive policies are distortionary. Equity is then most efficiently achieved through a portfolio of policies, including perhaps housing assistance for the poor. Since the optimal portfolio is sensitive to details of the economic environment, housing economists nowadays tend to base their policy advice as much on the results of controlled experiments and empirical studies as on *a priori* reasoning. While consideration of market imperfections expands the scope of potentially ameliorative government intervention, the political economy literature warns that government, too, is highly imperfect.

Government withdrawal from the housing sector has been especially pronounced in direct housing provision. The current view is that low-income housing projects were a mistake. They led to the alienation and ghettoization of the poor, and were excessively costly (since it is efficient for the poor to live in filtered-down housing). Low-income housing policy in the United States today encourages household mixing by race, ethnicity, and income, through subsidizing the nonpoor to live in the projects and the poor to move into middle-income neighborhoods. Other countries are experimenting with selling project units for owner occupancy.

A classic debate in housing economic policy is whether intervention is more effective on the demand or supply side of the market. Generally speaking, housing demand policies can achieve better targeting, but may drive up rents in the short run until the lagged supply response kicks in.

Virtually all housing policies apply to either rental or owner-occupied housing but not both. Most policy makers encourage homeownership, believing it to foster social behavior, civic responsibility, and commitment to neighborhood. The empirical evidence is, however, weak since controlling for the effect that responsible and well-functioning households are more likely to choose homeownership is difficult. Homeownership is encouraged through subsidized and guaranteed mortgages, mortgage interest deductibility under the income tax, tax inducements for first-time homeownership, and subsidies for home improvements.

There are several other elements common to the housing policy debate in most countries.

(a) *Rent control*: (reviewed in Arnott (1995)) Economists have been virtually unanimous in their opposition to traditional (now termed 'first-generation') rent controls, which impose ceilings on rents below market-clearing levels. Almost all jurisdictions which have had first-generation rent controls have either dismantled them or replaced them with second-generation pro-

grams which stress tenant security rather than rent reduction.

(b) *Tax policy*: Most countries use tax policies to achieve housing policy goals.

(c) *Eviction rights vs. security of tenure*: The nature of the landlord-tenant relationship bears the seeds of conflict. Policy concerning eviction for nonpayment of rent entails a classic tradeoff between risk sharing and incentives. Predictably, relative to the United States, European jurisdictions favor the tenant.

(d) *Zoning, building codes, and density and height restrictions*: These significantly affect the spatial pattern of rents and hence household location by income-demographic group. In many cities, height and redevelopment restrictions have forced up downtown property values, driving many poor households previously resident there to surrounding locations. Also, affluent North American suburbs have used minimum lot size zoning and other regulations to exclude the poor.

(e) *Antidiscrimination*: Housing audits have been used in the United States to identify discrimination. A pair of individuals, one of whom belongs to a potentially discriminated-against group, are paid to pose as prospective tenants or prospective homebuyers and are given the same resumé; differences in their treatment are then attributed to discrimination. Also, recent empirical research has attempted to identify discrimination by banks in mortgage lending.

(f) *Homelessness*: In most western countries, homelessness is now a very visible housing problem. While many causes are sociological—the weakened bonds of family and community, higher mobility, deinstitutionalization of the mentally ill, and increased drug use—economic causes are important too (O'Flaherty (1996)). The withdrawal of governments from the housing sector has caused low-quality housing rents in many cities to rise so much that homelessness is a rational response for some very low-income individuals and households.

Bibliography

- Alonso W 1964 *Location and Land Use: Towards a General Theory of Rent*. Harvard University Press, Cambridge, MA
- Anas A, Arnott R 1991 Dynamic housing market equilibrium with taste heterogeneity, idiosyncratic perfect foresight and stock conversions. *Journal of Housing Economics* 1: 2–32
- Anas A, Cho J R 1988 A dynamic, policy-oriented model of the regulated housing market: the Swedish proto-type. *The Review of Economic Studies with Regional Science and Urban Economics* 18: 201–231
- Anglin P, Arnott R 1999 Are brokers' commission rates on home sales too high? A conceptual analysis. *Regional Science and Urban Economics* 27: 719–749
- Arnott R 1995 Time for revisionism on rent control? *Journal of Economic Perspectives* 9: 99–120
- Arnott R, Davidson R, Pines D 1982 Housing quality, maintenance, and rehabilitation. *The Review of Economic Studies* 50: 467–494

- Brueckner J K 1986 The structure of urban equilibria: A unified treatment of the Muth–Mills model. In: Mills E S (ed.) *Handbook of Regional and Urban Economics*, Vol. 2, Chapter 20. North-Holland, Amsterdam, The Netherlands
- Brueckner J K 2000 Urban growth models with durable housing: An overview. In: Huriot J M, Thisse J F (eds.) *Economics of Cities: Theoretical Perspectives*. Cambridge University Press, Cambridge
- Capozza D, Li Y 1994 The intensity and timing of investment: the case of land. *American Economic Review* **84**: 889–940
- DiPasquale D, Wheaton W C 1996 *Urban Economics and Real Estate Markets*. Prentice Hall, Englewood Cliffs, NJ
- Goodman A C 1996 Topics in empirical urban housing research. In: Arnott R (ed.) *Encyclopedia of Economics*, Vol. 1, Harwood, Amsterdam, The Netherlands
- Grigsby W G 1963 *Housing Markets and Public Policy*. University of Pennsylvania Press, Philadelphia, PA
- Henderson J V, Ioannides Y 1983 A model of housing tenure choice. *American Economic Review* **73**: 98–113
- Jones L 1993 The demand for home mortgage debt. *Journal of Urban Economics* **33**: 10–28
- Mills E S 1972 *Studies in the Structure of the Urban Economy*. The Johns Hopkins University Press, Baltimore, MD
- Muth R 1961 Economic change and rural–urban land conversion. *Econometrica* **29**: 1–23
- Muth R F 1969 *Cities and Housing*. University of Chicago Press, Chicago
- Oates W E 1972 *Fiscal Federalism*. Harcourt Brace Jovanovich, New York
- O’Flaherty B 1996 *Making Room*. Harvard University Press, Cambridge, MA
- Olsen E O 1986 The demand and supply of housing service: A critical survey of the empirical literature. *Handbook of Regional and Urban Economics*, Vol. 2, Chapter 25
- Ortalo-Magné F, Rady S 1999 Boom in, bust out: Young households and the housing price cycle. *European Economic Review* **43**: 755–66
- Rosen S 1974 Hedonic prices and implicit markets. *Journal of Political Economy* **82**: 35–55
- Rothenberg J, Galster G C, Butler R V, Pitkin J 1991 *The Maze of Urban Housing Markets*. University of Chicago Press, Chicago
- Stein J C 1995 Prices and trading volume in the housing market: a model with down-payment effects. *Quarterly Journal of Economics* **110**: 379–406
- Sweeney J L 1974 A commodity hierarchy model of the rental housing market. *Journal of Urban Economics* **1**: 288–323
- Titman S 1985 Urban land prices under uncertainty. *American Economic Review* **75**: 505–14

R. Arnott

Human Behavioral Ecology

Human behavioral ecology (sometimes referred to as evolutionary ecology) is the application of evolutionary biological models to the study of behavioral variation in humans. Evolutionary approaches to the study of human behavior fall under a variety of names such as sociobiology, biosociology, biocultural or biosocial science, human ethology, socioecology, and

evolutionary psychology (Cronk 1991, p. 25). While some of these names are synonyms others represent different emphases that will be described below. Human behavioral ecology attempts to explain behavioral diversity as a consequence of environmentally contingent responses made by individuals in their attempts to maximize their inclusive fitness. Critical to this goal is the specification of environmental factors that constrain an individual’s attempt to maximize fitness. As a consequence, individuals develop behavioral strategies designed to solve a variety of adaptive problems set by nature such as producing food, mating, investing in offspring, and managing social interactions between offspring, mates, kin, and unrelated individuals.

1. Origins and Early History of the Field

Behavioral ecology is a relatively new discipline and, for the most part, it is practiced by anthropologists. The publication of E. O. Wilson’s *Sociobiology* (1975) along with the earlier work of biosocial anthropologists such as Robin Fox and Lionel Tiger (Cronk 1991) and evolutionary biologist Richard Alexander (1974) were critical in stimulating interest in applying Darwinian approaches to the study of human behavioral diversity. Very quickly Wilson’s sociobiological approach was found to be limiting because it was largely based on older ethological approaches which posited relatively simple, general, and rigid models of the relationship between genes and human behavior (Winterhalder and Smith 1992, p. 9). This was especially apparent in his final chapter of *Sociobiology* (1975, pp. 547–75) where he considered humans. Here he suggested that certain behavioral traits such as territoriality among foraging peoples were universal (1975, pp. 564–5). This claim and the approach that generated it was shown to be limited in a paper by Dyson-Hudson and Smith (1978), a paper which represented perhaps one of the earliest applications of behavioral ecology to the study of human behavioral variation. Using a cost-benefit model developed by the avian behavioral ecologist J. Brown, they showed that territoriality was variable among tribal populations and was determined by specific environmental factors consistent with evolutionary theory.

As will be more fully described below, environmental cost-benefit approaches in the context optimization models is at the core of much of behavioral ecology. It is vital to understand that behavioral ecology is a branch of evolutionary biology and the core of evolutionary biology is about changes in the frequencies of genes through time. However, behavioral ecology is agnostic about the causative role of genes in the study of human behavioral variation. This position was clarified by Alan Grafen (1984) with his notion of the phenotypic gambit. Phenotypic variation

is shaped by an individual's genotype interacting with a variety of cultural, developmental, environmental, and other non-genetic factors. The degree to which variation in any trait is a consequence of genetic variation is not directly relevant to the formulation or evaluation of behavioral ecological models. This approach is consistent with analyses in non-human behavioral ecology (Alcock 1989).

1.1 Relationship to Other Fields Using Evolutionary Approaches to the Study of Human Behavioral Variation

The new field of evolutionary psychology (Barkow et al. 1992), and, to some extent, its congener human ethology, is often inappropriately confused with behavioral ecology and sociobiology, and it is useful to distinguish between these complementary Darwinian approaches to the study of human behavioral variation. Behavioral ecology attempts to develop hypotheses regarding variation in behavioral strategies that individuals employ to maximize their inclusive fitness. In many studies the central focus is on the reproductive consequences of behavior. In contrast, evolutionary psychologists are interested in elucidating the cognitive mechanisms (or mental modules) that evolved in humans in the 'environment of evolutionary adaptedness.' This concept stresses the fact that all adaptations are to historic environments and whether traits developed in the past are adaptive in current environments is problematical. As a consequence, evolutionary psychologists seek to uncover those mental modules that were designed by natural selection to solve problems such as: What criteria do people use to select mates? How does mate selection criteria differ between the sexes? What factors induce individuals to continue or abandon mating relationships? From this perspective, the reproductive consequences of behavior are not of concern. Other differences distinguishing these fields include behavioral ecology's reliance on naturalistic fieldwork in contrast to questionnaire survey and laboratory experiment, and the relationship between behavior and fertility. Winterhalder and Smith (2000) should be consulted for a more detailed examination of the assumptions, methods, and goals of these fields.

Dual inheritance theory (Boyd and Richerson 1985) is founded on the realization that cultural traits or memes are in some sense identical to genes because both are replicators that use vehicles (bodies) to transmit themselves across generations and are subject to selective forces. Cultural traits differ from genetic traits in their mode of transmission: they may be transmitted from mind to mind through natural selection of carriers of the trait (the trait has a positive effect on fitness), through emulation of high status people who exhibit the trait, by coercion, or by psychological

mechanisms that make the trait desirable or habit forming (e.g., various addictions). In addition, cultural traits may be transmitted laterally within generations instead of lineally through generations allowing them to spread more rapidly. Compared to behavioral ecology and evolutionary psychology little empirical work has been done employing dual inheritance models.

1.2 Relationship to Traditional Social and Behavioral Sciences

Behavioral ecology is distinguished from traditional anthropological and social science theory in its emphasis on the individual as the unit of selection and the idea that much of the content of culture is a consequence of decisions by individuals to enhance their inclusive fitness. Ever since its founding as a discipline, anthropologists have argued that individuals sacrifice themselves to enhance the integration or survival of the social systems in which they participate or are much like puppets or robots manipulated by culture (Cronk 1991). Furthermore, any attempt to reduce or link cultural behavior to lower levels of analysis was specifically condemned. Human behavior was believed to have emergent properties irreducible to the facts of psychology and biology. While it is clear that some dimension of human behavior are not usefully reduced to underlying psychological or biological processes simple observation tells us that humans are biological and psychological entities and some of the behavioral diversity can be usefully linked to underlying levels of analysis.

1.3 Theoretical Armamentarium

Behavioral ecology is strongly informed by a series of theoretical breakthroughs that developed in the early 1960s and 1970s in the evolutionary analysis of social and foraging behavior. These include inclusive fitness (or kin selection, Hamilton 1964), reciprocal altruism (Trivers 1971), sexual selection and parental investment (Trivers 1972), sex biasing in parental investment (Trivers and Willard 1973), optimal foraging theory (Winterhalder and Smith 2000), and, most recently, life history theory (Stearns 1992). These papers clearly demonstrated that behavioral variation was an apt subject for investigation from an evolutionary perspective consistent with the modern synthesis and a gene centered view of evolution. Early on in the development of behavioral ecology there was considerable expectation that kin selection would inspire considerable research especially in the areas of food exchange and other forms of cooperative behavior. As will be evident below, few studies have made kin selection a central focus of investigation.

2. Areas of Application

2.1 Foraging Theory

Human dietary selectivity has been a fundamental area of interest for anthropologists because of its strong implication for the structuring of social relations among hunter-gatherers, group size, mobility patterns, and our understanding of economic dimensions of human evolution. One of the earliest applications of behavioral ecological research came with the application of optimal foraging theory (OFT) in the study of hunting, fishing, and gathering (or foraging behavior) among modern hunter-gathers and foraging horticulturalists. OFT allows researchers to develop a large set of fundamental hypotheses that predict which food resources foragers will pursue when encountered during a search (diet breadth), or where foragers will travel to search for resources (patch choice models) and how long they will stay in these places before moving to other areas (marginal value theorem). These models assume that foragers have a goal, which is to maximize their net rate of return while foraging because by doing so they are ultimately able to maximize their fitness. The rate of return is indexed by a currency such as calories or another attribute of food value (e.g., Hill 1988). This assumption is reasonable because food resources are fundamental to survival and reproduction for one of two reasons. When the environment is poor, humans can be said to be 'resources limited.' Therefore, maximizing one's net rate of return allows foragers to acquire the greatest amount of food while foraging. On the other hand, when the environment is relatively rich, humans may be 'time limited' such that efficient acquisition of food resources leaves foragers with more time to engage in alternative fitness enhancing activities such as child care (Winterhalder and Smith 2000).

There have been a number of qualitative and quantitative tests of optimal foraging models applied to the problem of diet breadth (Hill and Kaplan 1992). Qualitative tests predict directional tendencies in prey choice by showing that diet breadth (the number of different species that will be pursued upon encounter), for example, will expand or contract depending on the encounter rates of highly profitable species (Winterhalder and Smith 2000). Quantitative tests predict the precise number of species a forager will pursue under a variety of environmental conditions (Hill and Kaplan 1992), which species will enter or drop out of the optimal diet breadth, and the exact rank-ordering of species in the diet breadth. Most studies of diet breadth show a division of labor between men and women, and children and adults (Hill and Kaplan 1992) which seem to be based on game versus vegetable products for the former and skill and strength levels for the latter.

Although OFT was designed to account for foraging behavior in animals and then later extended to human foraging, recent applications have been made in the areas of the origins of agricultural and pastoral food production with considerable success (Winterhalder and Smith 2000). It has rekindled anthropological interest in the problem of conservation among native peoples. Thus far, research indicates that true conservation is rare among native peoples but even more importantly researchers are attempting to specify the conditions under which conservation is likely to emerge and an adaptive strategy.

2.2 Group and Resource Transfer

The movement of goods, and, to some extent, services, between families and individuals has received considerable attention by behavioral ecologists. Traditionally such research is subsumed under the terms of exchange or reciprocity. Behavioral ecologists term such activities transfers because it more neutrally characterizes resource movement between groups and/or individuals compared to traditional terms such as exchange or reciprocity (which assume that the transfer of good or services between individuals will be returned in some material form at some point in time). In the analysis of transfers, one assumes that giving or yielding a resource that one has acquired will be done such that acquirer of the resource gains something in return or avoids a cost in defending that resource. As a consequence, models of transfer assume that the benefits of giving are greater than the costs of monopolizing the resource. The models of transfer developed thus far include (a) scrounging (or tolerated theft); (b) risk minimization; (c) trade or exchange; and (d) the show-off hypothesis.

At one extreme, coercive models of scrounging or tolerated theft assume that the resource acquired cannot be easily or effectively defended against those who lack resources and that the benefit of giving or yielding a portion of the resource presents a higher payoff than the cost of defending it against another who are in greater need and thus more willing to contest the resource. At the other extreme are models of sharing, which assume that individuals give resources to others contingent upon an expectation that something of value will be returned to the giver in the future. For example, when resources cannot be predictably acquired (especially true in hunting) risk minimization or pooling strategies are employed through the mechanism of reciprocal altruism. Individuals who acquire resources share with those who have been unlucky in the hunt with the expectation that in the future should they return home without game others who have acquired game will share with them. Kin selection models have been used to explain why resource transfers tend to be biased toward close kin; sexual selection models (e.g., 'show off hypothesis') argue that actors attempt to acquire large but

high-variance resource packets (e.g., large game animals) to demonstrate their desirability as a mate; and trade models suggest that transfer occurs through specialization via the mechanism of reciprocal altruism.

2.3 Mating Strategies

Models of mating or marital strategies, in one form or another, are largely founded on the polygyny threshold model (Winterhalder and Smith 2000, p. 64). This model is relevant when males differentially control resources necessary for successful reproduction. The polygyny threshold model predicts that females will select mates according to the quantity of resources they control. In this situation, polygyny will emerge among those males who can offer more resources to females than unmarried or monogamous males. In short, wealthy males are much more likely to mate polygynously than poor males. In most cases, ownership of material resources (e.g., agricultural land) or high productive productivity (e.g., ownership of large herds) is what differentiates poor from wealthy men. However, in some societies where economic stratification is absent, some men are nevertheless exceptionally productive (e.g., excellent hunters) or have achieved political power or status (e.g., distinguished warriors, leaders, or shamans) and this may be the differentiator between polygynous and monogamous men.

The polygyny threshold model assumes female choice in marital affairs. In some cases this assumption is clearly suspect and male coercion may determine a woman's marital fate. At the same time, some researchers have identified counter strategies that females use to deal with coercive male mating tactics (Winterhalder and Smith 2000, pp. 60–1). Equally troubling is the fact that in some cases female fertility declines under polygyny (this is especially true for junior wives). At this point it is clear that polygyny will nearly always enhance a man's fertility but it frequently impairs a woman's. This latter issue is leading researchers to seek factors that could compensate for lowered fertility. For example, it may be the case that although a woman may bear fewer offspring under polygyny she may produce more grandchildren. This would be true if her sons were polygynous, married earlier than average, or married women with high reproductive value.

Although polygyny is permitted in nearly three-quarters of all anthropological populations, monogamy and the rare case of polyandry (a woman having more than one husband simultaneously) are being investigated through modified versions of the polygyny threshold model. In the case of monogamy it is becoming clear that extramarital mating (concubinage, affairs) and remarriage following divorce (serial monogamy) is more common among wealthy males. Although such systems are officially monogamous,

from a reproductive point of view they produce the high male reproductive variance characteristic of polygynous systems.

2.4 Reproduction: Mating, Parenting, and Life History Strategies

Given that the costs of reproduction vary greatly between males and females, they face fundamentally different trade-offs between the allocation of mating and parental effort in reproduction. The higher costs of reproduction for females limit the number of offspring they can produce and pursuit of additional males through mating effort will have little impact on the quantity or quality of offspring they can produce. As a consequence, they are concerned with the genetic quality of their mates (e.g., ability to deal with pathogens) since these qualities will be passed to offspring, and their ability and willingness to invest in offspring given the fact that humans require some degree of biparental care.

Studies of parental effort have focused on the issues of birth spacing, differential investment in offspring, and the interaction between mating and parenting. The ornithologist Lack was the first to recognize the trade-off between offspring number and offspring quality (or the ability of offspring to thrive and eventually reproduce). Since parental resources are limited, more offspring, through a shortening of the birth interval, will mean less investment per offspring and the potential for reducing each offspring's fitness. This trade-off suggests that there will be an optimal number of offspring that maximizes a parent's fitness and this number will vary with environmental circumstances. In general, one would predict that where parental resources are abundant birth intervals should be short and where parental resources are scarce birth intervals should increase. The crosscultural evidence from natural fertility traditional societies strongly supports the relationship between parental resources and reproductive success: wealthy parents or culturally successful parents have higher fertility than poor parents.

Paradoxically, the association between cultural success and fertility is absent in modern contraceptive societies. The situation has led behavioral ecologists to reexamine the demographic transition that began in Western countries during the eighteenth and early nineteenth centuries, which has now spread to developing nations. During this transition fertility fell faster than mortality rates fell yielding smaller families while family wealth increased. One solution to this paradox has been to investigate whether parents are investing more in children (e.g., education) thereby creating 'higher quality' children who will be able to compete successfully in highly competitive labor markets. At the same time, these children are delaying reproduction as they pursue higher education and gain entry into high status occupations, which require

further delays in reproduction as they work hard to achieve secure occupational niches. This theoretical perspective derives from life history theory described in Sect. 2.4.3.

In a seminal article, Irons (1979) suggested that there ought to be a correlation between cultural and reproductive success. Cultural systems are structured so as to encourage individuals to achieve certain kinds of culturally valued statuses. These achievements vary culturally depending on local environmental constraints. Both in traditional and in a number of recent societies, which have not entered the most recent demographic transition, culturally valued statuses are achieved through hunting prowess, religious rank, herd ownership, political leadership, occupational status, or land holdings. At this point there are nearly 20 studies showing in nearly all cases a positive correlation between cultural achievement and fitness (either fertility or survivorship of offspring) or factors strongly associated with fitness (number of wives, frequency of extra-pair copulations). However, it is very important to note that this relationship is demonstrated for male but not female reproductive success. Significantly, attributes that lead to high male fitness such as polygyny decrease female fitness at least over the short-term but perhaps not the long term.

2.4.1 Parental investment. In general the amount of investment an offspring receives from parents depends on (a) the relatedness between parent and offspring; (b) the effect of investment on the offspring; and (c) the effect of investment on parents' future reproduction and survival. These hypotheses have been positively evaluated using crosscultural and ethnographic cases studies. For example, the probability of infanticide, abuse, and neglect has been shown to be strongly sensitive to these factors. Infants who are unrelated to a woman's current husband ((a) above), or who have congenital defects ((b) above), or who are born in times of economic privation ((c) above) receive little or no investment. The thinking explicit in this approach is that parents adjust their investment in children according to their perceptions of a child's quality, paternity, current environmental conditions, and the needs of other offspring. Further investigations into differential parental investment have been inspired by the Trivers and Willard (1973) model where sex of offspring may influence investment. This model predicts that wealthy parents will invest more in the sex that has greatest variance in reproductive success (usually males) while poor parents will invest more in offspring with least reproductive variance (usually females). Research on societies as diverse as hunter-gatherers, pastoralists, to eighteenth-century Germany and nineteenth-century India has demonstrated the predicted differential investment by sex and distortions in sex ratios.

2.4.2 Parenting as mating effort. Providing care and giving food to children are conventionally seen as kinds of parental effort by males. However, recent research reveals that care and food provisioning are sometimes allocated to children who are unrelated to males. Consequently, such behavior may be fruitfully viewed as attempts to create or maintain a mating relationship. Essentially males are attempting to demonstrate their ability to invest in offspring and therefore convince a female that they would be a valuable partner.

2.4.3 Life history. Life history strategies deal with age specific schedules of mortality and fecundity and the traits that are directly a result of these schedules (Hill and Kaplan 1999, p. 397). It deals with important developmental events such as age at first and last reproduction, growth rates, birth intervals, and senescence. The theory is based on the assumption that effort allocated to somatic effort (growth, development, and maintaining the body) cannot be allocated to reproductive effort (e.g., mating and parenting). Formal research in this area is relatively new and has focused on the peculiar human trait of menopause, inter-birth intervals among foragers, and demographic transition theory.

The grandmother hypothesis attempts to explain the evolution of menopause, a life history trait that is exceptionally rare among mammals and well documented in humans. The model predicts that women will cease reproduction prior to death in order to insure required investment in current dependent offspring and to assist their children in the rearing of grand offspring. Although attractive, the model does not, as yet, show that continued reproduction until death is not a superior alternative.

Throughout history, humans have passed through a number of demographic transitions beginning with a high mortality and high fertility regime to today's worldwide transition to a low mortality and low fertility regime. A central issue behavioral ecologists are confronting is the apparent paradox between fertility and wealth. In tribal and in many state-level historical populations there is a consistent positive correlation between fertility and wealth. This means that the traditional human pattern was to convert wealth into reproduction. However, at the end of the twentieth century there is a global tendency for parents to reduce reproduction where wealth increases. Several explanations are offered for this trend. An initial set of explanations suggested that parents are now concerned with maximizing the number of offspring over more than one generation by reducing risk through an increase in heritable wealth. The 'human capital' perspective notes that in the current environment there is an increased requirement of parental investment needed to make children competitive in a modern

labor market where earning capacity is determined by parental investment in such things as education. Finally, more recent resolutions of the paradox focus on risk sensitive strategies that seek to lower the variance in expected completed family size. An important point evolutionary psychologists make in this problem is that, for the first time in human history, we possess safe and effective means of contraception, an alternative that humans did not previously have as they made their reproductive decisions. This has led to research demonstrating that there is a kind of correlation between wealth and reproductive success, at least for men. This research shows that wealthy men have greater numbers of premarital and extramarital unions than poor men. This lends some credence to the perspective that some of the adaptations we currently possess cannot be easily understood by looking at the reproductive consequences of behavior.

3. Summary

Human behavioral ecology is a young and explicitly Darwinian approach to the study of human behavioral variation. It redresses an absence of evolutionary considerations in the analysis of human behavior and social organization in the social and behavioral sciences. Since its inception in the late 1970s it has spawned the development of new journals and academic specialties explicitly devoted to training students to query the evolutionary bases of behavior. Its penetration into academic life has been uneven. It is most prominent in those disciplines such as anthropology and psychology where human evolution is considered relevant for a holistic understanding of human behavior or in economics where self-interest, methodological individualism, and maximization are key elements in modeling behavior. It has been resisted in departments such as sociology and political science where the findings on human evolution are either ignored or regarded as irrelevant.

See also: Adaptation, Fitness, and Evolution; Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Ecology, Cultural; Environmental Psychology: Overview; Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Evolution: Optimization; Evolutionary Social Psychology; Evolutionism, Including Social Darwinism; Human Ecology; Lifespan Development: Evolutionary Perspectives; Natural Environmental Psychology; Psychological Development: Ethological and Evolutionary Approaches

Bibliography

- Alcock J 1989 *Animal Behavior: An Evolutionary Approach*. Sinauer, Sunderland, MA
 Alexander R 1974 The evolution of social behavior. *Annual Review of Ecology and Systematics* 5: 325–83

- Barkow J, Cosmides L, Tooby J (eds.) 1992 *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press, Oxford, UK
 Boyd R, Richerson P 1985 *Culture and the Evolutionary Process*. University of Chicago Press, Chicago
 Cronk L 1991 Human behavioral ecology. *Annual Review of Anthropology* 20: 25–53
 Dyson-Hudson R, Smith E A 1978 Human territoriality: An ecological reassessment. *American Anthropologist* 80: 21–41
 Grafen A 1984 Natural selection, kin selection, and group selection. In: Krebs J R, Davies N B (eds.) *Behavioural Ecology: An Evolutionary Approach*. Sinauer, Sunderland, MA, pp. 62–84
 Hamilton W 1964 The genetical evolution of social behavior. *Journal of Theoretical Biology* 7: 1–52
 Hill K 1988 Macronutrient modifications of optimal foraging theory: An approach using indifference curves applied to some modern foragers. *Human Ecology* 16: 157–97
 Hill K, Kaplan H 1992 The evolutionary ecology of food acquisition. In: Smith E, Winterhalder B (eds.) *Evolutionary Ecology and Human Behavior*. Aldine de Gruyter, New York, pp. 167–202
 Hill K, Kaplan H 1999 Life history traits in humans: Theory and empirical studies. *Annual Review of Anthropology* 28: 397–430
 Irons W 1979 Cultural and biological success. In: Chagnon N, Irons W (eds.) *Evolutionary Biology and Human Social Behavior: An Anthropological Perspective*. Duxbury Press, North Scituate, MA, pp. 257–72
 Smith E, Winshie in press Conservation and subsistence in small-scale societies. *Annual Review of Anthropology*
 Stearns S 1992 *The Evolution of Life Histories*. Oxford University Press, Oxford, UK
 Trivers R 1971 The evolution of reciprocal altruism. *Quarterly Review of Biology* 46: 35–57
 Trivers R 1972 Parental investment and sexual selection. In: Campbell B (ed.) *Sexual Selection and the Descent of Man: 1871–1971*. Aldine, Chicago, pp. 136–79
 Trivers R, Willard D 1973 Natural selection of parental ability to vary the sex ratio of offspring. *Science* 179: 90–2
 Wilson E O 1975 *Sociobiology: The New Synthesis*. Harvard University Press, Cambridge, MA
 Winterhalder B, Smith E 1992 Evolutionary ecology and the social sciences. In: Winterhalder B, Smith E (eds.) *Evolutionary Ecology and Human Behavior*. Aldine de Gruyter, Hawthorne, NY, pp. 3–23
 Winterhalder B, Smith E 2000 Analyzing adaptive strategies: Human behavioral ecology at twenty-five. *Evolutionary Anthropology* 7: 51–72

R. Hames

Human Capital: Educational Aspects

The notion of ‘human capital’ is an economic concept, based on the distinction between two types of economic activity: investment and consumption. Economists use the term ‘investment’ to refer to an activity or use of resources which will generate income in the future, in contrast with ‘consumption,’ which produces immediate satisfaction or benefits, but does not create capacity to earn future monetary

income. Traditionally, economic analysis of investment and capital tended to concentrate on measurement of physical capital, namely buildings, machinery, or other equipment, which generate income by creating or enhancing productive capacity. More than 200 years ago classical economists, notably Adam Smith in *The Wealth of Nations* (1776), pointed out that education could be regarded as a form of investment, since it is likely to improve the productive capacity of workers, and thus increase their lifetime earnings, just as investment in physical capital, such as building a factory, or buying a new machine, will generate monetary income in the future.

This analogy between investment in physical capital and investment in human capital was recognized by several eighteenth- and nineteenth-century economists, but not fully developed until the 1960s when US economists such as Schultz and Becker published a series of influential works, notably *Investment in Human Capital* (Schultz 1961) and *Human Capital* (Becker 1964), which explored the idea and implications of investment in human capital by attempting to measure, for the first time, the rate of return to investment in education and training. This new emphasis on empirical analysis of human capital, described by another US economist, Mary-Jean Bowman (1966), as 'the human investment revolution in economic thought' proved to be a powerful influence, not only giving rise to a whole new branch of economics—the economics of education—but also affecting education policy in both industrialized and developing countries in a number of ways. This contribution summarizes research and literature on educational aspects of human capital, discusses ways in which estimates of rates of return and the links between education, employment, and economic growth have influenced education policy, particularly on the financing of education, and examines a number of criticisms of the concept of human capital, including the idea that education often represents a 'screening device,' rather than a genuinely productive form of investment, and the more recently developed concept of 'social capital,' which challenges narrow definitions of human capital (see *Education and Economic Growth; Education and Employment; Education and Income Distribution*).

1. Measuring the Rate of Return to Investment in Education

The notion that education is investment in human capital represents more than a simple analogy between education and investment in physical capital. It implies that it is possible to measure the returns to investment

in education, and to apply cost-benefit analysis to decisions about education expenditure, in the same way as rates of return are used to analyze the profitability of investment in conventional physical capital. Cost-benefit analysis is designed to express the costs and benefits of a project in monetary terms, and to calculate its expected profitability, or rate of return, by comparing all future benefits with the total costs of the investment. As noted above, there were several early references to the idea of human capital in eighteenth- and nineteenth-century economic literature, but the first serious attempts to measure the economic costs and benefits of education, and compute rates of return, were in the USA in the 1960s. (Schultz 1961, Becker 1964). These were followed quickly by estimates of the rate of return to education in many other countries. Psacharopoulos (1973) presented international comparisons of rates of return to education, both as a social and as a private investment, in 32 countries, and subsequently published several attempts to update and summarize estimates of rates of return around the world, including more than 50 industrialized and developing countries (see, for example, Psacharopoulos and Woodhall 1985, Psacharopoulos 1994).

Such estimates of the rate of return to educational investment for the individual (the private rate of return) or to society as a whole (the social rate of return) rest on several crucial assumptions. Cost-benefit analysis of education assumes that the economic benefits of education can be measured in monetary terms by using the average earnings differentials between more and less educated workers as a measure of their superior productivity. Workers with higher levels of education have better employment opportunities, and therefore higher average lifetime earnings, than those with less education, and these additional earnings, after various statistical adjustments, are used by economists as a proxy measure of the direct economic benefits of education, and compared with the private and social costs of education (including costs of tuition and earnings forgone by students in school or university). Most rate of return estimates concentrate on these direct monetary benefits, and either ignore or sidestep the problem of measuring the indirect benefits to society (sometimes called 'spillover' benefits or 'externalities') of investing in education. More attention has been paid recently to indirect benefits, but most rate of return estimates still rely mainly on the direct monetary benefits of investment in human capital.

International comparisons of rates of return to education (for example Psacharopoulos and Woodhall 1985) reveal certain general patterns:

(a) Social returns are consistently lower than the private rate of return.

(b) Social and private rates of return to primary education tend to be higher than rates of return to secondary or higher education.

(c) The rate of return to education is higher in developing countries than in developed countries.

(d) The rate of return to investment in education is higher than the average rate of return to physical capital in developing countries, though not necessarily in developed countries.

2. Implications of Rates of Return for Education Policy

Analysis of costs, benefits, and rates of return on investment in human capital raises important implications for both economic and education policies. The idea that education contributes to economic growth by developing and disseminating knowledge, and by improving the quality and productivity of the labor force, is now widely acknowledged and has influenced the economic strategies and educational planning of many individual governments, as well as international agencies and multilateral agencies such as the World Bank. For example, a study for the World Bank (Psacharopoulos and Woodhall 1985, p. 3) notes that:

Education, like other forms of investment in human capital, can contribute to economic development and raise the incomes of the poor just as much as investments in physical capital, such as transport, communications, power, or irrigation. The World Bank ... has long recognized the importance of investment in education and has been active in this field since 1962. When the Bank was first established in 1944, however, education was not counted among the productive purposes for which it was authorized to provide investment capital. This attitude prevailed right through the 1950s, when official statements of Bank policy took the view that the Bank should concentrate its lending on projects designed to make a direct contribution to the productive capacity of its members, rather than finance projects such as the construction and equipment of schools, colleges and universities. Even though such projects were considered essential to the development of a country, the Bank's proper role, as a bank, was said to be confined to providing loans for the more directly productive sectors of the economy. Gradually, this attitude changed, however, both within the World Bank and in the world at large.

More recently, the 1998–9 World Development Report (World Bank 1999), devoted to *Knowledge for Development*, emphasized that 'knowledge is crucial for development,' arguing that education as a form of investment in human capital was a 'key ingredient' in the economic growth and success of East Asian economies such as Hong Kong (China) the Republic of Korea and Singapore (sometimes called the East Asian Miracle economies). (See also *Education and Economic Growth* for a more extended discussion of this topic.)

Policies determining investment by governments, lending by agencies such as the World Bank, and aid policies of bilateral and multilateral agencies have been influenced not only by the assumed and often

demonstrated relationship between education and economic growth, but also by the consistent finding that the rate of return to primary education is the highest among all educational levels, particularly in developing countries. Studies of the economic benefits of primary education were influential in building support for the World Conference on Education for All at Jomtien in 1990, and for the subsequent emphasis on achieving universal primary education as an economic as well as an educational goal (see Colclough and Lewin 1993).

Rate of return estimates also have important implications for policies on education finance. The fact that private rates of return to educational investment are higher than the social rate of return, particularly for higher education, has been used to justify the introduction of tuition fees or student loans at the university level in several countries. For example, when the UK government first announced the introduction of student loans in 1988, the Department for Education and Science (1988, p. 10) presented estimates of social and private rates of return and argued that:

the return on the individual graduate's personal investment in higher education ... in recent years has been in the region of 25 percent, taking all subjects together. The same analysis shows that the return to society on its investment in the graduate's higher education has been much lower: between 5 percent and 8 percent. The individual graduate benefits more than the community as a whole from the latter's investment in the former's higher education. The division of costs, on the other hand, goes the other way. That is why there is scope for the individual student to bear more of the cost of the investment.

Critics argue that such arguments ignore, or at least tend to underestimate, the indirect benefits of education, which accrue to society as a whole as much as (or even more than) the individual. These include the inter-relationships between education, family health, fertility and child mortality, which are now recognized as crucial components of the benefits of education for girls and women in developing countries. An early attempt to analyze differences between rates of return for males and females, (Woodhall 1973) emphasized that 'human capital' should not be interpreted as 'investment in man,' being just as relevant to women as to men, but there was little progress in developing measures of social or indirect economic benefits of education until the 1990s, when a study of *Women's Education in Developing Countries* (King and Hill 1993) included estimates of the returns to women's education which demonstrated clearly the importance of social as well as monetary returns to education for women and girls. This led Larry Summers, then chief economist of the World Bank, to declare, in a Foreword to his book 'I have become convinced that once all the benefits are recognized, investment in the education of girls may well be the highest return investment available in the developing world.'

(Summers 1993, p. v). This has crucial implications for policies to improve equity by reducing barriers to female participation in education.

3. *Objections to Human Capital Theories: Education as a 'Screening Device'*

Although the concept of human capital has a long history in economic theory it remains a controversial idea, when applied to education, and there have been many objections and criticisms to the notion of measuring the rate of return to educational investment. The earliest discussions of human capital suggested that education or training raised the productivity of workers, and hence increased their lifetime earnings, by imparting useful knowledge and skills. However, this assumption was challenged by critics who argued that the earnings of educated workers reflected their superior ability, rather than the specific knowledge and skills acquired during their education. Moreover, critics argued that highly educated workers were more likely to come from higher social or income groups, to work in favored occupations and in urban rather than rural areas. Many estimates of rates of return to education now adjust the observed earnings differentials of educated people to allow for the observed influence of other factors, such as ability or social class background, on earnings, or in the absence of suitable data, they simply assume that 60 percent, rather than the full 100 percent of observed earnings differentials, are due to education, rather than ability or other factors (an adjustment that is often called the 'alpha coefficient' in the literature on the economics of education).

Many critics go further, however, and suggest that education does not improve productivity by imparting knowledge and skills, but simply acts as a 'screening device,' which enables employers to identify individuals of high innate ability or with certain personal characteristics, attitudes, or motivation, which employers value and which they therefore reward by means of higher earnings. This objection to theories of human capital, often called the 'screening hypothesis,' has attracted considerable controversy. Dore for example, in *The Diploma Disease*, an influential critique of the effects of education in developing countries, (Dore 1997, p. 93), presents many criticisms of what he calls the 'rate of return argument,' culminating in the assertion:

We observe people who have certain kinds of education subsequently (on average) getting higher levels of income than other people with less education. But how do we know that this enhanced earning power is *because* of their education (implying on a broader scale, that more education in a society would make that society more productive)? Is the 'investment' metaphor genuinely justified? Is it true that the education

changes a man, *enhances* his capacity to work and in this way alters the price at which he can sell his labours?

Dore gives numerous examples of educational qualifications being used as a 'screening device' or 'filter' in selecting job applicants—an aspect of what he calls the 'diploma disease'—and contrasts the 'screening hypothesis' with 'traditional "human capital theorists" who persist in believing that people earn more because their education has *changed* them and that the "investment" assumptions of their rate of return analysis are correct' concluding that 'different institutional practices, different social structures, different educational traditions and degrees of equality of opportunity make the intervening mechanisms between schooling and earnings levels very different in different countries.' (Dore 1997, pp. 94–5)

Such criticism was challenged, however, on the grounds that while a 'weak' version of the screening hypothesis is undoubtedly true, since employers may well use educational qualifications as a proxy for other characteristics when selecting their work force, there is no evidence to support the 'strong' version of the hypothesis, that education has no direct effect on productivity. The fact that employers continue to pay educated workers more than the uneducated throughout their working lives refutes this.

Even so, the idea of education as a screen or filter has been important in influencing theories of human capital. Blaug (1976, p. 850) in a review which he describes as a 'slightly jaundiced survey of the empirical status of human capital theory' predicts that:

in time, the screening hypothesis will be seen to have marked a turning point in the 'human investment revolution in economic thought,' a turning point to a richer, still more comprehensive view of the sequential life cycle choices of individuals.

4. *Concepts of Social and Cultural Capital*

Just as economists now recognize that the concept of human capital must be enlarged to take account of the effect of education on individual attitudes and motivation, as well as knowledge and skills, and that the indirect benefits (externalities) of education may be even more important than the direct economic and monetary benefits, sociologists and scholars from other disciplines now emphasize the complementarities between human capital, social capital and other forms of investment. Bourdieu (1986) argued that 'capital can present itself in three fundamental guises: as *economic capital*, which is immediately and directly convertible into money and may be institutionalized in the form of property rights; as *cultural capital* which is convertible, on certain conditions, into economic capital and may be institutionalized in the form of

educational qualifications, and as *social capital*.' (Bourdieu 1986, p. 47). At about the same time Coleman (1988, pp. 14–15) argued that analysis of the purely economic effects of human capital is deficient and needs to take into account social structures and relationships in the family and community as well as effects of investment on the individual: 'norms, interpersonal trust, social networks, and social organization are important in the functioning not only of society but also of the economy.' (Coleman 1988, reprinted in Dasgupta and Serageldin 2000, p. 14).

More recently the concept of 'social capital' has attracted considerable attention, including a new study for the World Bank (Dasgupta and Serageldin 2000) that emphasizes the complementarities between human capital and social capital, and between concepts from sociology, anthropology, and political science as well as economics, in analysis of the effects of investment and social action. The overwhelming conclusion of this review is that educational aspects of human capital must also be included in such a 'multifaceted perspective.'

See also: Education and Economic Growth; Education and Employment; Education and Income Distribution; Education, Economics of; Educational Institutions and Society; Educational Productivity

Bibliography

- Becker G S 1964 *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. Princeton University Press, Princeton, NJ
- Blaug M 1976 The empirical status of human capital theory: A slightly jaundiced survey. *Journal of Economic Literature* **14**: 827–55
- Bourdieu P 1986 The forms of capital (translated by Nice R). In: Richardson J E (ed.) *Handbook of Theory of Research for the Sociology of Education*. Greenwood Press, New Haven, CT, pp. 241–58 [reprinted in Halsey A H, Lauder H, Brown P and Wells A S (eds.) 1997 *Education: Culture, Economy, Society*. Oxford University Press, Oxford, UK and New York, pp. 46–58]
- Bowman M J 1966 The human investment revolution in economic thought. *Sociology of Education* **39**: 111–38
- Colclough C, Lewin K 1993 *Educating All the Children*. Clarendon Press, Oxford, UK
- Coleman J S 1988 Social capital in the creation of human capital. *American Journal of Sociology* **94** (suppl.) [reprinted in Dasgupta P, Serageldin I (eds.) 2000 *Social Capital: A Multifaceted Perspective*. The World Bank, Washington, DC, pp. 13–39]
- Dasgupta P, Serageldin I (eds.) 2000 *Social Capital: A Multifaceted Perspective*. The World Bank, Washington, DC
- Department of Education and Science (UK) 1988 *Top-up Loans for Students*. Her Majesty's Stationery Office, London
- Dore R 1997 *The Diploma Disease: Education, Qualification and Development*, 2nd edn. Institute of Education, University of London, UK

- King E M, Hill M A (eds.) 1993 *Women's Education in Developing Countries: Barriers, Benefits and Policies*. John Hopkins University Press, Baltimore, MD and London
- Psacharopoulos G 1973 *Returns to Education: An International Comparison*. Elsevier, Amsterdam
- Psacharopoulos G 1994 Returns to investment in education: A global update. *World Development* **22**: 1325–43
- Psacharopoulos G, Woodhall M 1985 *Education for Development: An Analysis of Investment Choices*. Oxford University Press, Oxford, UK and New York
- Schultz T W (ed.) 1961 Investment in human capital. *American Economic Review* **51**: 1–17
- Smith A 1776 *The Wealth of Nations*
- Summers L 1993 Foreword. In: King E M, Hill M A (eds.) *Women's Education in Developing Countries: Barriers, Benefits and Policies*. John Hopkins University Press, Baltimore, MD and London
- Woodhall M 1973 Investment in education: A reappraisal of the concept of human capital. *International Review of Education* **XIX**: 9–29
- World Bank 1999 *Knowledge for Development: World Development Report 1998/9*. Oxford University Press for the World Bank, Oxford, UK and New York

M. Woodhall

Human Cognition, Evolution of

Human cognition is the way it is, in large part, because it evolved the way it evolved. Unfortunately, for science, however, the key evolutionary processes took place long ago and cognition does not fossilize. This means that scientific investigations of human cognitive evolution must rely on a variety of more-or-less indirect sources of data. Among the most important of these are: (a) the cognition of nonhuman primates; (b) the fossils and material artifacts left behind by humans at various periods of their evolution; and (c) the cultural-cognitive processes at work in modern human populations.

Following common practice in the various branches of comparative biology and psychology, it may be assumed that if all the species of a given evolutionary taxon possess a certain physical trait or behavioral propensity, then the common ancestor for that taxon also possessed that trait or propensity. In the case of humans, the taxon of direct relevance is primates, with special attention to *Pan troglodytes* and *Pan paniscus* (referred to collectively as chimpanzees), because human beings shared a common ancestor with them a mere 6 million years ago. The cognitive processes common to chimpanzees and human beings thus represent the best evidence we have for the cognitive processes of the very first hominids.

For most of the day, every day, primates forage for food, quite often for patchy resources such as fruits and new leaves and buds (which only grow on some

plants at some times of the year), and sometimes for insects or other animal prey who can move freely on their own. In order to forage efficiently, therefore, primates must:

- (a) remember 'what' is 'where' in their local environments, for example, which fruits are in which trees at what times;
- (b) take creative detours and shortcuts to get to the food as quickly as possible;
- (c) track the visible and sometimes invisible movements of prey and other objects;
- (d) recognize and categorize different foods and objects on the basis of perceptual features;
- (e) quantify relative amounts of foods on the basis of perceptual features; and
- (f) use insight in problem solving, for example, in extracting food from substrates.

A variety of empirical studies—both observational in the wild and experimental in captivity—demonstrate that virtually all anthropoid primates have all of these cognitive skills (see Tomasello and Call 1997 for a review). Primates also must use sophisticated cognitive skills to deal effectively with their social environments. Among the most important of these (again, see Tomasello and Call 1997, for a review) are the abilities to:

- (a) recognize individuals within the social group;
- (b) form social relationships with group mates based on such things as kinship, friendship, and dominance rank;
- (c) recognize the 'third-party social relationships' of other group mates based on such things as kinship, friendship, and dominance rank;
- (d) predict the behavior of individuals based on such things as their emotional state and their direction of locomotion;
- (e) use many types of social and communicative strategies to outcompete group mates for valued resources;
- (f) cooperate with conspecifics in problem-solving situations and in forming social coalitions and alliances; and
- (g) engage in various forms of social learning from conspecifics.

Indeed, the fact that the primate social environment presents such a wide array of cognitive challenges has led some theorists to posit that the truly unique features of primate cognition—involving such things as understanding and dealing with relational information—have come from their adaptations not to the physical world but to the social world (Humphrey 1976, Tomasello and Call 1997). Empirical work on primate cognition—coupled with the working assumptions of comparative biology—has thus established that by the time hominids set out on their own evolutionary pathway some 6 million years ago, they already possessed a plethora of cognitive skills, derived mainly from adaptations in the context of foraging for food and dealing effectively with group mates. In brief,

they already conceptualized a sensory-motor world of permanent objects—arrayed in a representational space and categorized and quantified in some fundamental ways—and a social world of animate conspecifics with whom they could cooperate and compete in various complex ways.

However, if we compare contemporary chimpanzees and human beings there are still some cognitive skills that are unique to humans and that therefore must have arisen after the human–chimpanzee divergence 6 million years ago. Most important, in the physical world human beings, and only human beings, understand the causality of events; that is, they understand events in terms of the unobservable 'forces' that explain why they happen in the way they do (Visalberghi and Tomasello 1998). Understanding causality enables individuals to interact with their physical words in much more flexible ways, for example, by creating or blocking key antecedent events in order to produce a desired consequent. Analogously, in the social world human beings, and only human beings, understand the intentionality of behavior; that is, they understand behavior in terms of the unobservable mental 'forces' (thoughts, beliefs, desires) that explain why individuals behave the way they do (Visalberghi and Tomasello 1998). Understanding intentionality enables individuals to interact with conspecifics in much more flexible ways, for example, by attempting to manipulate their knowledge or motivational states so that they will act in desired ways. But perhaps even more importantly, understanding conspecifics as intentional agents constitutes the social–cognitive foundation for human culture, which gives human cognition its species-unique social–collective dimension (see Sect. 3).

1. Hominid Material Artifacts

For the first 4 million of its 6 million years of existence, the human lineage consisted exclusively of one or another species of australopithecine. Australopithecines were small in stature, bipedal, with ape-sized brains, and no stone tools; many modern paleo-anthropologists thus refer to them as 'bipedal apes.' Although we know little of the behavior of these earliest hominids, there are no clear indications that they cognized the world in anything other than the typical great ape manner.

With the birth of the genus *Homo* some 2.5 to 2 million years ago, the human lineage began to show signs of some unique cognitive adaptations. Most important, humans began to make stone tools in ways that required at least some degree of planning and causal understanding of the way rocks could be modified and used instrumentally (Schick and Toth 1993). Thus, early in this period *Homo habilis* made Oldowan (pebble) tools, which required some understanding of how rocks break when struck in certain

ways, and which were often made in one location and then carried to another for subsequent use. Later, *Homo erectus* made Acheulean hand axes (bifaced and symmetrical), which showed a clear ability to hold in mind a definite end product during the process of manufacture and in addition an ability to understand some of the basic causal forces at work in stone tool making. During the early parts of this period (especially from 2 to 1.5 million years ago), the brains of hominids expanded in size by about one third, perhaps because of the cognitive demands of making and using these complex technological products.

A second major brain expansion about 500,000 to 200,000 years ago increased hominid brain size by a factor of one third once again, basically to its modern size. This took place in a set of hominids known collectively as archaic *Homo sapiens*, followed soon by modern *Homo sapiens*. These hominids invented some new techniques for making stone tools, initially the Levallois technique in which a large stone core is specially prepared so that smaller flakes of a special type may be systematically struck off, and later the technique of making blade tools (in much the same way) whose great length to width ratio made them much more multipurpose. Following this in many populations were specialized microlith tool industries, with all kinds of highly specialized tools for highly specialized purposes (each made in a specialized way). Interestingly, some of these flakes, blades, and microliths were hafted onto wooden or bone handles, which made them even more efficient cutting tools. Of special importance from a cognitive point of view, manufacturing these hafted tools required their makers to understand, before the manufacturing process began, that different parts of the tool had different instrumental functions. These so-called composite tools may thus indicate the ability of archaic and early modern *Homo sapiens* to understand and plan for the different causal roles of the different components of a tool.

Another feature of hominid tool use suggests something of how individuals learned to make and use their tools. Both Oldowan and Acheulean types of tools showed great uniformity of use at particular historical periods across all groups of hominids who used them, no matter how far apart from one another they may have lived. In the case of the Oldowan tools of *Homo habilis* (and relatives), this uniformity was confined to Africa and lasted for hundreds of thousands of years. In the case of the Acheulean tools of *Homo erectus*, this uniformity extended all over Africa, Western Europe, and the near East and India, and lasted for over one million years. But with the rise of the more modern techniques of tool making and use, especially the blade and microlith technologies uniquely associated with modern humans, different populations of human beings began to make and use their tools in population specific ways. This new pattern of distribution among social groups suggests the possibility of the cultural

transmission of tool making and using techniques in the different populations, in much the manner of contemporary human populations.

We cannot know much about the social lives of the earliest modern humans because we have so few fossils and artifacts. But if indeed truly cultural organization emerged only with modern humans, then it is reasonable to suppose that there was some change in social cognition that led to this new form of social organization. The main reason to believe this is that before nine months of age human infants are 'acultural' creatures in the sense that they are not yet able to appropriate the cultural artifacts and practices of those around them actively. Soon a major change in their social cognition occurs and they begin to understand other persons as intentional beings like themselves, with mental states like their own. This is indicated by the emergence a whole suite of new behaviors such as:

- (a) following the gaze direction and pointing gestures of adults;
- (b) engaging in shared attentional interactions with adults;
- (c) imitating adult actions on external objects; and
- (d) directing adult attention through pointing and other gestures, all of which indicate a new understanding that other persons have intentional relations to the world (Carpenter et al. 1998).

Because of this new understanding, infants begin to acquire in earnest the tools, linguistic symbols, and other cultural artifacts and activities of their social groups (Tomasello et al. 1993). (Note that children with severe autism never develop this social cognitive ability and they, in a sense, remain acultural individuals; Hobson 1993.) If we assume this same relationship between social cognition and cultural learning in human evolution, we can use the artifactual evidence for the emergence of human cultures to infer that a change in social cognition occurred with modern humans.

2. Culture and Cognition

With the advent of modern humans and their cultural organization, human cognition took on a radically different character. First, with the advent of skills of cultural learning and teaching (resulting from the social-cognitive ability to understand others as intentional and mental agents), human beings began to create artifacts and social practices that accumulated modifications over time. Thus, the so-called 'cultural explosion'—which took place in some, but not all, human populations at about 40,000 years ago—was characterized by many new technological inventions and very rapid changes and improvements in existing technologies. Tomasello et al. (1993) hypothesized that this kind of cumulative cultural evolution, in which the complexity of material and symbolic artifacts ratchets up very quickly over historical time, is

only possible with the human form of cultural learning in which the learner understands the intentions of another individual in the sense that she understands what the behavior is 'for.' The need for this kind of understanding is clearest in the case of tools and symbols since both of these artifacts clearly point to things beyond themselves: the tool to its instrumental function in modifying the physical world and the symbol to its communicative function in modifying other people's attention and behavior.

Once this kind of cultural learning was operative, human populations could, in effect, pool their cognitive resources in a way unique in the animal kingdom. Chimpanzees and other nonhuman primates benefit from observing the activities of their conspecifics, but because they do not understand the intentions of others, what they learn from these observations is limited (Tomasello 1996, see also Kummer and Goodall 1985). Human beings, on the other hand, are able to acquire relatively faithfully many of the tools, symbols, and social practices of others, so that inventions spread relatively rapidly and are preserved in the group effortlessly until some new and improved innovation comes along, which is then acquired faithfully, and perhaps modified further, by another generation. An important result of this ratcheting up of cultural inventions over historical time, is that human children are born into a world of artifacts and social practices that embody something resembling the accumulated wisdom of their entire social group throughout its entire cultural history (Donald 1991).

When children acquire the use of a material or symbolic artifact, or learn to engage in a conventional social practice in their culture, they internalize the intentional states of the adult inventors and users of the artifact or social practice (Vygotsky 1978). This internalization process is of special importance when children acquire linguistic symbols, because a human language embodies all of the myriad ways that the people of a social group have found it useful to categorize and construe their worlds for purposes of communicating with one another. Because the same entity may be construed in many different ways linguistically—this object is a dog, an animal, a pet, or a pest—in acquiring a natural language children are induced to take multiple simultaneous perspectives on entities in the world, with some of these perspectives being common to all languages and some of them being unique to particular languages. When internalized by the child, then, the process of linguistic communication creates in each individual a species-unique form of perspectively based cognitive representation, commonly known as symbolic representation (Tomasello in press).

As a way of highlighting the cognitive importance of processes of cumulative cultural evolution over historical time and their interaction with processes of cultural learning and internalization over ontogenetic

time, we may imagine a human child raised on a desert island in isolation from all other human beings, with a normal brain, body, and sense organs, but with no access to tools, other material artifacts, language, graphic symbols, writing, Arabic numerals, pictures, people who could teach her things, people whose behavior she could observe and imitate, or people with whom she could collaborate. It is likely that as an adult this child's cognitive skills would look very little like those of contemporary human beings (Tomasello in press). The point is thus that modern human beings have the cognitive skills they do because they inherit things from their progenitors both biologically and culturally (cf. the Dual Inheritance Model of Boyd and Richerson 1985, Durham 1991). They inherit biologically the ability to understand conspecifics as intentional agents—and the cultural learning processes that this enables—and these abilities then in turn open the way to the cultural line of inheritance and children's internalization of the ways of understanding the world embodied in their culture's accumulated artifacts and social practices.

3. Conclusion

Much of this account, in common with other accounts of human evolution, is necessarily speculative. But it is based on a solid body of research demonstrating precisely how human cognition is similar to and different from that of nonhuman primates, and on a relatively straightforward search for the two most important species-unique aspects of human cognition in the human artifactual record. First, the earliest humans some 2 million years ago began to demonstrate some new skills of causal understanding in the context of complex tool manufacture and use. Second, and even more important, the earliest modern humans some 150,000 years ago began to demonstrate some new skills of social cognition—understanding conspecifics as intentional and mental agents—that fundamentally changed the process of human cognitive evolution because they enabled uniquely powerful processes of cultural learning and cumulative cultural evolution. These cultural processes, in turn, transformed the ontogenetic niche in which human cognition developed ontogenetically from a social into a truly cultural niche including artifacts, symbols, and social practices with accumulated histories. The ability to pool cognitive resources in these ways may even have been the competitive advantage that modern humans had over other contemporaneous hominids. The ultimate outcome of these evolutionary and cultural processes in humans today is a form of cognition that is culturally and symbolically constituted in ways unprecedented in the animal kingdom.

See also: Animal Cognition; Brain, Evolution of; Cognitive Psychology: History; Cognitive Psychology:

Overview; Emotions, Evolution of; Evolution of Cognition: An Adaptationist Perspective; Intelligence, Evolution of; Intelligence, Genetics of: Cognitive Abilities; Intelligence, Genetics of: Heritability and Causation; Lifespan Development: Evolutionary Perspectives

Bibliography

- Boyd R, Richerson P 1985 *Culture and the Evolutionary Process*. The University of Chicago Press, Chicago
- Carpenter M, Nagell K, Tomasello M 1998 Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development* **255**
- Donald M 1991 *Origins of the modern mind*. Harvard University Press, Cambridge, MA
- Durham W 1991 *Coevolution: Genes, Culture, and Human Diversity*. Stanford University Press, Stanford, CA
- Hobson P 1993 *Autism and the Development of Mind*. Erlbaum, Hillsdale, NJ
- Humphrey N K 1976 The social function of intellect. In: Bateson P, Hinde R A (eds.) *Growing Points in Ethology*. Cambridge University Press, Cambridge, UK, pp. 303–21
- Kummer H, Goodall J 1985 Conditions of innovative behaviour in primates. *Philosophical Transactions of the Royal Society of London Series B* **308**: 203–14
- Schick K, Toth N 1993 *Making Silent Stones Speak*. Simon and Schuster, New York
- Tomasello M 1996 Do apes ape? In: Galef J, Heyes C (eds.) *Social Learning in Animals: The Roots of Culture*. Academic Press, New York
- Tomasello M in press *The Cultural Origins of Human Cognition*. Harvard University Press, Cambridge, MA
- Tomasello M, Call J 1997 *Primate cognition*. Oxford University Press, Oxford, UK
- Tomasello M, Kruger A, Ratner H 1993 Cultural learning. *Behavioral and Brain Sciences* **16**: 495–552
- Visalberghi E, Tomasello M 1998 Primates causal understanding in the physical and psychological domains. *Behavioural Processes* **42**: 189–203
- Vygotsky L 1978 Mind in society. In: Cole M (ed.) *The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA

M. Tomasello

Human Development and Health

Both 'health' and 'human development' are broad terms, covering a lot of conceptual and empirical ground. Both will be defined quite broadly here. Health will encompass the broad range of indicators of physical and mental illness, as well as disability and mortality. Human development is associated with two relevant bodies of scholarship (see Lerner 1998 for extensive discussions). One view, seen most often in

the early history of growth-oriented child development, defines human development as the unfolding of characteristics and capacities as one ages. A second view, used more often in current-day developmental psychology and by social scientists, defines human development as the patterns of skills, attitudes, and behaviors that emerge over the life course. These definitions are not discordant; what distinguishes them most are the assumptions and theoretical contexts on which the definitions rest.

Except for the field of developmental psychopathology (Cicchetti and Cohen 1995), the dominant psychological view is a developmental one that focuses on the patterns typically observed across individuals who are not subjected to unusual traumas or disruptions. Generally it is assumed that developmental processes and outcomes are characteristic of the species; hence, variability is expected to be modest, although both personal experiences and the broader social and cultural contexts can affect the nature and timing of developmental phenomena to a limited extent. The strongest and most generalizable developmental patterns are observed in the early years of life when it is possible to predict with considerable accuracy developmental milestones such as walking, talking, and the capacity for abstract thinking. There has been less interest in and study of developmental changes during adulthood, which are fewer in number, less dramatic, and less consistent across individuals. In fact, it was the fields of lifespan psychology and psychological gerontology that made a major contribution to reorienting developmental psychology to a fully contextual view (see *Lifespan Development, Theory of*, Schaie 1996).

In line with a contextual view of development, social scientists expect substantial variation in the ways in which skills, attitudes, and behaviors are patterned across individuals and the life course. Although they do not discount the possibility of universal or nearly universal laws of development, life course scholars assume that social and cultural forces are the primary determinants of human behavior—and that those forces will exert strong influences on the timing and nature of even those behaviors that are characteristic of the species. Because of dramatic differences in social systems and cultural traditions, wide individual and group variations in patterns of behavior over time are expected. Even when confronted with patterns of behavior that are strongly associated with age, social scientists will tend to view those patterns as the result of strong cultural proscriptions and/or the influence of powerful social forces (Elder 1992).

1. The Link Between Health and Human Development

Scholars studying developmental processes first became interested in illness as a threat to normal

development. Serious illnesses and/or disabilities can constrain the individual's ability to master developmental tasks and/or the timing of mastering them (e.g., Cicchetti and Cohen 1995, Newacheck and Taylor 1992). Moreover, failure to master early developmental tasks can disrupt the achievement of later ones, resulting in a snowballing pattern of developmental disadvantage. Although early research focused largely on disruptions to highly specific developmental tasks (e.g., cognitive development, motor skills), recent research casts a wider net, including, for example, studies of how severe childhood illness can delay or disrupt the development of social skills, self-esteem, and academic achievement (e.g., Gortmaker et al. 1990).

Another research tradition examines the relationships between health and human development from the opposite point of view: how patterns of development influence subsequent health. In this tradition, successful accomplishment of developmental tasks (be they biologically driven or socially/culturally prescribed) is viewed as a form of resource accumulation (human and social capital) that affects health later in the life course. A broad body of research now documents the links between early achievements—including intelligence, educational achievement, self-esteem, feelings of self-efficacy, and the social skills required for successful social relationships—and health in middle and later life (e.g., Clausen 1993, Elder 1999).

Although these are distinct research traditions, both are part of a broader and deeper intellectual transformation of how health and illness are understood by both scientists and the larger public. From the middle of the nineteenth century until the middle of the twentieth century, the *medical model* dominated our understanding of the nature and causes of illness. The medical model views illness as biological dysfunction caused by the failure of the organism's biological defenses to prevent damage or disturbance. It is a complex model, recognizing that threats to the body's integrity can be generated from either external forces (injuries, viral and bacterial antagonists) or internal dysfunctions (congenital deficits, cellular dysregulation). But it is a purely biological model that ignores potentially health-relevant social, psychological, and cultural factors.

Since the 1950s a multidisciplinary model, sometimes called the *biopsychosocial model* of illness, has become the dominant perspective on illness among both scientists and the larger public. The major tenet of this model is that illness is a product of the complex and interacting effects of biological, psychological (including behavioral), social, and cultural factors. An important corollary of this model is that biological intervention alone typically cannot prevent, cure, or manage illness. Both the causes of and appropriate interventions for illness and premature mortality occur at multiple levels. Research on the reciprocal and

complex links between human development and health has been one important illustration of the utility of the biopsychosocial model of illness.

2. Current State of Knowledge

2.1 The Effects of Health on Development

A broad body of research documents the fact that childhood illness, especially chronic illness, is associated with increased risk of developmental problems. This has been demonstrated for a variety of physical illnesses including epilepsy (Hermann et al. 1989), juvenile onset diabetes (Hoare 1984), and cystic fibrosis (Thompson et al. 1994), to name but a few. Problems and delays are observed most frequently for cognitive development and learning, emotional control and maturity, the ability to form successful attachments with peers, and behavioral problems at school and at home. Mental illness during childhood is associated with similar developmental problems (Rutter and Rutter 1993). Although the number of studies that examine the consequences of childhood chronic illness on long-term developmental outcomes is small, research to date suggests that developmental problems tend to multiply over time, supporting psychologists' predictions that early developmental disruptions bode poorly for mastery of later developmental tasks.

There also is strong evidence that childhood mental illness is associated with poor *adult* outcomes. Turnbull et al. (1990), for example, report that the onset of major psychiatric illness before the age of 16 is associated with lower levels of educational attainment, lower occupational attainment and income, and higher rates of separation and divorce. Family formation is also affected by childhood mental illness. Forthofer et al. (1996) report that childhood and adolescent mental illness is associated with earlier, but less stable marriages.

The effects of chronic childhood physical illness on adult achievements are less clear and appear to vary across illnesses. Children suffering from illnesses that can be managed effectively typically attain levels of education, occupation, and income similar to those of their healthy peers (Newacheck and Taylor 1992). In contrast, illnesses that result in frequent incapacity and/or threats to survival (such as cystic fibrosis) and those that involve the brain (such as epilepsy) are associated with poorer adult socioeconomic outcomes.

Although childhood physical and mental illnesses are associated with developmental problems in a number of domains, cognitive problems, especially learning disabilities, have the most severe consequences for both short- and long-term outcomes, as demonstrated in studies comparing children who did

and did not experience learning disabilities as a result of epilepsy (Hermann et al. 1989). Children with epilepsy who did not have learning disabilities resembled healthy children with regard to social competence, depression, and behavior problems at school and at home. In contrast, children who experienced learning disabilities as a result of epilepsy exhibited developmental problems in a wide range of areas. Relative to other developmental problems, learning disabilities also are most strongly related to low levels of socioeconomic achievement during adulthood.

There is also an intergenerational component to the effects of health on development during childhood and adolescence. A broad body of research consistently reports that children who live with a parent who suffers from chronic physical illness, chronic or intermittent mental illness, and/or disability are at increased risk of developmental problems or delays (e.g., Hirsch et al. 1985). Indeed, the effect sizes for parental illness are as great as, and sometimes greater than, those for childhood illness. Again, a variety of developmental problems are observed, including cognitive and learning problems, behavior problems at school and at home, general social competence, and the ability to make successful attachments to peers.

Given the strong evidence that chronic illness (of self or parent) places children and adolescents at risk of developmental problems, recent research appropriately focuses on the processes and mechanisms that account for these relationships. This is a highly complex issue because there are multiple pathways by which illness affects development. For children who are chronically ill, characteristics of the illness, its *sequellae*, and the nature and efficacy of medical management all play a role in helping to protect against or, alternatively, increasing the risk of developmental problems. As noted above, some illnesses have direct effects on developmental outcomes (e.g., when diseases cause cognitive problems or constrict the child's ability to form bonds with peers). Illness severity has an impact on development beyond the symptoms and impairments generated by the disease. For example, the greater the level of absence from school as a result of illness, the more severe the developmental problems (McCubbin 1988). Medical treatments also play a role in developmental outcomes. Some medications have side effects that place the child at risk of developmental obstacles (e.g., inability to concentrate); other treatments (e.g., chemotherapy) can impair functioning in multiple domains. Finally, the stigma associated with chronic illness (Hermann et al. 1989) can lead to rejection from peers and discomfort in public places (e.g., diabetic children require special diets and cannot participate in many 'normal' childhood activities).

Beyond the constraints of the illness and its treatment, the family plays a strong role in increasing or lessening the ill child's risk of developmental problems. Quality of parenting and the parents' methods of

coping with a chronically ill child are especially important (McCubbin 1988). The entire dynamic of family life is changed by the presence of a chronically ill child. Parenting requires more time, special skills, and maintaining a delicate balance between 'normalizing' the child's life as much as possible, while also providing high levels of supervision and medical management. Demands change over time as the illness waxes or wanes and as the child ages. Providing care to a chronically ill child over long periods of time is draining and affects all areas of parent's lives (e.g., the marital relationship, relationships with other children, labor force participation, financial security) (Avison et al. 1991).

Family dynamics also appear to explain the conditions under which children experience developmental problems when a parent is chronically ill. Two processes seem to be especially important in this regard. First, if parental illness is accompanied by high levels of marital discord, the likelihood of developmental problems in the children is greatly increased (e.g., Mann and MacKenzie 1996). Second, the risk of developmental problems increases exponentially if the parent's illness is associated with being 'emotionally unavailable' to the child (e.g., Wootton et al. 1997). Indeed, developmental delays can be observed by the age of 2 when the parent is emotionally estranged or distant from the child (Radke-Yarrow et al. 1985).

3. The Effects of Development on Health

There also is impressive evidence that developmental outcomes during childhood and adolescence have strong effects on health in middle and late life. There are three primary pathways by which early developmental outcomes exert persisting effects on health: socioeconomic achievements, social integration, and social support.

Of all the social factors related to health, socioeconomic status (SES) is the most consistent and powerful (House et al. 1994, Link and Phelan 1995). SES typically is operationalized as a combination of income and education, and sometimes occupation as well. All three SES components are related to health and longevity, although they seem to operate in somewhat different ways. The relationship between income and health is multifaceted, for example, income is related to access to and quality of healthcare, and to goods and services that promote health (e.g., safe housing, adequate nutrition). Occupation is, of course, a primary determinant of income, but also affects health via prestige and greater autonomy and control in occupational settings (e.g., Link et al. 1993). Education typically exhibits the strongest relationships with health outcomes, in part via its effects on occupation and income, and more directly via better health habits and increased knowledge and use of

information about health and healthcare (Ross and Wu 1995). SES during adulthood is strongly related to early development. As noted earlier, school performance and educational attainment are major developmental outcomes during early life.

Social integration refers to the attachments individuals sustain with the larger society and are typically measured in terms of occupational, organizational, and community roles. A broad body of research, beginning with Durkheim's classic work on suicide, has documented strong links between social integration and health and longevity (House et al. 1988). Again, early developmental outcomes are linked directly to social integration during adulthood. Sustaining social integration requires social competence, reasonable intelligence, and the self-discipline required for adequate role performance—all of which are developmental outcomes that are established, in large part, during early life.

There is impressive evidence that health is related robustly to social relationships. The effects of social ties on health operate in a variety of ways. First, high-quality relationships with family and friends are directly related to better health and longevity (e.g., House et al. 1988). Second, close relationships are the foundation from which social support emerges during times of stress. Social support takes a variety of forms ranging from consolation and reassurance to the provision of information and tangible assistance. High levels of social support help to protect against the onset or exacerbation of illness and facilitate recovery from illness (de Leon et al. 1999, George et al. 1989). The ability to form appropriate attachments and the social skills needed to sustain social bonds are established early in life.

Thus, early development sets the stage for subsequent health in a variety of ways. Moreover, the benefits of successful development tend to increase over time, whereas the disadvantages that result from unsuccessful development also spiral over time. This pattern, known as cumulative advantage/disadvantage (Ross and Wu 1996), results in diverging levels of health across adulthood, until very late life when biological factors may act as 'levelers,' narrowing the differences between the advantaged and disadvantaged.

Health in middle and late life is affected by both proximal and distal factors. Unquestionably, proximal and contemporaneous factors are typically more powerful predictors of morbidity and mortality than are distal patterns. Nonetheless, life course research has added to the knowledge base in two ways. First, sometimes, early events and experiences have persisting demonstrable effects on health and longevity, even when more proximal factors are taken into account. Second, the indirect effects of early events and experiences often help to elucidate the selection processes that result in more proximal antecedents of morbidity and mortality.

4. Methodological Challenges

A variety of methodological issues complicate efforts to better understand the links between development and health; two especially important ones are considered here. First, adequate depiction of the complex and time-dependent relationships between development and health requires longitudinal data collected over a significant portion of the life span. Such studies are especially important for identifying the effects of development on health because it is not until middle age that a significant proportion of the population exhibits health problems. Few studies to date assess developmental and health outcomes over the appropriate age range (i.e., from childhood/adolescence until at least late middle age). Such studies are very expensive in both time and money, making a substantial increase in the future unlikely.

Selective mortality also poses a serious threat to our understanding of the links between development and health, both of which are related strongly to mortality. Extant research is based on survivors, which has two important implications. First, it is likely that research findings *underestimate* the effects of developmental outcomes on subsequent health because the most unhealthy are likely to have died. Second, selective mortality may account for some of the observed patterns among survivors. For example, the narrowing of SES differences in health during late life may be due to earlier mortality among low SES persons rather than a 'biological leveling' at the end of life (Bartley and Plewis 1997). There is no way to overcome this problem. However, given longitudinal data covering large segments of the life course, the effects of selective survival can be estimated. The possible effects of selective survival should be considered in the interpretation of all empirical findings.

5. Future Directions

Future research on the links between health and human development is likely to focus on two major topics. First, efforts will continue to trace the specific pathways and mechanisms by which development and health affect each other. Although extant research has made important contributions to this issue, substantial components of these processes remain unknown and replication of findings to date is especially important because of reliance on cross-sectional and short-term longitudinal research designs. Second, future research will address the extent to which and ways in which developmental and health problems in early life can be compensated for or reversed. Some contributions to this issue will result from naturalistic research that identifies compensatory factors that permit early disadvantage, in either development or health, to be overcome or minimized. Other contributions will result from research in which interventions designed to

prevent developmental or health problems or to compensate for them are evaluated in terms of feasibility, cost, and effectiveness. Given the complex web of factors associated with early developmental and health problems, the challenges of arresting or reversing trajectories of disadvantage will be great, but the benefits will be great as well.

See also: Adolescent Health and Health Behaviors; Adult Development, Psychology of; Chronic Illness, Psychosocial Coping with; Coping across the Lifespan; Coronary Heart Disease (CHD), Coping with; Developmental Psychology; Health Behavior: Psychosocial Theories; Health Behaviors; Health Behaviors, Assessment of; Human Development, Successful: Psychological Conceptions; Illness: Dyadic and Collective Coping; Lifespan Development, Theory of; Midlife Psychological Development; Well-being and Health: Proactive Coping

Bibliography

- Avison W R, Noh S, Speechley K N 1991 Parents as caregivers: Caring for children with health problems. In: Albrecht G A, Levy J L (eds.) *Advances in Medical Sociology*. JAI Press, Greenwich, CT, Vol. 2, pp. 65–94
- Bartley M, Plewis L 1997 Does health-selective mortality account for socioeconomic differences in health? Evidence from England and Wales, 1971–1991. *Journal of Health and Social Behavior* **38**: 376–86
- Cicchetti D, Cohen D J (eds.) 1995 *Developmental Psychopathology*. Wiley, New York, 2 Vols
- Clausen J A 1993 *American Lives*. University of California Press, Berkeley, CA
- Costa P T, McCrae R R 1996 Mood and personality in adulthood. In: Magai C, McFadden S H (eds.) *Handbook of Emotion, Adult Development, and Aging*. Academic Press, San Diego, pp. 369–83
- de Leon C F M, Glass T A, Beckett L A, Seeman T E, Evans D A, Berkman L F 1999 Social networks and disability transitions across eight intervals of yearly data in the New Haven EPESE. *Journal of Gerontology: Social Sciences* **54B**: S173–89
- Elder G H 1992 Life course. In: Borgatta E F, Borgatta M L (eds.) *Encyclopedia of Sociology* pp. 1120–30
- Elder G H 1999 *Children of the Great Depression*, 25th Anniversary edn. Westview Press, Boulder, CO
- Forthofer M S, Kessler R C, Story A L, Gotlib I H 1996 The effects of psychiatric disorders on the probability and timing of first marriage. *Journal of Health and Social Behavior* **37**: 121–32
- George L K, Blazer D G, Hughes D C, Fowler N 1989 Social support and the outcome of major depression. *British Journal of Psychiatry* **154**: 478–85
- Gortmaker S L, Walker D K, Weitzman M, Sobal A M 1990 Chronic conditions, socioeconomic risks, and behavioral problems in children and adolescents. *Pediatrics* **85**: 267–76
- Hermann B P, Whitman S, Dell J 1989 Correlates of behavior problems in children with epilepsy. In: Hermann B P, Seidenberg M (eds.) *Childhood Epilepsies: Neuro-psychological, Psychosocial, and Intervention Aspects*. Wiley, New York, pp. 143–57
- Hirsch B J, Moos R H, Reisch T M 1985 Psychosocial ad-

- justment of adolescent children of a depressed, arthritic, or normal parent. *Journal of Abnormal Psychology* **94**: 154–64
- Hoare P 1984 Does illness foster dependency? A study of epileptic and diabetic children. *Developmental Medicine and Child Neurology* **26**: 14–19
- House J S, Landis K, Umberson D 1988 Social relationships and health. *Science* **241**: 540–5
- House J S, Lepowski J M, Kinney A M, Mero R P, Kessler R C, Herzog A R 1994 The social stratification of aging and health. *Journal of Health and Social Behavior* **35**: 213–34
- Lerner R M (ed.) 1998 *Handbook of Child Psychology. Vol. 1: Theoretical Models of Human Development*. Wiley, New York
- Link B, Lennon M C, Dohrenwend B 1993 Socioeconomic status and depression: The role of occupations involving direction, control, and planning. *American Journal of Sociology* **98**: 1351–87
- Link B G, Phelan J 1995 Social conditions as fundamental causes of disease. *Journal of Health and Social Behavior extra issue*: 80–94
- Mann B J, MacKenzie E P 1996 Pathways among marital functioning, parental behaviors, and child behavior problems in school-age boys. *Journal of Clinical Child Psychology* **25**: 183–91
- McCubbin M A 1988 Family stress, resources, and family types: Chronic illness in children. *Family Relations* **37**: 203–9
- Newacheck P, Taylor W R 1992 Childhood chronic illness: Prevalence, severity, and impact. *American Journal of Public Health* **82**: 364–71
- Radke-Yarrow M, Cummings M, Kuczynski L, Chapman M 1985 Patterns of attachment in two- and three-year olds in normal families and families with parental depression. *Child Development* **56**: 884–93
- Ross C E, Wu C 1995 The links between education and health. *American Sociological Review* **60**: 719–45
- Ross C E, Wu C 1996 Education, age, and the cumulative advantage in health. *Journal of Health and Social Behavior* **37**: 104–20
- Rutter M, Rutter M 1993 *Developing Minds: Challenge and Continuity Across the Life Span*. Basic Books, New York
- Schaie K W 1996 *Adult Intellectual Development: The Seattle Longitudinal Study*. Cambridge University Press, New York
- Thompson R J, Gustafson K E, George L K, Spock A 1994 Change over a 12-month period in the psychological adjustment of children and adolescents with cystic fibrosis. *Journal of Pediatric Psychology* **19**: 189–203
- Turnbull J E, George L K, Landerman R, Swartz M S, Blazer D G 1990 Social outcomes related to age on onset among psychiatric disorders. *Journal of Consulting and Clinical Psychology* **58**: 832–9
- Wootton J M, Frick P J, Shelton K K, Silverthorn P 1997 Ineffective parenting and childhood conduct problems: The moderating role of callous-unemotional traits. *Journal of Consulting and Clinical Psychology* **65**: 301–8

L. K. George

Human Development, Bioecological Theory of

The bioecological model, together with its corresponding research designs, is an evolving theoretical system for the scientific study of human development

over time. Within the bioecological theory, development is defined as 'the phenomenon of continuity and change in the biopsychological characteristics of human beings, both as individuals and as groups. The phenomenon extends over the life course, across successive generations, and through historical time, both past and future.' The term 'future' raises a question: How is it possible to investigate scientifically phenomena that have not yet taken place? The question, of course, is hardly new; indeed it pervades every field of scientific endeavor. However, we are the only species that has, over historical time, developed the capacity to engage successfully in scientific inquiry, and thereby, in many respects, has been able to change the nature of the world in which we live. As a result, within certain limits, we humans have altered the nature and course of our own development as a species (Bronfenbrenner and Evans 2000, Bronfenbrenner and Morris 1998).

To place the bioecological theory of human development into a larger context, it is important to recognize that many of the general perspectives advanced and elaborated in this theory, are also parts of other related lines of theoretical and empirical inquiry into human development. Examples are lifespan psychology (Baltes et al. 1999), cultural psychology (Cole 1995), Magnusson's developmental theory of contextual-interactive holism (Magnusson and Stattin 1998), and, especially, the work of Robert Cairns, who through communications and publications extending over three decades, has played a major role in the evolution of the four defining properties of bioecological model: process-person-context-time. Cairns is of course best known as the founder and principal protagonist of developmental science. Examples of his books and articles that have been most relevant to the evolution of the ecological model are cited in the bibliography (Bergman et al. 2000, Cairns 1970, Cairns and Cairns 1994). The specific profile of the theory of Bronfenbrenner and his colleagues is its comprehensive interdisciplinary and integrative focus on the age periods of childhood and adolescence and its explicit interest in applications to policies and programs pertinent to enhancing youth and family development.

1. Defining Properties of the Bioecological Model

The term 'evolving' highlights the fact that the model, along with its corresponding research designs, has itself undergone a process of development over its own 'life course.' Another defining property of the bioecological model specifies that it deals with two closely-related but nevertheless fundamentally different developmental processes, each taking place over time. The first defines the phenomenon under investigation: namely, that of continuity and change in the bio-

psychological characteristics of human beings. The second focuses on the development of the scientific tools—the theoretical models and corresponding research designs required for assessing the continuity and change.

These two tasks cannot be carried out independently, for they are the joint product of emerging and converging ideas, based on both theoretical and empirical grounds—a process called 'developmental science in the discovery mode' (Bronfenbrenner and Evans 2000, pp. 999–1000). In the more familiar 'verification mode' the aim is to replicate previous findings in other settings to make sure that the findings still apply. By contrast, in the discovery mode, the aim is to fulfill two broader but interrelated objectives:

(a) Devising new alternative hypotheses and corresponding research designs that not only call existing results into question, but also stand a chance of yielding new, more differentiated, more precise, replicable research findings and, thereby, producing more valid scientific knowledge.

(b) Providing scientific bases for the design of effective social policies and programs that can counteract newly emerging developmentally disruptive influences. This has been an explicit objective of the bioecological model from its earliest beginnings.

A major challenge to today's bioecological model is to discover how such new working hypotheses and corresponding research designs can be developed for the future. One answer lies in the possibility that, despite historical change, some elements of the model, and their inter-relationships, may remain constant across both time and space. From this perspective, today's model has several distinctive defining properties that become the foundation for the rest. Some are of relatively recent origin; others date back to the model's earliest formal beginnings (Bronfenbrenner and Evans 2000, Bronfenbrenner and Morris 1998, for reviews). Each is expressed here in the form of a Proposition.

1.1 Proposition I

An early critical element in the definition of the ecological model is experience. The term is used to indicate that the scientifically relevant features of any environment for human development include not only its objective properties but also the way in which these properties are subjectively experienced by the persons living in that environment. This equal emphasis on an 'experiential' as well as an 'objective' view springs neither from any antipathy to behaviorist concepts, nor from a predilection for existential philosophical foundations. It is dictated simply by a hard fact. Very few of the external influences significantly affecting human behavior and development can be described solely in terms of objective physical conditions and

events (Bronfenbrenner and Evans 2000, Bronfenbrenner and Morris 1998).

A critical term in the foregoing formulation is the word 'solely.' In the bioecological model, both objective and subjective elements are posited as driving the course of human development; neither alone is presumed sufficient. Moreover, these elements do not always operate in the same direction. It is therefore important to understand the nature of each of these two dynamic forces, beginning on the phenomenological or experiential side. Both of the underlined terms are relevant because, while related to each other, they are typically applied to somewhat different spheres. The former is more often used in relation to cognitive development and pertains mainly to changes in how the environment is perceived at successive stages of the life course, beginning in early infancy and proceeding through childhood, adolescence, adulthood, and, ultimately, old age.

By contrast, 'experience' pertains more to the realm of feelings; for example, anticipations, forebodings, hopes, doubts, or personal beliefs. These too, emerge in early childhood, continue through life, and are characterized by both stability and change. They can relate to self or to others, and especially to family, friends, and other close associates. They can also apply to the activities in which one engages; for example, those that one most or least likes to do. But the most distinctive feature of such experiential qualities is that they are 'emotionally- and motivationally-loaded', encompassing both love and hate, joy and sorrow, curiosity and boredom, desire and revulsion, often with both polarities existing at the same time but usually in differing degrees. A significant body of research evidence indicates that such positive and negative subjective forces, evolving in the past, can also contribute in powerful ways to shaping the course of development in the future (Bronfenbrenner and Evans 2000, Bronfenbrenner and Morris 1998).

But these are not the only powerful forces at work. There are others that are more objective in nature. This does not mean, however, that they are necessarily either more or less influential, mainly because the two sets of forces are interdependent and affect each other. Like their subjective counterparts, these more objective factors also rely for their assessment on corresponding theoretical models and associated research designs that have evolved over time. These more objective relationships are documented below in the form of two propositions (Bronfenbrenner and Evans 2000, Bronfenbrenner and Morris 1998). The first specifies the theoretical model, and provides concrete examples; the second foreshadows the corresponding research designs for their assessment.

1.2 Proposition II

Over the life course, human development takes place through processes of progressively more complex

reciprocal interaction between an active, evolving biopsychological human organism and the persons, objects, and symbols in its immediate external environment. To be effective, the interaction must occur on a fairly regular basis over extended periods of time. Such enduring forms of interaction in the immediate environment are referred to as proximal processes. Examples of such processes include feeding or comforting a baby; playing with a young child; child-child activities; group or solitary play; reading, learning new skills; athletic activities; problem solving; caring for others; making plans; performing complex tasks; and acquiring new knowledge and know-how.

For the younger generation, participation in such interactive processes over time generates the ability, motivation, knowledge, and skill to engage in such activities both with others and on one's own. For example, through progressively more complex interaction with their parents, children increasingly become agents of their own development, to be sure only in part.

In sum, proximal processes are posited as the primary engines of development. The next defining property speaks to the corresponding research designs.

1.3 Proposition III

The form, power, content, and direction of the proximal processes producing development vary systematically as a joint function of the characteristics of the developing person (including genetic inheritance); the environment—both immediate and more remote—in which the processes are taking place; the nature of the developmental outcomes under consideration; and the continuities and changes occurring in the environment over time, through the life course, and during the historical period in which the person has lived.

Propositions II and III are theoretically interdependent and subject to empirical test. An operational research design that permits their simultaneous investigation is referred to as a Process-Person-Context-Time model (PPCT).

In the corresponding research designs for the bioecological model, the element of time has special importance. To show that development has actually occurred, the research design must demonstrate, or at least make plausible, that the elements in the design, and their dynamic relationships to each other, have influenced the 'biopsychological characteristics of the developing person over an extended period of time.'

For example, a rich data archive generously made available by Small and Luster (1990) from their state-wide studies of youth at risk in Wisconsin has made possible some reanalyses of working hypotheses derived from the newly-emerging formulations.

Figure 1 depicts the results from a reanalysis of the effects of parental monitoring on the academic

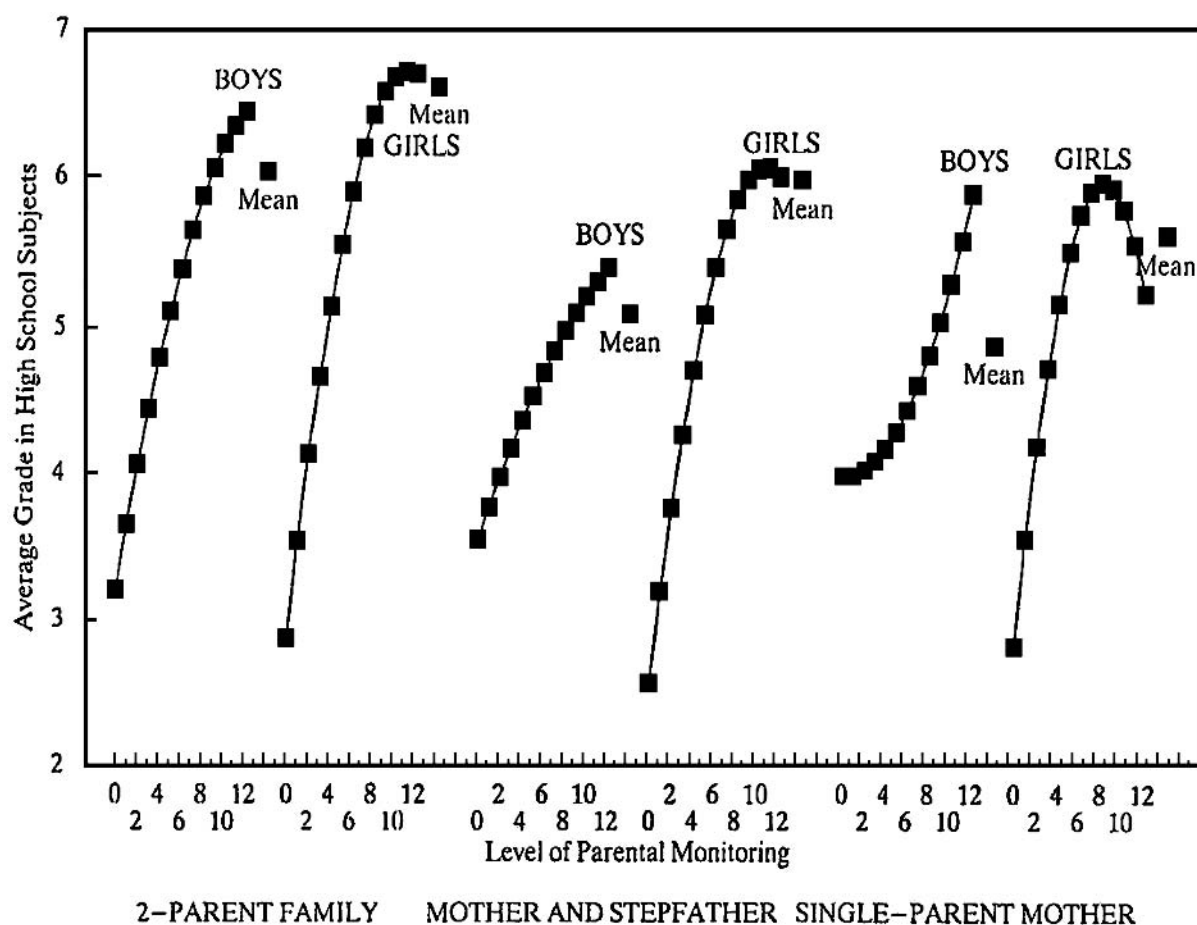


Figure 1
Parental monitoring and high school grades by gender: mothers with more than a high school education

achievement of high school students living in one of the three most common family structures found in the total sample of over 2500 cases. The students were between 14 and 16 years of age. The participants were also differentiated by gender, and stratified by two levels of mother's education, with completion of high school as the dividing line. The figure shows the results only for the high school graduates.

Parental monitoring was defined as 'the effort by parents to keep informed about and set limits on their children's activities outside the home.' Levels of parental monitoring, ranging from 0 to 12, are shown on the horizontal axis, and grade point average (GPA) appears on the vertical scale. The markers to the right of each curve record the average GPA for each of the six groups.

The data were fitted to a curve with a declining slope, rather than simply with a straight line, based on the following considerations derived from the bioecol-

ogical model: Higher levels of academic performance require mastery of progressively more complex tasks, and hence are more difficult to achieve. As can be seen in Fig. 1, the relation between parental monitoring and school grades does indeed show a declining curvilinear trend. This effect, however, is far stronger for girls than for boys, particularly in the most advantaged ecological niche—families with two biological parents in which the mother had some education beyond high school.

Both of these results are consistent with two working hypotheses derived from the bioecological model. The first stipulates that proximal processes (in this instance, parental monitoring) are likely to have greater impact in two-parent families than in those in which the mother is a single-parent or the father is a step-parent. The second hypothesis posits a stronger and longer-lasting influence of the family on the development of females than of males.

In addition, a distinctive feature of the pattern for girls is a marked flattening of the curve at higher levels of monitoring, and, in the case of daughters of single-parent mothers, even becomes a turnaround. This finding suggests that under such circumstances, the demands on the girls may become so great that the existing proximal processes are not equal to the task, and as a result, bring fewer educational returns. Finally, an analysis (not shown) of data on students whose mothers had no more than a high-school education revealed a similar pattern, but the constructive influence of monitoring was appreciably weaker, and its greater benefit to girls was reduced. Nevertheless, daughters of mothers with less than a high-school education both in single parent and in stepfather families still had higher GPA scores than did sons. Moreover, within each level of mothers' education statistically significant differences by family structure were found in school achievement, with students growing up in two-parent families receiving the highest grades and those from single-parent families the lowest.

To turn to a related issue: although proximal processes function as the engines of development, the energy that drives them comes from deeper sources that take us back to the experiential world of Proposition I (Bronfenbrenner and Evans 2000, Bronfenbrenner et al. 1996, Bronfenbrenner and Morris 1998). Both the subjective and the objective forces exert an especially strong influence on development during the formative years (from early infancy to young adulthood). A substantial body of research over the past century indicates that, two or three decades ago, these forces lay mainly within the family, with parents acting as the principal caregivers and sources of emotional support for their children, and with other adult family members living in the home being next in line. To a lesser extent, other relatives, family friends, and neighbors also functioned in this role.

However, there has been a marked change in this pattern over the past three decades. Parents, and other adult family members as well, have been spending increasing amounts of time commuting to and working at full-time jobs (in which overtime is increasingly often required or expected).

The nature of this trend and its relevance for human development are conveyed in the Propositions that follow. (For brevity's sake, the term 'child' is used below to encompass the entire period from infancy through young adulthood.)

1.4 Proposition IV

In order to develop—intellectually, emotionally, socially, and morally—a child requires, for all of these, the same thing: participation in progressively more complex activities, on a regular basis over an extended

period of time in the child's life, with one or more persons with whom the child develops a strong, mutual emotional attachment, and who are committed to the child's well-being and development, preferably for life (Bronfenbrenner and Evans 2000, Bronfenbrenner and Morris 1998). The prerequisites stipulated in Proposition IV then lead to the developmental consequences described in the next Proposition.

1.5 Proposition V

The establishment of a strong mutual emotional attachment leads to internalization of the parent's activities and expressed feelings of affection. Such mutual ties, in turn, motivate the child's interest and engagement in related activities in the immediate physical, social, and—in due course—symbolic environment that invite exploration, manipulation, elaboration, and imagination. The next Proposition broadens the family circle.

1.6 Proposition VI

The establishment and maintenance of patterns of progressively more complex interaction and emotional attachment between parent and child depend to a substantial degree, on the availability and involvement of another adult, a third party, who assists, encourages, spells off, gives status to, and expresses admiration and affection for the person caring for and engaging in joint activity with the child. It also helps, but is not absolutely essential, that the third party be of the opposite sex from that of the other person caring for the child, because this is likely to expose and involve the child in a greater variety of developmentally-instigative activities and experiences (Bronfenbrenner et al. 1996). Where this is an attachment to two or more parent figures, each can serve as a third party to the other.

The research evidence for this Proposition came mainly by default. It was produced by demographic data documenting a rapid rise in the proportion of single-parent households. The trend began in the 1980s, and then continued at an even faster rate through most of the 1990s. The overwhelming majority of such homes were those in which the father was absent and the mother bore primary responsibility for the upbringing of the children.

A large number of investigations of developmental processes and outcomes in families of this kind have since been conducted across a range of cultural and social class groups. In general, the findings lead to two complementary conclusions:

(a) Even in families living in favorable socio-economic circumstances, children of single-parent mothers, or of fathers, for whom no other person is

acting reliably in a 'third party' role are at greater risk for experiencing one or more of the following developmental problems such as: hyperactivity or withdrawal; lack of attentiveness; difficulty in deferring gratification; poor academic achievement; school misbehavior; and frequent absenteeism.

(b) At a more serious level, there is the so-called 'teenage syndrome' of behaviors that tend to be associated together: dropping out of school; involvement in socially-alienated or destructive peer groups; smoking, drinking, frequent sexual experience, adolescent pregnancy, a cynical attitude toward work, and—in the more extreme cases—drugs, suicide, vandalism, violence, and criminal acts. Most of these effects are more pronounced for boys than for girls (Bronfenbrenner et al. 1996).

Not all single-parent families, however, exhibited these disturbed relationships and their disruptive effects on children's development. Systematic studies of the exceptions have identified what may be described as a general 'immunizing' factor. For example, children of single parents were less likely to experience developmental problems especially in families in which the mother (or father) received strong support from other adults living in the home. Also helpful were nearby relatives, friends, neighbors, members of religious groups, and, when available, staff members of family support and child care programs. What mattered most was not only the attention given to the child—important as this was—but also the assistance provided to the single parent or by others serving in the supportive roles cited in Proposition VI. It would seem that, in the family dance, 'it takes three to tango.'

But dancing is not the whole story. By the 1980s, theory and research in the ecology of human development had documented an accelerating trend toward greater permissiveness in styles of child rearing in American families. At the same time, successive scientific investigations had revealed progressively greater developmental advantage for strategies that placed increased emphasis on parental discipline and demand. The interpretation that emerged from analyses of the available data suggested that widespread application of these research findings would serve as an effective response to the developmentally disruptive changes taking place in contemporary society.

At this point, it is important to mention two other bodies of research that contributed significantly to the development of bioecological theory and its corresponding research designs. The first is now of long standing. In 1974 the sociologist Glen H. Elder, Jr., in his classic volume *Children of the Great Depression* (1974, 1999), extended the concept of development beyond the formative years to encompass the entire life course.

The second addition has yet to be fully exploited. Bronfenbrenner and Ceci (1994), taking the bioecological model as their point of departure, suggested an empirically testable alternative for the established

scientific paradigm used in behavior genetics. The proposed alternative model: (a) allows for nonadditive synergistic effects; (b) employs direct measures of the environment; and (c) proposes proximal processes as mechanisms of person–environment interaction through which genotypes for developmental competence are transformed into phenotypes. The model predicts that: (a) estimates of heritability (h^2) for developmental competence increase markedly with the magnitude of proximal processes; (b) heritability measures the proportion of variation in individual differences that are attributable only to 'unactualized' genetic potential, with the degree of 'unactualized' potential remaining unknown; (c) actualized genetic potential (h^2) will vary with the quality of the environment, and will increase as that quality is improved (for example, through providing job opportunities, health services, and intervention programs in low-income neighborhoods).

The authors also suggested that high levels of such patterns of parental behavior as 'neglect, abuse, or domination' could serve as powerful mechanisms for actualizing genetic potentials for developmentally maladaptive behaviors that both disrupt proximal processes and produce developmental disarray.

1.6.1 The bioecological model in the discovery mode: future perspectives. This section is based on Propositions and working hypotheses derived from the bioecological model for which, as yet, there are few empirical data. The section begins with future prospects for addressing the second stated goal of the bioecological model: that of 'providing needed scientific bases for the design of effective social policies and programs that can counteract newly-emerging developmentally-disruptive influences.'

In accord with the latter objective, the section begins with an unorthodox proposal. By and large, thus far, theory and research on human development have been concerned with the influence of the older generation on the development of the younger. In the Proposition that follows, the direction is reversed. (It should be noted that the basic idea underlying this Proposition is not new, and is foreshadowed both in the theory of Vygotsky (1978), and in the contemporary 'action' theory of Brandtstädter (1998, 1999).)

1.7 Proposition VII

The psychological development of parents is powerfully influenced by the behavior and development of their children. This phenomenon occurs through the life course; is more evident during the formative years when most children are living at home in the care of their parents; and often becomes especially pronounced during adolescence, when the young begin to

strive for independence both as individuals and as members of peer-groups. Such behavior is particularly likely to occur amongst those adolescents or youth who have comparatively little contact with their parents or other caring adults earlier in life. Although many studies have focused on the development of such alienated young people, the impact of the latter's behavior on the subsequent development of their parents has yet to receive the systematic investigation that it deserves. The converse of the foregoing Proposition—the influence of the successful transition of children through adolescence and young adulthood on the constructive development of their parents—has regrettably received even less scientific attention.

1.8 Proposition VIII

Over the life course, the process of attachment exhibits a turnaround. In the beginning, it is the children who are the beneficiaries of the parents' irrational commitment, whereas toward the end the roles are reversed. Then it is the elderly parents who receive the love and care of their now middle-aged children. If, however, there was no attachment at the beginning, there may be no attachment at the end.

In this regard, developmental science has yet to address a curious omission both with respect to theory and research design. Resort to search engines in psychology and related fields has thus far failed to identify any investigations of the influence of parent-child attachment on the future development of the parent in contrast to that of the child.

This is not quite the case, however, for the next proposition in the discovery mode. The theoretical model, the corresponding research design, as well as half of the necessary empirical data are already available. The only problem is to find or conduct a study that meets the requirements of this model.

1.9 Proposition IX

If an investigation conducted in the past meets the requirements of the bioecological model, including assessment of developmental outcomes 'over an extended period of time,' then replication of the study at a later point in time would reveal whether the processes under investigation were still valid or had been nullified or superseded by subsequent historical changes. When the latter occurs, the investigator is confronted with the challenge of proposing alternative working formulations for explaining the observed phenomena.

At the conclusion of this article, we move from the domain of theory and research design to the world of reality and action. In the bioecological model, these two worlds have never been far apart. Especially from

1970 to 2000, these two worlds have become ever closer to each other. At a more general level, the findings from both domains reveal what has been referred to as 'growing chaos' in the lives of children, youth, families, schools, the world of work, and the ever-greater commuting in between (Bronfenbrenner and Evans 2000, p. 1022). The most recent report of this phenomenon appears in an article by Bronfenbrenner and Evans (2000). It contains the following summary about the nature of 'chaos' and its developmental consequences.

'Chaos integrates the various elements involved, and foreshadows the role [of chaos] in the bioecological model in terms of what is called 'chaotic systems.' Such systems are characterized by frenetic activity, lack of structure, unpredictability in everyday activities, and high levels of ambient stimulation. Background stimulation is high, and there is a general lack of routinization and structure in daily life. The environment is also a major source of interruption of proximal processes in the form of residential noise, crowding, and classroom design' (p. 121).

At the turn of the century, we are left with a troubling question: From the perspective of the bioecological model, what is the prospect for the future development of our species? The answer to that question lies with the willingness of the USA and other economically developed countries to heed the emerging lessons of developmental science. At the moment, it is difficult to know what the answer will be. The future could go either way.

Given this alternative, then surely it becomes the responsibility of developmental science to communicate such knowledge as we possess, and to do so in words that can still find an echo. Here is a first draft.

In the USA it is now possible for a youth, female as well as male, to graduate from high school, or university, without ever caring for a baby; without ever looking after someone who was ill, old, or lonely; or without comforting or assisting another human being who really needed help. The developmental consequences of such a deprivation of human experience have not as yet been scientifically researched. But the possible social implications are obvious, for—sooner or later, and usually sooner—all of us suffer illness, loneliness, and the need for help, comfort, and companionship. No society can long sustain itself unless its members have learned the sensitivities, motivations, and skills involved in assisting and caring for other human beings.

See also: Adulthood: Developmental Tasks and Critical Life Events; Biopsychology and Health; Development: Social-anthropological Aspects; Developmental Sciences, History of; Genetics and Development; Human Development and Health; Human Evolutionary Genetics; Life Course in History; Life Course: Sociological Aspects; Lifespan

Development, Theory of; Lifespan Theories of Cognitive Development; Observational Methods in Human Development Research; Plasticity in Human Behavior across the Lifespan; Psychological Development: Ethological and Evolutionary Approaches; Vygotskij's Theory of Human Development and New Approaches to Education

Vygotsky L S 1978 *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA

U. Bronfenbrenner

Bibliography

- Baltes P B, Staudinger U M, Lindenberger U 1999 Lifespan psychology: Theory and application to intellectual functioning. *Annual Review of Psychology* **50**: 471–507
- Bergman L R, Cairns R B, Nilsson L G, Nysted L (eds.) 2000 *Developmental Science and the Holistic Approach*. Lawrence Erlbaum Associates, Mahwah, NJ
- Brandtstädter J 1998 Action perspectives on human development. In: Lerner R M (eds.) *Theoretical Models of Human Development. Handbook of Child Psychology*, 5th edn. Wiley, New York, Vol. 1, pp. 807–63
- Brandtstädter J 1999 The self in action and development: Cultural, biosocial, and ontogenetic bases of intentional self-development. In: Brandtstädter J, Lerner R M (eds.) *Action and Self-Development: Theory and Research Through the Life Span*. Sage Publications, Thousand Oaks, CA, pp. 37–65
- Bronfenbrenner U, Ceci S J 1994 Nature-nurture reconceptualized in developmental perspective: A bio-ecological model. *Psychological Review* **101**(4): 568–86
- Bronfenbrenner U, Evans G W 2000 Developmental science in the 21st century: Emerging questions, theoretical models, research designs, and empirical findings. *Social Development* **9**(1): 115–25
- Bronfenbrenner U, Morris P A 1998 The ecology of developmental processes. In: Lerner R M (ed.) *Handbook of Child Psychology Theory*, 5th edn. Wiley, New York, Vol. 1, pp. 993–1028
- Bronfenbrenner U, McClelland P, Wethington E, Moen P, Ceci S J 1996 *The State of Americans: This Generation and the Next*. Free Press, New York, pp. 150–88
- Cairns R B 1970 Towards a unified science of development. *Contemporary Psychology* **15**: 214–15
- Cairns R B, Cairns B D 1994 *Lifelines and Risks: Pathways of Youth in our Time*. Cambridge University Press, Cambridge, UK
- Cole M 1995 Culture and cognitive development: From cross-cultural research to creating systems of cultural mediation. *Culture and Psychology* **1**: 25–54
- Elder G H, Jr 1974 *Children of the Great Depression*. University of Chicago Press, Chicago
- Elder G H, Jr 1998 The life course and human development. In: Lerner R M (ed.) *Handbook of Child Psychology. Theoretical Models of Human Development*, 5th edn. Wiley, New York, Vol. 1, pp. 939–91
- Elder G H 1999 *Children of the Great Depression* (25th Anniversary edn.). University of Chicago Press, Chicago
- Magnusson D, Stattin H 1998 Person-context interaction theories. In: Lerner R M (ed.) *Handbook of Child Psychology. Theoretical models of Human Development*, 5th edn. Wiley, New York, Vol. 1, pp. 685–759
- Small S, Luster T 1990 Youth at risk for parenthood. Paper presented at the *Creating Caring Community Conference*. Michigan State University, East Lansing, MI

Human Development, Successful: Psychological Conceptions

From its outset human development is dependent on the presence of essential and beneficial conditions in the physical and social environment. Some of these conditions are fundamental and the same for all human beings (oxygen, water, food), while others are more complex and necessary to varying degrees (e.g., support of caregivers, language education, culturally specific competencies). Human beings also differ from one another in their psychological and physical needs: congenital and acquired vulnerability and deficiencies need special support and encouragement, some of them may even require specialist intervention or preventive treatment (e.g., specialized diets in the case of inherited metabolic disturbances such as phenylketonuria). At any event, there needs to be a minimum degree of fit between the individual and his or her environment before development can actually take place. Within these bounds, however, a very wide range of human development can unfold: An individual's predispositions and qualities encounter an immeasurably vast range of diverse environmental influences (socialization, critical and non-normative life events, historical influences, etc.). These influential factors operate throughout an individual's ontogenesis (Baltes et al. 1998). Human development never reaches completion, is multidirectional and seldom irreversible, while at the same time features a high degree of interindividual variability and intraindividual plasticity (Baltes 1987, Baltes and Baltes 1990). Moreover, there is now a broad consensus that human development does not take place uniformly in an irreversible sequence of different stages and phases, but is a process that individuals themselves actively influence in substantial ways (Brandtstädter 1998).

In view of this broad range of possibilities, the question of how and under what circumstances human development can succeed proves to be complex in several respects. The purpose of this chapter is to deal with two questions. The more important, of course, is the question of what makes development successful: which are the prerequisites of an 'art of life' (Staudinger 1999)? To answer this question, however, the term 'successful' needs to be clarified.

1. What is 'Successful' Development? The Problem of Criteria

Since the work of David Hume (1978), many have argued that while empirical sciences like psychology may, for reasons of principle, apply criteria which are value judgments, they cannot justify or legitimize the use of such criteria. According to this view, no matter how good science may be, it can only describe and explain the world, but it cannot judge it. Is the question of 'successful' development hence unanswerable from the start, at least from the viewpoint of an empirical science?

Actually, the relationship between empirical description and explanation on the one hand and value judgments or normative standards on the other is in fact more complex than that. Even if empirical facts do not constitute sufficient conditions for making evaluations, as a rule they are nevertheless necessary ones (Brandtstädter and Schneewind 1977). Any decision or action is predicated not only upon preferences and value judgments but also upon expectations which take account of empirical facts. For instance, the demand for a juvenile who commit crimes to be 'educated' by a punishment assumes that punishment actually serves to influence his future actions. This is an assumption that can, and indeed must, be empirically tested. In this respect science plays a key role in answering the question of successful development, even if the criterion for measuring 'success' has had to be determined on a prescientific basis.

Yet it is open to question whether assessments of human action and biographies really are as completely arbitrary as is suggested by a Humean perspective. Without wishing to become involved in the finer points of the philosophical debates surrounding the 'deducibility' of normative statements from descriptive ones, it can nevertheless be said that all positions purporting to offer proof of the inherent distinction between 'is' and 'ought' in fact tend to presuppose this 'gap' of principle rather than to prove its existence (Mavrodes 1968). In search for firm foundations for value judgments as well as for empirical descriptions, the last line of defense is the practice of simply not calling certain statements into question because the grounds for doubting them appear more dubious than the statements themselves.

So although we can forget any prospect of finding 'the' best biography *per se* in the foreseeable future, indisputably negative examples can easily be identified within the variations realized. For instance, there is no doubt that a person who persistently and repeatedly violates others for no reason, leads a less 'productive' life than one who shows reasonable consideration for others' interests and needs when taking his or her actions and decisions. Likewise, one can hardly describe an old age marred by illness and loneliness, causing distress particularly to the aged person but also to his or her relatives and friends, as 'successful'

without arousing the suspicion that this is an artificial, academic discussion or simply cynicism. The details of examples such as these will be controversial, and here too particular constellations will lead to changes in our assessments. Nevertheless, such examples show that in many cases value judgments ('better'-'worse') are scarcely more doubtful than the 'brute' facts.

However, the search for universal criteria of successful development which can be applied to all people irrespective of their individual, cultural and historical circumstances, is likely to be fruitless, because it follows from the actional perspective of human development that people's 'success' depends to no small degree on what they themselves consider to be successful. Within the development-regulating normative rules and developmental tasks of a culture, which incidentally are themselves subject to cultural development processes, the individual organizes ways of shaping his or her own life, looking for and construing orientations which give meaning to his or her personal development (Brandtstädter 1998).

These considerations point toward a solution to the question of criteria. Beyond the fulfillment of the fundamental conditions essential for life, the prerequisites for successful development are clearly those meta-competencies and meta-resources which ensure that individuals can make use of opportunities and encouragement, can compensate for deficiencies and vulnerabilities, can choose and realize their own developmental goals, and thus can actively shape their development within the framework defined by immediate and potential possibilities (Brandtstädter and Schneewind 1977). This metaperspective implies a multicriteria approach on the one hand, and the integration of subjective and objective criteria of 'success' on the other (Baltes and Baltes 1990).

2. What Makes Development Successful?

If human development is basically the product of culture and, at the same time, the result of individual action, an essential prerequisite for successful development is the capacity to act, i.e., to choose developmental goals, to make decisions, to form intentions and to put them into practice, and also to overcome resistance and to revise strategies where necessary. Acting successfully entails first the precise competencies needed to carry out a particular action in practice. Additionally, an adequate representation of these competencies and potentials in the person's self-concept is a prerequisite for an appropriate and autonomous choice to be made from among the possible developmental options. Thus, processes of self-regulation perform an integrating, 'orchestrating' function for human development (Staudinger et al. 1995).

Self-regulating processes can essentially be classified into two modes (Brandtstädter and Greve 1994). In the assimilative mode, goals are maintained even

against resistance, and strategies are pursued tenaciously until they are achieved. Obviously, development can only succeed if people are sufficiently willing and capable; anyone who always gives up immediately will achieve neither essential tasks nor those of his or her choosing. In the accommodative mode, when faced with goals that are truly or are perceived to be unachievable, or with irreversible losses and deterioration, preferences and orientations are adapted in a way that avoids meaninglessness and depressive tendencies, and preserves subjective well-being, but at the same time brings new, achievable goals into view, which revive the capacity and motivation to act. These aspects are illustrated with examples from the research fields of juvenile delinquency and successful aging.

2.1 Autonomy and Social Integration: The Sample Case of Juvenile Delinquency

Although difficulties of coping with developmental tasks occur throughout the lifecourse, juvenile delinquency serves as a good example with which to illustrate such difficulties. Challenges such as physical changes and sexual maturation, the need to develop interpersonal skills (e.g., for relationships with the opposite sex), acquiring education and, in particular, forming a personal and social identity are not just the storms to be weathered during adolescence; these challenges are vitally bound up with laying the groundwork for successfully solving the tasks and meeting the demands of adult life (Crockett and Crouter 1995). In particular, social developmental tasks play a particularly prominent role during the years of youth (Petersen and Leffert 1995). Especially at a juvenile age, the dualism involved in social integration is patently evident: at one and the same time, social integration is an aim in itself and a prerequisite for successfully fulfilling the developmental task of preparing adulthood. However, both the necessity and the opportunity to change and challenge social contexts are characteristic features of adolescence too. Since the dividing line between creatively transgressing the social rules and a deviant or pathological disregard for social limits can be a fuzzy one, these dilemmas of individuation vs. integration and of autonomy vs. adaptation may serve as a background for the explanation of juvenile delinquency (Coie and Dodge 1998). Delinquent juveniles, in particular those whose deviance is 'adolescence-limited' (Moffitt 1993), have evidently not managed to cope successfully with these tasks (Petersen and Leffert 1995). Thus, personal resources of coping with discrepancies between the social demands on the one hand and the personal goals, desires and competencies on the other gain a key function in explaining juvenile delinquency and, as a consequence, successful juvenile development. In particular, the goodness of fit between

the individual and his or her social environment thus becomes an important precondition and, at the same time, a central criterion for developmental success of adolescence.

2.2 Resources of Resilience: The Sample Case of Successful Aging

Going yet another step further, human development in general can be regarded as a hierarchical, interactive sequence of micro- and macroscopic coping processes. While current coping processes and resources are shaped by developmental conditions, coping reactions can also act as forces in creating future development. That leads on to the investigation of developmental constellations of individual and social resources that make an individual either resilient against or vulnerable to new strains as they arise. The discussion of the conditions under which people cope successfully with aging illuminates this aspect very well. Numerous studies have shown that people's ability to function and their sense of well-being are not greatly impaired as they grow old, if at all. Indicators of overall psychological status such as self-esteem, general sense of well-being, or being convinced of having control over one's own life are variables which, taken together, do not show any substantial covariance with age (e.g., Brandtstädter et al. 1993, Staudinger et al. 1995). As a consequence, the stereotype of aging being a process strongly characterized by deficiencies and loss has gradually gone out of favor as being too one-sided and incomplete, to be superseded by a more multifaceted view of 'productive' and 'successful' aging (Baltes and Baltes 1990). These findings are most surprising when set against the well-documented fact that, whether in terms of cognitive development, changing sensory functions or morbidity, aging does entail a real increase in the losses and impairment people suffer (Bond et al. 1995).

This apparent contradiction between the growing strains and losses on the one hand and the stability of the self and well-being on the other is resolved once the personal constellations of 'resilience' (Rutter 1990) are made the focus of the theoretical perspective (Baltes and Baltes 1990, Brandtstädter 1998, Staudinger et al. 1995). In addition to the model of developmental self-regulation already mentioned (Brandtstädter et al. 1993), the main model providing important insights that help to explain resilience in old age is that of 'selective optimization with compensation' (Baltes and Baltes 1990). As losses occur, they can be partly cushioned or alleviated by way of selective improvement (training, expertise), and compensated for as people reorient their lives or change their strategies. Against the background of the actional conception of development, it is important to point out that resilience is not just a reactive quality, but is often brought to bear proactively, before the troubling event

actually occurs or the degenerative tendency sets in. For instance, the fact that people have less chance of becoming crime victims as they grow older is partially explained by an increasing tendency to behave more cautiously (Greve 1998).

3. Conclusion: Courses and Conditions of Successful Development

In terms of cognitive functions, social integration, and self-regulation, human development features a considerable amount of plasticity (Baltes et al. 1998), which not only makes up the core prerequisite for the dynamics of resilience and the coping processes discussed here, but indeed is also a general precondition for the very possibility of lifespan development. Self-regulation processes are both a central criterion for adaptation processes on the one hand and they substantially control them on the other.

Personal well-being and psychological health, and social integration can only be maintained and improved if the individual can be protected against unfavorable conditions by resilient resource constellations, if personal and social coping resources can help to cushion and compensate for the crises and losses that will inevitably occur, and if the individual capacity to act is largely maintained, whether by stabilizing the required conditions or by finding a new orientation toward achievable objectives. Actually, all of these aspects can be seen both as the products and the producers of successful development.

Development essentially consists in upholding and implementing the individual's capacity to adapt to the discrepancies in new situations and developmental tasks as he or she seeks (via assimilation or accommodation) to attain a solution to, or dissolution of, the discrepancies that are generating a crisis or creating a burden, if these are unavoidable in the long run. Thus, successful development implies a progressive adaptation which simultaneously maintains or indeed expands this plasticity and adaptivity. As far as we know today, this is possible—and usual—at least up to the latest years of very old age (Baltes 1997). That leads to the somewhat tautological-sounding statement that successful development essentially means ensuring that development, i.e., progressive adaptation, will always be possible and really will occur.

The emerging perspective of adaptivity and activity as the roots of successful human development suggests essential new research perspectives focusing on sources of resiliency and vulnerability at any stage of human development. Moreover, even if the preconditions and processes are known which provide successful development, the ways to improve the individual's capacity to acquire or to use them are still to be investigated. Thus, studies in applied developmental science are needed in order to discover techniques and

strategies of individual and social improvement of both of these core conditions, with respect to special interventions, e.g., in the case of juvenile delinquency, as well as with respect to primary prevention of developmental crises, e.g., in the case of successful aging in very old age.

Bibliography

- Baltes P B 1987 Theoretical propositions of life-span developmental psychology: On the dynamics between growth and decline. *Developmental Psychology* 23: 611–26
- Baltes P B 1997 On the incomplete architecture of human ontogeny. *American Psychologist* 52: 366–80
- Baltes P B, Baltes M M 1990 Psychological perspectives on successful aging: The model of selective optimization with compensation. In: Baltes P B, Baltes M M (eds.) *Successful Aging. Perspectives from the Behavioral Sciences*. Cambridge University Press, New York
- Baltes P B, Lindenberger U, Staudinger U M 1998 Life-span theory in developmental psychology. In: Lerner R M (ed.) *Handbook of Child Psychology: Theoretical Models of Human Development*, 5th edn. Wiley, New York, Vol. 1, pp. 1029–143
- Bond L A, Cutler S J, Grams J (eds.) 1995 *Promoting Successful and Productive Aging*. Sage, Thousand Oaks, CA
- Brandtstädter J 1998 Action perspectives on human development. In: Lerner R M (ed.) *Handbook of Child Psychology: Theoretical Models of Human Development*, 5th edn. Wiley, New York, Vol. 1, pp. 807–63
- Brandtstädter J, Greve W 1994 The aging self: Stabilizing and protective processes. *Developmental Review* 14: 52–80
- Brandtstädter J, Schneewind K A 1977 Optimal human development: Some implications for psychology. *Human Development* 20: 48–64
- Brandtstädter J, Wentura D, Greve W 1993 Adaptive resources of the aging self: Outlines of an emergent perspective. *International Journal of Behavioral Development* 16: 323–49
- Coie J D, Dodge K A 1998 Aggression and antisocial behavior. In: Eisenberg N (ed.) *Handbook of Child Psychology: Social, Emotional, and Personality Development*. Wiley, New York, Vol. 3, pp. 779–862
- Crockett L J, Crouter A C 1995 Pathways through adolescence: An overview. In: Crockett L J, Crouter A C (eds.) *Pathways Through Adolescence*. L Erlbaum Associates, Mahwah, NJ, pp. 1–12
- Greve W 1998 Fear of crime among the elderly: Foresight, not fright. *International Review of Victimology* 5: 277–309
- Hume D 1978 [1739–40] *A Treatise of Human Nature*. Clarendon Press, Oxford, UK
- Mavrodes G I 1968 On deriving the normative from the nonnormative. *Papers of the Michigan Academy of Science, Arts, and Letters* 53: 353–65
- Moffitt T E 1993 Adolescence—limited and life-course—persistent antisocial behavior: A developing taxonomy. *Psychological Review* 100: 674–701
- Petersen A C, Leffert N 1995 What is special about adolescence? In: Rutter M (ed.) *Psychosocial Disturbances in Young People*. Cambridge University Press, Cambridge, UK, pp. 3–36
- Rutter M 1990 Psychosocial resilience and protective mechanisms. In: Rolf J, Masten A S, Cicchetti D, Nuechterlein K H, Weintraub S (eds.) *Risk and Protective Factors in the Development of Psychopathology*. Cambridge University Press, Cambridge, UK, pp. 181–214

Staudinger U M 1999 Social cognition and a psychological approach to an art of life. In: Blanchard-Fields F, Hess T (eds.) *Social Cognition and Aging*. Academic Press, New York, pp. 343–75

Staudinger U M, Marsiske M, Baltes P B 1995 Resilience and reserve capacity in later adulthood: Potentials and limits of development across the life span. In: Cicchetti D, Cohen D J (eds.) *Developmental Psychopathology*. Wiley, New York, Vol. 2, pp. 801–47

W. Greve

Copyright © 2001 Elsevier Science Ltd.

All rights reserved.

Human Ecology

Human ecology addresses the relationships of humans to their environments. This demarcation concept unites an array of otherwise unconnected scholarly traditions. The traditions are distinguished by their different interpretations of ecology itself, by the different theory frames they employ for the study of human ecology, and by their respective objects of explanation or problematics. This article begins with the different interpretations as a preview for sorting and examining the different theory frames and their problematics.

1. Interpretations of Ecology

Interpretations of the concept and substance of ecology differ according to the unique vantage points provided by independent scholarly traditions.

1.1 The Biological Interpretation

Ecology was originally given a biological interpretation in 1866 when Ernst Haeckel defined the science in accord with the work of Darwin, who had attempted to explain evolution in terms of the adaptations of organisms to the biotic and abiotic features of their environments. A generation later another foundation concept appeared: since evolutionary studies indicated there was an ecology of living organisms before humans evolved, and humans themselves are living organisms, there must therefore be a human ecology. On the biological interpretation, humans are part and parcel of the natural world, and are subject to the same biophysical dynamics that govern the organization of other living communities. For most of the twentieth century, however, although gradually shifting in the last couple of decades, biologists took only a passing interest in the scientific study of the ecology of humans *per se*.

1.2 The Metaphorical Interpretation

Social scientists, by contrast, were quick to seize on the idea of a human ecology with independent investigations begun in the 1920s, in India, by Radhakamal Mukerjee, who followed the biological interpretation, and in the USA by Robert Park, whose students did not. Aware of theories of community succession that their biological colleagues were then developing for natural communities, in which society was used as a metaphor to understand natural organization, some early US sociologists came to use nature as a metaphor to understand social organization. Setting aside the rhetorical uses of ‘ecology’ as a synonym for ‘contextual’ or ‘environmental,’ two variants of metaphorical interpretation can now be found in social science. The first applies bioecological concepts to the study of human communities *per se*, reasoning by analogy. Here, institutional or organizational settings are taken as contexts or environments for humans within their groups, populations, or societies, which are supposed to be organized parallel to natural communities. This variant generally overlooks the fact that natural communities contain multiple species and humans are a single species, preferring to finesse this by elaborating biological species and evolutionary analogies for societal divisions of labor. The second variant, discussed in Sect. 2.3, treats ecological knowledge as a social construct. Left unaccounted for in the metaphorical interpretation are biological connections between human societies and other natural systems. The metaphorical interpretation is a blueprint for the study of the internal workings of society through the adaptations of its populations or groups by means of their social organizations and shared cultural knowledge (see *Biology's Influence on Sociology: Human Sociobiology*).

1.3 The Ideological Interpretation

With the reawakening of the modern environmental movement worldwide and its growth in the last decades of the twentieth century, an ideological interpretation of ecology has appeared. In this, the biological and metaphorical interpretations of ecology are freely interwoven and supplemented with value judgments gauged against standards supposedly congruent with ecological science. Some versions of the ideological interpretation use nature as an ideal to evaluate and reform social and political organization, economic practice, and personal conduct, while others try to frame environmental ethics so as to integrate ecological science with politics, economics, philosophy, education and various religions and cultures. The ideological interpretation is neither monolithic nor uniform, but its advocates generally fall to the left of ideological spectra. The common thread is that ecology is more than science or metaphor: It is a world view with ethical implications for personal, cultural,

and institutional change (see, for example, Evernden 1992, Oelshlaeger 1994, Orr 1992).

2. Theory Frames for Human Ecology

The multifarious research traditions of human ecology may be understood in terms of the different theory frames they have employed, which reflect their respective interpretations of ecology. Different theory frames address different problematics.

2.1 The Evolutionary Theory Frame

The evolutionary theory frame assumes that ecological forces act on genetic and cultural variation among individuals and populations so as to effect a selective survival and reproduction. Using this assumption, neo-Darwinian principles of natural selection are applied to understand how individuals adapt to their natural environments, how different social species reflect common principles of social organization in response to ecological contingencies, and how populations and cultures are affected by pathogenic disease.

Individual behavior and social organization are classic problematics in social science. The application of evolutionary theory attempts to provide a fuller explanation of these. One strand of theory develops a human evolutionary ecology patterned on the behavioral and evolutionary ecology of the biological sciences. In this, an adaptive design is assumed to be the product of the elemental behavior of individuals, such as foraging, territorial defense, time allocation and mate selection, all of which are presumed to satisfy economic cost-benefit principles (Smith and Winterhalter 1992). It has yet to be demonstrated whether neo-Darwinian evolutionary-economic models have application beyond the case of hunter-gatherers. Because organic traits (such as brain size, bipedalism, and vocal chords) underlie cultural capabilities, and culture is understood as a human adaptation, some evolutionary models assume the co-evolution of biological and cultural traits in ecological context (Boyd and Richerson 1985, Durham 1991). Another strand of theory focuses on problematics of social organization among different species, with a view to discovering common principles but without presuming that the properties of social organization can be reduced to the summed behavior of individuals. In addition to comparisons with social primates, a research tradition of long standing, Machalek's (1999) transpecific sociology enables the comparison with humans of the demographic constraints, division of labor, and homeostatic mechanisms that appear in colonies of social insects, and of common strategies of environmental exploitation across social species (see *Cultural Evolution: Overview; Human Ecology*).

Problematics of epidemiology, ancient, and contemporary, fall within an evolutionary theory frame inasmuch as most scholars consider pathogens to be the principle agents of natural selection operating on humans since sedentary lifestyles were adopted 10,000–12,000 years ago. The most startling and horrible example of the human ecology of disease is the cultural and demographic extirpation of untold millions of native Americans in times of post-Columbian exploration, due to epidemic pathogens imported by Europeans and their African slaves, and from which the tribes had no inherited or acquired immunities. In modern times, medical authorities are again sounding alarms about the development and spread of epidemic pathogens, old and new, because normal genetic mutations eventually produce drug-resistant variants of old as well as new, and now dangerous, variants of formerly benign organisms. Since microbic pathogens are part of the living environment of all humans, human culture and society have to adapt to ever-changing conditions of pathogenic disease just as pathogens adapt to humans—an inherently ecological condition. The incidence of pathogenic diseases is known to be affected by agricultural practices, dietary customs, migration patterns, population densities, climate alteration, cultural hygiene practices, lack of potable water, sexual mores, modern transportation that transcends normal environmental barriers, and technological methods of intervention (see Ewald 1994, Kipple 1997).

Intervention methods typically disrupt population equilibria that might normally exist between pathogens and humans, thus preventing natural environmental controls from exerting themselves. Instead of stabilizing natural environments, technology implemented through the social organization of medicine and public health rather attempts to control pathogens so as to stabilize their relationship to humans. This method in itself, by intervening to change natural biota, provides a stimulus for microbic evolutionary adaptation. The general principle is that changed environments bring changed frequencies of microbic pathogens, human intentions notwithstanding. Some scholars (Mascie-Taylor 1993) have tended to conceive human ecology entirely in these epidemiological terms, however, this has not been predominant.

2.2 The Environmental Theory Frame

The environmental theory frame dominates modern research traditions of human ecology. It too presumes a biological interpretation of ecology, but in some cases substantially supplemented by considerations of social science targeted to problematics of subsistence and sustainability.

Anthropologists and geographers have a long tradition of investigating how human behavior, culture, traditions, settlements, built environments, spatial

location, territory, and humans' sense of place are affected by efforts to procure subsistence from natural—typically regional—environments. Research programs investigating cultural ecology, ecological anthropology, environmental geography, and environmental history consistently demonstrate how the development of culture and society vary with settlement patterns, technology and means of material resource acquisition, and how natural environments are modified by sociocultural variations in subsistence practices. A cultural materialism, which viewed culture and its symbolism as epiphenomenal, was implicit or explicit in many of these classic approaches, as was a dualistic construal of nature and culture. Modern approaches, especially political ecology, are different (see Biersack 1999, Scoones 1999). They focus on the interaction of spatial and temporal environmental dynamics with people situated in political, social, institutional, and cultural context, and emphasize the effects of human agency within the push-pull parameters of social, economic and political inequalities, on the one hand, and local or regional environmental transformations, on the other. Political ecology conceives human-environment interactions as mutual, thus linking human ecology to political economy and crossing from problematics of subsistence to those of sustainability (see *Feminist Political Ecology*).

No problematic has so dominated modern environmental discourse as the question of limits to growth, specifically, whether industrialism is sustainable into the indefinite future as it is adopted round the world. Biologists began the debate in the 1960s by importing from population ecology concepts of carrying capacity limits, and they averred that cascading world population growth augured for environmental catastrophe. Some social scientists amplified this neo-Malthusian argument, noting that industrialism is not a population but a way of life, one which produces too much consumption not just too many consumers.

Attentive to anthropogenic environmental degradation (aquifer depletion, biodiversity loss, deforestation, ozone depletion, global warming, etc.), the new social science tradition of examining environmental interfaces with human production and reproduction is now secure with the appearance of numerous textbooks and treatises on environmental sociology, environmental psychology, and ecological economics (see Costanza et al. 1997, Harper 1996). These specialties tend to conceive human-environment interactions within the existing theory traditions of their established disciplines. Recently, ominous neo-Malthusian analyses have been challenged by more optimistic theories of ecological modernization (Spaargaren and Mol 1992).

Theory aside, problematics of sustainability are routinely encountered as matters of practice in the concerted effort to promote sustainable development in economically less secure nations of the world. Specifically promoted by the United Nations, sus-

tainable development is taken to mean eliminating world poverty while not compromising local environmental stabilities (Brundtland Commission 1987). In fact, sustainable development as practiced has promoted the transfer of industrial technology to promote economic growth. Where the ideal has succeeded, and by consensus it has not much, in effect sustainable development programs have promoted further industrialization—thus rekindling the question of the long term sustainability of human-environment interactions as local, regional, and global environmental degradation compounds, interactively cascading from industrialism's increased energy capture and waste disposal. The concern over environmental degradation directs the problematic of sustainability, which currently is targeted more towards local than toward global applications (see *Human-Environment Relationships*).

2.3 The Analogic-Symbolic Theory Frame

The analogic-symbolic theory frame works with a metaphorical interpretation of ecology, with different variants selectively emphasizing analogies or symbolic constructs as central to theory development. Generally speaking, the theory frame takes biological ecology just seriously enough to borrow some of its content (sometimes merely its words) so that nature can serve as a metaphor to understand society, thus advancing social science. In the classic sociological tradition, 'environment' usually had no biological referent, and neither does 'community' or 'ecosystem,' which are favorite analogs in modern sociological human ecology. As noted earlier, US sociology provided the first sustained research tradition of human ecology, something of an urban geography, concerned as it was with spatial relations in cities, that beginning in mid-twentieth century the influential Amos Hawley gradually transformed into an ecological demography. In this, territorially circumscribed human populations are characterized by technologically based organizational adaptations in the sustenance producing activities of populations within their social, political, and economic environments. These orienting concepts have not changed in more than a generation (see Micklin and Posten 1998). They reflect a social and demographic, but not a biological, materialism in the conception of purely sociological problematics. Direct descendants, such as organizational ecology, and distant cousins, such as family ecology, likewise depend on theoretical analogies to biological phenomena in order to relate social entities to social environments (see *Population Ecology*).

A wholly independent and more recent set of problematics relating society and environment applies a social constructionist interpretation to environmental social science, counterweighting its biological interpretation. In this, natural environments become

material for the analysis of symbolic meanings and understandings by which social groups perceive and define environmental events as real and respond to them collectively. One-to-one theory analogs with bioecology are not sought, but the metaphorical interpretation of ecology runs deep because symbolized realities are accorded primacy over biophysical realities. Natural environments are taken to have significance only as people socially construct, perceive, define, and negotiate them as real, and make contestable claims about them. The problematics address social responses to natural environments insofar as environmental issues are defined through shared meanings (Hannigan 1995).

Theory variants that are exclusively analogic or symbolic usually seek to separate not integrate human ecology and bioecology, so as to strengthen social science *per se*. Implicitly, sometimes explicitly, many variants reflect the species prejudice that nonhuman organisms represent 'lower' forms of life and the cultural prejudice that Western industrialism represents the 'highest' form of social progress. These two prejudices are explicitly criticized in many variants of the interactive theory frame.

2.4 The Interactive Theory Frame

The interactive theory frame for human ecology is the newest, most diverse and most numerous in its scholarly problematics. Its growing traditions may only be listed here. They include assorted environmental ethics, such as deep ecology, the land ethic, and animal rights; assorted spiritual ecologies, including native US land wisdom, Christian-based ecological process theology, and strains of Buddhist, Taoist, and Jainist thought; some applications of complexity theory for evolving interactive systems; Gaian ecology; politically grounded approaches, such as social ecology and green Marxism; ecofeminism; and more (see, for example, Bookchin 1995, Benton 1996, Callicott 1994, Tucker and Grim 1993, Warren 1996).

All of these assume some variant of the ideological interpretation of ecology, reflecting a wish to integrate ecological knowledge with selected priorities in human affairs. Some are based in science, while others are based in philosophy, politics, religion, or tradition, but almost none maintains a hard and fast distinction between objective and subjective knowledge. Almost all take a holistic rather than reductionistic approach to their problematics, in which the traditional boundaries of scientific disciplines, and even the classic boundaries between science and the humanities, are regarded as anachronistic. Most reject any nature-culture dualism, and many disavow anthropocentrism. Multiple distinct problematics are defined, depending on the point of view. The overriding theme construes human ecology as knowledge of the complex interactive connections that integrate human activities with

environmental contexts. The overall result is an eclectic combination of biological, environmental, moral, and even esthetic, approaches to human ecology (see *Nature, Concepts of: Environmental and Ecological; Ecology, Deep; Feminist Political Ecology*).

3. Conclusion

Human ecological scholarship on the biological, environmental, and ideological interpretations, with their derivative problematics, is now being increasingly writ into the curricula of educational institutions. Calls for 'green education' are being implemented internationally, programs in environmental science that include a human component are expanding, and one US undergraduate institution of higher learning, the College of the Atlantic, devotes its curriculum entirely to human ecology. While still stymied by the traditional boundaries of academic disciplines, and sometimes by parochial scholarship, human ecology—the subject, not the name—is experiencing dynamic growth, as befits the coming century of the environment.

See also: Darwinism; Social; Ecological Economics; Ecological Imperialism; Ecology: Organizations; Ecology, Political; Environmental Challenges in Organizations; Sustainability Transition: Human-Environment Relationship; Sustainable Development

Bibliography

- Benton T (ed.) 1996 *The Greening of Marxism*. Guilford Press, New York
- Biersack A 1999 Introduction: from the 'new ecology' to the new ecologies. *American Anthropologist* 101: 5–18
- Bookchin M 1995 *The Philosophy of Social Ecology*. Black Rose Books, Montreal, Canada
- Boyd R, Richerson P J 1985 *Culture and the Evolutionary Process*. University of Chicago Press, Chicago
- Brundtland Commission (World Commission on Environment and Development) 1987 *Our Common Future*. Oxford University Press, Oxford, UK
- Callicott J B 1994 *Earth's Insights*. University of California Press, Berkeley, CA
- Costanza R, Cumberland J, Daly H, Goodland R, Norgaard R 1997 *An Introduction to Ecological Economics*. St. Lucie Press, Boca Raton, FL
- Durham W H 1991 *Coevolution: Genes, Culture, and Diversity*. Stanford University Press, Stanford, CA
- Evernden N 1992 *The Social Creation of Nature*. Johns Hopkins University Press, Baltimore, MD
- Ewald P W 1994 *Evolution of Infectious Disease*. Oxford University Press, Oxford, UK
- Hannigan J A 1995 *Environmental Sociology: A Social Constructionist Perspective*. Routledge, London
- Harper C L 1996 *Environment and Society: Human Perspectives on Environmental Issues*. Prentice-Hall, Upper Saddle River, NJ

- Kipple K F 1997 *Plague, Pox, and Pestilence*. Weidenfeld and Nicholson, London
- Machalek R 1999 Elementary social facts: emergence in non-human societies. In: Freese L *Advances in Human Ecology*. JAI Press, Greenwich, CT, Vol. 8, pp. 33–64
- Mascie-Taylor C G N (ed.) 1993 *The Anthropology of Disease*. Oxford University Press, Oxford, UK
- Micklin M, Poston D L Jr. (eds.) 1998 *Continuities in Sociological Human Ecology*. Plenum, New York
- Oelshlaeger M 1994 *Caring for Creation*. Yale University Press, New Haven, CT
- Orr D W 1992 *Ecological Literacy: Education and the Transition to a Postmodern World*. State University of New York Press, Albany, NY
- Scoones I 1999 New ecology and the social sciences: what prospects for a fruitful engagement? *Annual Review of Anthropology* 28: 479–507
- Smith E A, Winterhalder B (eds.) 1992 *Evolutionary Ecology and Human Behavior*. Aldine de Gruyter, New York
- Spaargaren G, Mol A P J 1992 Sociology, environment, and modernity: ecological modernization as a theory of social change. *Society and Natural Resources* 5: 323–44
- Tucker M E, Grim J A (eds.) 1993 *Worldviews and Ecology*. Bucknell University Press, Lewisburg, PA
- Warren K J (ed.) 1996 *Ecological Feminist Philosophies*. Indiana University Press, Bloomington, IN

L. Freese

Human Ecology: Insights on Demographic Behavior

1. Introduction

Human ecology is the ‘the study of the form and the development of the community in human populations’ (Hawley 1950, p. 68, Hawley 1971) for which the unit of analysis ‘is not the individual but the aggregate which is either organized or in the process of being organized’ (Hawley 1950, p. 67) and where the concept of ‘community’ is defined as a ‘territorially based system of functional interdependencies that results from the collective adaptation of a population to its environment’ (Berry and Kasarda 1977, p. 12). Human ecology proposes that populations adapt to social, economic, political, and biophysical environments through organizational innovations, and that systems grow ‘to the maximum size and complexity afforded by the technology for transportation and communication’ in reducing the time cost of maintaining the accessibility and exchange required by interdependent populations (Hawley 1986, p. 7). It is important to recognize that it is this theoretical framework, not the mere fact that aggregate data are used, that places research in the human ecological tradition.

2. Contemporary Human Ecology

The intellectual lineage of human ecology traces from the emergence of plant and animal ecology in the late nineteenth and early twentieth centuries and its application to human populations in the work of Park and Burgess (1921), McKenzie (1933), and other scholars at the University of Chicago. The approach of these scholars, while innovative, was ultimately limited in scope. Perhaps most limiting was the initial insistence that cultural variables be excluded from the purview of human ecology in favor of a focus on spatial patterns resulting from a process of unplanned competition. By contrast, the insights into demographic phenomena described herein depend almost entirely on *contemporary human ecology* (hereafter, *human ecology*) which denotes the body of theory and research that began with the publication in 1950 of Amos H. Hawley’s *Human Ecology*, a landmark volume which led to a major reorientation of the field. Hawley’s work dramatically shifted the focus from the description of physical features of the city and mapping of social phenomena to ‘the ways in which human populations organize in order to maintain themselves in given environments, thus relegating spatial analysis to a minor though still useful position in the discipline’ (Hawley 1986, p. 3). It is not spatial arrangements themselves, but rather the social relations often reflected in spatial patterns and processes, that are of interest to human ecologists (Frisbie and Kasarda 1988 for further discussion).

Albeit something of an oversimplification, the human ecological paradigm can be thought of as encompassing four broad categories of variables: population, organization, environment, and technology—codified by Duncan (1959, 1964) as the ecological complex bearing the acronym POET. It is crucial to recognize that, in human ecology, population is defined generically to mean ‘a number of units sharing one or more given characteristics... (which acquires) operational significance only when clothed with a set of institutions and located in a space-time context’ (Hawley 1998, pp. 11–12). At least since Hawley’s redirection of the field (1950), the primary unit of analysis for human ecology has been the organized population, not the individual units comprising populations. In addition, human ecology places a much heavier emphasis on analysis of the social and economic environment, as compared to the biophysical environment (Hawley 1998). This distinguishes (but does not isolate) human ecological concerns from those which characterize bioecology and the environmental movement.

The term ‘demographic behavior’ is used broadly here to refer, not only to the basic processes of fertility, mortality, and migration, but also to *social demographic* phenomena, with social demography defined as ‘the analysis of how general social and cultural factors are related to population structure and pro-

cess' (Ford and De Jong 1970, p. 4). In this context, an impressive range of human ecological insights into a large number of demographic issues might be identified. This article primarily emphasizes contributions to the study of the systemic attributes and processes involved in population growth and societal development, the distribution and redistribution of population, and intra- and interurban morphology. While this selectivity is logical in light of human ecology's emphasis on systems, it warrants mention that the well-known contributions of human ecology to research on social networks and residential segregation are given little attention inasmuch as these are the subjects of other articles included in this volume (see *Social Networks and Fertility* and *Residential Concentration/Segregation, Demographic Effects of*, respectively). Further, while much of the present discussion concentrates on research conducted in the US, human ecology has been drawn upon rather extensively in cross-national research (Frisbie and Al-Khalifah 1998).

3. Human Ecology's General Theoretical Contribution to Macrolevel Demography

It is clear that human ecology has contributed substantially to the preservation and advance of macrolevel theory in the social and behavioral sciences, generally, and in demography, in particular (Poston and Frisbie 1998). However, in the mid-twentieth century, two events combined to diminish the interest of demographers and other social scientists in macrolevel analysis. The first was the publication of work dealing with what came to be known (erroneously) as the 'ecological fallacy.' Robinson (1950) argued that results of research in which aggregates are the units of analysis cannot be used to infer relationships among individuals and that a large proportion of research that employs aggregate data does so only because information on individuals is not available. The second was actually a sequence of events involving the emergence of modern techniques of sampling and survey analysis which greatly enhanced the opportunity for rigorous microlevel studies (Linz 1969). For a time, demography listed far toward reliance on individual-level research, even in the face of warnings by renowned scholars that the discipline was ill-served by a fixation at the microlevel (e.g., Ryder 1980).

A more balanced view prevails, partially due to the efforts of 'ecological demographers' both early on (Schnore 1961) and later (Namboodiri 1988). As it turned out, Robinson's thesis was only partially correct. The problem identified by Robinson is not 'ecological' in nature, but rather an issue that arises whenever cross-level generalizations of any kind are contemplated. Moreover, many of the central tenets of human ecology are inherently macrolevel constructs

devoid of any counterpart at the individual level. For example, the concept of the division of labor makes sense only when applied to groups. Analogous reasoning applies to many demographic variables. Individuals are born, they die, and in between they may migrate, but only groups, not individuals, have birth, mortality or migration rates.

A signal contribution of human ecology has thus been the development of a theoretical framework for a wide range of demographic studies of organized aggregates. Examples include investigations of the relationship between commodity exchange and position of cities in the US metropolitan hierarchy (Eberstein and Frisbie 1982), as well as in the world-wide system of cities (Meyer 1986), the effects on the volume and rate of migration, of functional niche and extent of the division of labor (Murdock et al. 1992, Poston and Mao 1996), changes in the social stratification of suburbs (Guest 1978, Krivo and Frisbie 1982), segregation of racial and ethnic populations (Massey and Denton 1989), and the 'birth' and 'death' rates of formal organizations (Carroll 1984, Hannan and Freeman 1977). Such work, in turn, offers a substantive grounding for multilevel analyses, in which, e.g., the behavior of individuals is partially explained in terms of the social or ecological context in which they exist.

4. Specific Insights on Demographic Phenomena

4.1 Population Growth

Perhaps no demographic issue has seized and held the attention of scholars, policy makers, the media, and the general public as that of rapid population growth. Interest in the 'population explosion' intensified circa 1970 in the wake of publications which predicted exhaustion of resources and irremediable environmental pollution if population and economic growth continued unchecked (Erllich 1968, Meadows et al. 1974). A wide range of less apocalyptic, but still highly disturbing, publications focused attention on less developed countries (LDCs) and the actual or anticipated deleterious consequences of rapid population growth for development and increases in the standard of living in these nations—effects epitomized by the piling up of excess rural populations in peripheral slums of huge cities (Davis 1975).

Early on, some scholars challenged the view that continued rapid population growth spelled worldwide collapse (see Davis 1990 for a summary of the position of 'alarmists' and 'skeptics'). Regardless of whether one tends toward pessimism or optimism, it is evident that demography has to come to grips with the Malthusian contention that population growth will inevitably be limited by exhaustion of natural and/or nonrenewable resources. Before one can meaningfully discuss its utility for better understanding population

growth, it is necessary to recognize that, in human ecological theory, population is regarded as a necessary, but not sufficient, condition. An important analytical implication is that population

cannot operate independent of organizational circumstances. In and of itself it cannot cause war, resource exhaustion, or environmental pollution, as Malthusians have argued. Such outcomes are explained as due to maladaptations or malfunctioning of organization (Hawley 1986, p. 26).

Human ecologists have also called attention to certain errant assumptions that are perhaps more apt to be found in popular or 'vulgar' interpretations than in Malthus' work itself (Keyfitz 1982). Among the most troublesome of these is the naïve perspective that the concept of manageable population size can be operationalized as a 'simple ratio of organisms to units of a resource or resources'—naïve because it fails to take into account 'the manner in which the population is organized to use the resources' (Hawley 1950, p. 151). This flaw apparently underlies the conclusion that concepts such as 'carrying capacity' and 'limits to growth,' have failed to provide either 'analytic power, or foresight, instead of hindsight' (Davis 1990, p. 13).

Constructive insights are found in the human ecological model of system change, growth, and equilibrium. In *system* terms, growth refers to 'maturation ... through the maximization of the potential for complexity and integration implicit in the technology for movement and communication possessed at a given point time' (Hawley 1986, p. 52). This perspective has direct roots in Boulding's (1953) theory of nonproportional change which suggests that a limit to growth occurs, not when some ratio of organisms to resources is reached, but when the communications and transportation costs involved in maintaining the coherence of a system become prohibitively high. Growth of systems can be expected to take the form of a logistic curve which describes a period of rapid growth which gives way to slow or no growth. Equilibrium in this view is a *temporary* state, and population increase will resume in combination with system expansion as additional organizational and technological inputs allow breaking through the upper asymptote imposed by the costs of movement and coordination of activities (Hawley 1986, 1998).

Applied specifically to demography, this insight might be translated as follows:

As to demographic growth, its initial acceleration is a natural byproduct of...economic transformation since, predictably, the fruits of economic improvement are in part used for obtaining better health and lowered mortality. Malthusian outcomes, however, need not be the inevitable result of rapid population growth. Such growth is transitory: given economic success, the spontaneous onset of "demographic transition" can be confidently expected. Pressure built into the reward mechanisms of modern industrial society induce behavioral

changes that eventually lead to low fertility, hence to low or no population growth (Demeny 1990, p. 417).

Recent worldwide fertility declines are consistent with this perspective. A 1999 United Nations publication shows that the world total fertility rate (TFR = the number of children a female surviving to the end of her reproductive life would bear given prevailing age-specific fertility rates (Keyfitz and Flieger 1990, p. 16)), after hovering around 5.0 during the period 1950–70 had fallen precipitously to 2.93 in 1990–95, with an expected TFR = 2.71 for 1995–2000 (UN 1999). By 1975–80, fertility in developed nations was already below replacement level (TFR = 1.91; for replacement, TFR is approximately 2.1, allowing for mortality of children). In LDCs, the rate for 1990–95 was 3.27, but this represented a rather astonishing decline from TFR = 6.00 in 1965–70. Based on its medium variant, the UN (1999, pp. 520–1) projects that average fertility among today's LDCs will fall to bare replacement level within the first four to five decades of the twenty-first century.

Of course, even with this major slowdown in population growth, 50 years into the twenty-first century, the world will still have a population much larger than today's roughly six billion (UN 1999, US Agency for International Development 1999). Some scholars (e.g., Bongaarts and Feeney 1998) have argued that current low fertility rates are, at least in some contexts, temporarily, rather than permanently, depressed to below replacement, while others believe the contemporary trend is likely to be permanent or at least rather persistent (Lesthaeghe and Willens 1999).

5. Urbanization

The concentration, deconcentration, and organization of urban populations have been central to ecological theory and research. As noted, Hawley defined human ecology as the study of the form and development of the community, of which urban places are a prime example. Although neither the *highly stylized spatial* models of urban form and growth developed by Burgess and Park, nor those of other scholars, have remained central to urban studies, these heuristic representations had a basis in analytical constructs that are important building blocks for ecological theories of the distribution and interaction of populations in both intraurban and interurban systems.

5.1 Intraurban Systems: Land Use and Population Distribution

Among the most directly relevant of human ecological insights for understanding *intrametropolitan* systems are those having to do with land use and population concentration/deconcentration. Of special interest is

Burgess' notion of geographical gradient as a basis for the hypothesis predicting an inverse relationship between intensity of land use by a population and distance from center to the periphery of cities (cf. Hawley 1971). This relationship, formalized by Colin Clark (1951), can be expressed in the form of natural logarithms as

$$\ln d_x = \ln d_o - bx \quad (1)$$

where d_x is the density of population at distance x from the center, d_o is the central density, and b is an empirically derived density gradient. Reflected in this equation is a theory of land use which begins with the premise that functional differentiation and interdependence of populations and organizations in a city require mutual accessibility. Since, *ceteris paribus*, the center of a city is the point most accessible to all other parts of the city, there will be competition for core location, thereby driving up the cost of land at or near the center which, in turn, requires that more intensive use must be made of space. As ease of access declines with distance from the core, so too does the cost of land and the intensity of its use (Berry and Kasarda 1977, Hawley 1971, Turner 1994). The magnitude of the insight afforded by the human ecological theory of urban land use for demographers and others interested in the distribution and redistribution of population is seen in the fact that the validity of this model has been demonstrated by research using data covering a period of over a century and a half and for more than 100 cities (Frisbie and Kasarda 1988, p. 634).

Ecological theory also predicts that reduction in the friction of movement of population, materials, and ideas over time and space will cause the slope of the gradient to become less and less steep. As modern means of transportation developed and dramatic advances were achieved in accumulating, processing, and disseminating information, simple models based on a core-to-periphery gradient in a monocentered city have become progressively less relevant as metropolitan areas have increasingly become multinodal. But, because it is the systemic relationships reflecting the need for mutual accessibility of functionally interdependent populations, not spatial patterns, *per se*, that are central to the theory, the relevance of the ecological perspective remains in the face of rapidly changing urban systems. To illustrate, in the 1970s, as documented by numerous scholars and summarized by Edmonston and Guterbock (1984), several factors seemed inimical to the continuation of the rapid pace of suburbanization. Transportation costs had risen dramatically due to the reduction of petroleum supplies by major producing nations, while income growth in constant dollars slowed notably and housing prices soared—all of which raised the cost of moving to, and commuting from, the urban periphery. In addition, demographic composition had shifted such

that the types of families that had traditionally fueled movement to suburbs, i.e., larger households headed by couples of prime child-bearing age, became proportionately fewer. Taking the coefficients of Clark's equation as dependent variables for more than 200 metropolitan areas, Edmonston and Guterbock demonstrated that mean central densities of cities continued to decline over a 25-year time span and that 'the density gradient dropped consistently' (1984, p. 918). That is, the initial concentration at the core of cities occurred due to the need for accessibility. The large negative sign of the density gradient reflected the less developed transportation and communications technology available during the early stage of city building, but the slope flattened as technological innovations emerged. When the cost of movement over time and space rose sharply in the 1970s, the deconcentration of population nevertheless continued because, by that time, jobs and market outlets (industrial parks and shopping malls), recreational facilities, and utilities had also expanded well into the urban periphery (Berry and Kasarda 1977, Edmonston and Guterbock 1984). Thus, in the face of altered social and economic conditions, the human ecological paradigm continues to be a reliable guide for research on population distribution and redistribution. The reason, in part, is due to the conceptualization of space as a time-cost variable. What organized populations seek to minimize is not *distance* between interdependent entities, but the cost in time, effort, and other resources required for regular interaction (Stephan and McMullin 1981).

Another crucial aspect of intraurban population distribution in the US is residential segregation by race and ethnicity. As mentioned above, this topic is addressed in detail later in this volume. However, a brief discussion is in order here because it is demographic analysis that has most fully documented segregation, and a substantial portion of the explanation for these findings has been provided by human ecological theory. A prime example is seen in the work of Massey and colleagues who, over many years, have traced the sharp separation of African Americans from whites (and the lower levels of separation of Latinos and Asians from whites), taking into account variation in socioeconomic status, nativity, and other factors (Massey 1979, Massey and Denton 1989, 1993, Farley and Frey 1994). Based on an examination of segregation in the US and five other nations, Massey concluded that 'the operation of common processes of *ethnic* concentration and dispersal are well-summarized by the ecological model' (Massey 1985, p. 315; emphasis added), especially as the model pertains to chain migration, invasion, and succession of populations, the urban economy, and the housing market. Although human ecological theory is much less helpful in accounting for black-white segregation, Massey and Eggers (1990) have usefully addressed this phenomenon in research on 'the ecology of inequality.'

5.2 Interurban Systems: National and World Patterns

A useful point of departure for many demographers interested in *intermetropolitan* organization has been the classification of cities according to horizontal and vertical dimensions. The horizontal dimension refers to functional specialization of cities in various goods and services at various stages of production for the market, as operationalized in terms of types of industrial activities (Duncan et al. 1960, Frisbie and Kasarda 1988). The vertical dimension refers to the level of dominance exercised by cities in a metropolitan hierarchy as measured by performance of coordinative functions, including financial services and organization of the market (Duncan et al. 1960, Eberstein and Frisbie 1982, Wanner 1977) on a scale that may be regional, national, or international in scope (Frisbie and Kasarda 1988, Wilson 1984).

Moving down the ladder of abstraction leads to a crucial hypothesis, viz., that a positive association will exist between the performance of higher-order functions and extensiveness of involvement in the urban web of interdependence. An empirical test (Eberstein and Frisbie 1982), using data on 171 US urban centers, with metropolitan function defined as performance of financial, commercial, administrative, and distributive activities and extensiveness of interdependence measured in terms of volume and variety of trade flows and number of exchange linkages among urban centers, provided considerable support for this hypothesis. In a study of the world system of cities, Meyer concluded that '(t)he hypotheses of ecological dominance were confirmed' (1986, p. 553). Specifically, Meyer found that, with a focus on the key functions of banking and finance, world core cities (e.g., New York, London, Tokyo) 'dominate the peripheral metropolises of South America, and within the core the upper level metropolises exert even greater dominance than the lower level metropolises' (1986, p. 553). In addition to the general support provided for ecological theory, such findings reinforce the social demographic point that 'while all metropolises are large cities, not all large cities are metropolises. Population size is a concomitant; function is the keynote' (Vance and Sutker 1954, p. 115).

6. Migration

'Change without movement is impossible' (Hawley 1950, p. 324). Based on this tenet, human ecology views migration (whether within or across national boundaries) as a means through which something approaching a balance between population size and life chances may be *temporarily* achieved. Many of the most fundamental concepts employed in migration research are rooted in human ecological theory, including the recognition that, because migration

involves a destination as well as a place of origin, it is necessary to consider both 'push' and 'pull' factors. Recently, human ecological models, which also include population size and distances between origin and destination as predictors, have been able to account for well over half the variation in migration streams of interstate migration in the US (Poston and Mao 1996) and interprovincial migration in China (Poston and Mao 1998). The ecological view of the push of 'overpopulation' as an impetus to migration refers not only to an excess of population to land, food, or some other natural resource, but more generally to availability of ecological niches in the form of job opportunities, especially as these are generated by the movement of capital (Hawley 1950).

Another contribution is seen in the notion of 'chain migration,' a phenomenon to which classical ecologists called attention. Chain migration means that 'migrants do not select destinations randomly, but follow social networks... to specific jobs in particular neighborhoods of selected cities' (Massey 1985, p. 318). The concept has enhanced our understanding of how residential segregation occurs and of how migrants adapt at destination through drawing on 'social capital' provided by co-ethnics (Massey 1985, 1991).

Clearly, the human ecological perspective must be supplemented by other theories of migration (Portes and Fernandez-Kelly 1989). One problem with the 'push—pull' model is that it cannot explain, '(g)iven the same set of expelling forces and factors of attraction, why is it that some individuals leave while others stay' (Portes 1991, p. 76). This weakness is inherent in the macrolevel approach of human ecology. Microlevel studies essentially seek an answer to the question 'Who moves and why?' By contrast, for human ecological studies of migration, the central question is 'Where do people go and why?'

While ecological theories of migration involve much more than a simple 'push—pull' hypothesis, there is, nevertheless, a certain tendency to neglect population movements that are forced or 'impelled' by political agents at origin (Petersen 1958). An analogous gap is evident at destination due to the lack of attention paid by ecologists to the impact of immigration policies instituted by receiving countries seeking to restrict, and/or influence the composition of, immigration flows (Freeman and Bean 1997). Further, despite the importance attributed by Hawley to capital flows as a determinant of immigration (1950, p. 330), research in human ecology seems to have failed to appreciate fully the importance of the penetration of capital for the creation of economic imbalances that stimulate immigration (Portes and Fernandez-Kelly 1989).

The theoretical contributions of human ecology to this line of research did not cease with the appearance of Human Ecology. Hawley has extended his own ideas concerning the spatial movement and redistribution of population, particularly with respect to urban-oriented migration (1971, 1986). More recently,

as part of a comprehensive theory of 'assembling of human populations' that involves geopolitical as well as ecological processes, Turner has set forth a set of propositions that envisions the rate of net immigration as 'a positive and additive function of: (a) the level of transportation and communication technologies, (b) the rate and extent of contact with external populations, (c) the differentiation as well as volume and velocity of markets, (d) the diversity of the population, and (e) the size of population settlements' (Turner 1994, p. 88). Coupled with the notions of chain migration and population flows preceded by capital flows, Turner's formulation seems highly pertinent to many of the international population flows that have characterized the twentieth century.

7. Summary

Contemporary human ecology has sustained and helped extend the theoretical foundation of macrolevel demography. It has also contributed in numerous ways to the understanding of specific demographic phenomena, as illustrated by reference to research on population growth, urban form and process, and migration, with brief attention given to residential segregation by race and ethnicity. While other examples could have been employed, it is palpable that human ecology provides a paradigm that, in spite of certain weaknesses and oversights, greatly facilitates comprehension of how populations grow and distribute themselves within, and adapt to, the natural and social environments.

See also: Demography, History of; Ecological Economics; Ecological Fallacy, Statistics of; Human Ecology; Migration, Economics of; Migration, Theory of; Population Ecology; Population Pressure, Resources, and the Environment: Industrialized World; Residential Segregation: Sociological Aspects; Urban Geography

Bibliography

- Berry B J L, Kasarda J D 1977 *Contemporary Urban Ecology*. Macmillan, New York
- Bongaarts J, Feeney G 1998 On the quantum and tempo of fertility. *Population and Development Review* **24**: 271–91
- Boulding K E 1953 Toward a general theory of growth. *Canadian Journal of Economics and Political Science* **19**: 326–40
- Carroll G R 1984 Organizational ecology. *Annual Review of Sociology* **10**: 71–93
- Clark C 1951 Urban population densities. *Journal of the Royal Statistical Society Series A* **114**: 490–6
- Davis K 1975 Asia's cities: problems and options. *Population and Development Review* **1**: 71–86
- Davis K 1990 Population and resources: fact and interpretation. *Population and Development Review* **16**(suppl.): 1–21

- Demeny P 1990 Tradeoffs between human numbers and material standards of living. *Population and Development Review* **16**(suppl.): 408–21
- Duncan O D 1959 Human ecology and population studies. In: Hauser P M, Duncan O D (eds.) *The Study of Population*. University of Chicago Press, Chicago
- Duncan O D 1964 Social organization and the ecosystem. In: Faris R E L (ed.) *Handbook of Modern Sociology*. Rand-McNally, Chicago
- Duncan O D, Scott W R, Lieberman S, Duncan B 1960 *Metro-polis and Region*. Johns Hopkins University Press, Baltimore, MD
- Eberstein I W, Frisbie W P 1982 Metropolitan function and interdependence in the US urban system. *Social Forces* **60**: 676–700
- Edmonston B, Guterbock T M 1984 Is suburbanization slowing down?: Recent trends in population deconcentration in US metropolitan areas. *Social Forces* **62**: 905–25
- Erlich P 1968 *The Population Bomb*. Ballentine, New York
- Farley R, Frey W H 1994 Changes in segregation of Whites from Blacks during the 1980s: Small steps towards a more integrated society. *American Sociological Review* **59**: 23–45
- Ford T F, De Jong G F 1970 *Social Demography*. Prentice-Hall, Englewood Cliffs, NJ
- Freeman G P, Bean F D 1997 Mexico and US worldwide immigration policy. In: Bean F D, de la Garza R O, Roberts B R, Weintraub S (eds.) *At the Crossroads: Mexico and US Immigration Policy*. Lanham, Boulder, CO
- Frisbie W P, Al-Khalifah H M 1998 Ecology's contribution to cross-national theory and research. In: Micklin M, Poston Jr D L (eds.) *Continuities in Sociological Human Ecology*. Plenum, New York
- Frisbie W P, Kasarda J D 1988 Spatial processes. In: Smelser N J (ed.) *Handbook of Sociology*. Sage, Newbury Park, CA
- Guest A M 1978 Suburban social status: Persistence or evolution. *American Sociological Review* **43**: 251–64
- Hannan M T, Freeman J 1977 The population ecology of organizations. *American Journal of Sociology* **82**: 929–64
- Hawley A H 1950 *Human Ecology: A Theory of Community Structure*. Ronald, New York
- Hawley A H 1971 *Urban Society: An Ecological Approach*. Ronald, New York
- Hawley A H 1986 *Human Ecology: A Theoretical Essay*. University of Chicago Press, Chicago
- Hawley A H 1998 Human ecology, population and development. In: Micklin M, Poston Jr D L (eds.) *Continuities in Sociological Human Ecology*. Plenum, New York
- Keyfitz N 1982 The evolution of Malthus's thought. In: Keyfitz N (ed.) *Population Change and Social Policy*. Abt Books, Cambridge, MA
- Keyfitz N, Flieger W 1990 *World Population Growth and Aging*. University of Chicago Press, Chicago
- Krivo L J, Frisbie W P 1982 Measuring change: the case of suburban status. *Urban Affairs Quarterly* **17**: 419–44
- Lesthaeghe R, Willens P 1999 Is low fertility only a temporary phenomenon in the European Union. *Population and Development Review* **25**: 211–28
- Linz J J 1969 Ecological analysis and survey research. In: Dogan M, Rokkan S (eds.) *Quantitative Ecological Analysis in the Social Sciences*. MIT Press, Cambridge MA
- Massey D S 1979 Residential segregation of Spanish Americans in United States urbanized areas. *Demography* **16**: 553–63
- Massey D S 1985 Ethnic residential segregation: A theoretical synthesis and empirical review. *Sociology and Social Research* **69**: 315–50

- Massey D S 1991 Economic development and international migration in comparative perspective. In: Diaz-Briquets S, Weintraub S (eds.) *Determinants of Emigration From Mexico, Central America and the Caribbean*. Westview Press, Boulder, CO
- Massey D S, Denton N A 1989 Hypersegregation in US metropolitan areas: Black and Hispanic segregation along five dimensions. *Demography* **26**: 373–91
- Massey D S, Denton N A 1993 *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press, Cambridge, MA
- Massey D S, Eggers M L 1990 The ecology of inequality: Minorities and the concentration of poverty, 1970–1980. *American Journal of Sociology* **95**: 1153–88
- McKenzie R D 1933 *The Metropolitan Community*. McGraw-Hill, New York
- Meadows D H, Meadows D L, Randers J, Behrens III W W 1974 *Limits to Growth: A Report for the Club of Rome Project on the Predicament of Mankind*, 2nd edn. Universe Books, New York
- Meyer D R 1986 The world system of cities: Relations between international financial metropolises and South American cities. *Social Forces* **64**: 553–81
- Murdock S H, Backman K, Hwang S, Hamm R R 1992 Sustenance specialization and dominance in international and national ecosystems: Implications for post-1980 migration in counties in the United States. In: Jobes P C, Stinner W F, Wardwell J M (eds.) *Community, Society and Migration: Noneconomic Migration in America*. University Press of America, Lanham, MD
- Nambodiri K 1988 Ecological demography: Its place in sociology. *American Sociological Review* **53**: 619–33
- Park R E, Burgess E W 1921 *Introduction to the Science of Sociology*. University of Chicago Press, Chicago
- Petersen W 1958 A general typology of migration. *American Sociological Review* **23**: 256–66
- Portes A 1991 Unauthorized immigration and immigration reform: Present trends and prospects. In: Diaz-Briquets S, Weintraub S (eds.) *Determinants of Emigration from Mexico, Central America, and the Caribbean*. Westview Press, Boulder, CO
- Portes A, Fernandez-Kelly M P 1989 Images of movement in a changing world: A review of current theories of international migration. *International Review of Comparative Public Policy* **1**: 15–33
- Poston Jr D L, Frisbie W P 1998 Human ecology, sociology, and demography. In: Micklin M, Poston Jr D L (eds.) *Continuities in Sociological Human Ecology*. Plenum, New York
- Poston Jr D L, Mao M X 1996 An ecological investigation of interstate migration in the United States. *Advances in Human Ecology* **5**: 303–42
- Poston Jr D L, Mao M X 1998 Interprovincial migration in China, 1985–1990. *Research in Rural Sociology and Development* **7**: 227–50
- Robinson W S 1950 Ecological correlations and the behavior of individuals. *American Sociological Review* **15**: 351–7
- Ryder N B 1980 Where do babies come from? In: Blalock Jr H M (ed.) *Sociological Theory and Research: A Critical Appraisal*. Free Press, New York
- Schnore L F 1961 The myth of human ecology. *Sociological Inquiry* **31**: 128–39
- Stephan G E, McMullin D R 1981 The historical distribution of county seats in the United States: A review, critique, and test of time-minimization theory. *American Sociological Review* **46**: 907–17
- Turner J H 1994 The assembling of human populations: Toward a synthesis of ecological and geopolitical theories. *Advances in Human Ecology* **3**: 65–91
- United Nations 1999 *World Population Prospects*. United Nations, New York, Vol. 1
- United States Agency for International Development 1999 *Popbriefs*. USAID, Washington, DC
- Vance R B, Sutker S S 1954 Metropolitan dominance and integration in the urban South. In: Vance R B, Demerath N J (eds.) *The Urban South*. University of North Carolina Press, Chapel Hill, NC
- Wanner R A 1977 The dimensionality of the urban functional system. *Demography* **14**: 519–37
- Wilson F D 1984 Urban ecology: Urbanization and systems of cities. *Annual Review of Sociology* **10**: 283–307

P. Frisbie

Human Evolutionary Genetics

Human evolutionary genetics is the inference of human history and evolution through the study of genetic variation. While much biological variation among humans is readily observable, variation at the DNA level typically is hidden. Only in the last decades of the twentieth century, with the development of molecular biological techniques, has this variation become detectable. The hidden patterns of human genetic variation have been shaped by events and processes of the past. For instance, the differing long-term census sizes of the human and chimpanzee species have led to dramatically different levels of genetic diversity within the two species. Chimpanzees are more genetically diverse, indicating that the species has undergone less severe bottlenecks than has the human species. While genetic data may not indicate why the human species went through a bottleneck, they provide at least a partial solution to the puzzle that is human evolutionary history.

The field of human evolutionary genetics is built on the premise that one can infer evolutionary history through the analysis of human genetic variation in the context of relevant theory and nongenetic data. Patterns of genetic variation are considered to be the outcome of a set of processes that can be modeled, primarily according to population genetics and molecular evolutionary theory. The goal is to ‘undo’ the process, in a sense, in order to infer the past. The extent to which nongenetic data are incorporated, either explicitly or implicitly, cannot be overstated. Geographic, linguistic, archaeological, demographic, historical, and climatological data, along with findings from other areas in biology, are incorporated to varying degrees as human evolutionary geneticists draw conclusions from the genetic data.

1. *Two Directions within Human Evolutionary Genetics*

The study of genetic variation provides two categories of insight into human history and evolution. Each category reflects an emphasis on one set of evolutionary processes. While the pattern of variation of some DNA segments may reflect both sets of processes, one set is often presumed to have predominated.

1.1 *Phenotypic Evolution*

A longstanding focus of human evolutionary genetics is phenotypic evolution, including changes in morphology, physiology, and behavior. By comparing patterns of genetic variation of humans and closely related species, for example, researchers can infer the processes that led to the current human phenotype (to the extent that this phenotype has an inherited genetic basis). Also included in this category is the study of the evolution of susceptibility to disease. To date, most findings regarding phenotypic evolution have been in this area. The more general question of what, at the DNA level, distinguishes humans from other species has only recently reached the forefront of research in human evolutionary genetics but promises to be an exciting research direction. Critical in this area of research is an understanding of the relationship between the variation at the DNA level and the variation at the phenotypic level.

1.2 *Population Processes and Events*

A second category, exemplified by the human–chimpanzee comparison above, includes population processes and events. Researchers attempt to infer, for example, population sizes of the past, rates of migration between groups, and the timing of population movements. This was a primary focus of the field during the latter half of the twentieth century, and remains so into this century. A central question has been the timing and location of the origin of anatomically modern humans.

2. *‘Population’ and ‘Group’*

Essential to the practice of human evolutionary genetics are definitions of the terms ‘population’ and ‘group.’ The less precisely defined term, ‘group,’ is used here to mean any collection of individuals. In a theoretical framework, the term ‘population’ is defined very precisely, as a set of individuals constituting a mating pool. All individuals of the appropriate sex in the population are considered to be equally available as potential mates. Groups of humans rarely, if ever, fit this definition of a population. The boundary between one population and another is obscure. In

practice, therefore, human evolutionary geneticists delineate populations along linguistic, geographic, sociopolitical, and/or cultural boundaries. A population might include, for example, all speakers of a particular Bantu language, all inhabitants of a river valley in Italy, or all members of a caste group in India.

3. *‘DNA Segment’*

The term ‘DNA segment’ is defined here as any length of DNA that is passed intact (without recombination of paternal and maternal DNA) from parent to child. Under this definition, any single nucleotide constitutes a DNA segment, as does the entire non-recombining portion of the Y chromosome. The entire mitochondrial (mt) DNA genome can also be considered an independently evolving DNA segment. Essential to this definition is the consistency of history across the nucleotides of a DNA segment. That is, without recombination, all nucleotides of a given individual’s DNA segment have been inherited via the same set of ancestors (e.g., father, paternal grandfather, etc., for the non-recombining portion of the Y chromosome). Success in inferring human evolutionary history depends in part on the extent to which we can infer the history of a particular DNA segment. It also depends on the number of independently-evolving DNA segments studied, as discussed below.

While there are, as exemplified above, stretches of the human genomes (nuclear and mitochondrial) that fit the above definition of a DNA segment, most stretches of the nuclear genome have the potential to undergo recombination. When human evolutionary geneticists analyze DNA sequences such as the Human Leukocyte Antigen (HLA) region on chromosome 6, they must either assume that recombination has occurred rarely enough to be ignored, or bring recombination into their analytical framework.

4. *Summarizing Genetic Variation*

In summarizing genetic variation, initially researchers count the number of copies of each genetic variant in each group under study. For instance, frequencies of the variants A, B, and O (also called alleles) of the ABO blood group system, one of the earliest polymorphic (variable) genetic systems studied, have been estimated for thousands of human groups around the world.

4.1 *Within Group Genetic Diversity*

Gene diversity (heterozygosity), defined as the chance that two DNA segments drawn randomly from a

group are different, summarizes within-group variation and is estimated from allele frequencies. Other measures of diversity take into account additional information, such as the number of differences between two DNA segments (nucleotide diversity), or the total number of variable positions along a DNA segment (number of segregating sites), within a group. Molecular evolution and population genetic models allow the interpretation of such diversity estimates in terms of mutation rates, population size changes, natural selection, and migration.

4.2 Among and Between Group Diversity

Human evolutionary geneticists often condense information regarding the genetic variation among a set of human groups into a single measure, typically called F_{ST} (Hartl and Clark 2000). Estimates of F_{ST} and similar statistics lie behind the oft-mentioned finding that the greatest proportion of the total genetic variation in the human species lies within human groups rather than among them. This has been interpreted as evidence for a relative lack of isolation among human groups. While F_{ST} summarizes variation among groups, genetic distances summarize the genetic difference between a pair of populations (Hartl and Clark 2000). Relatively small genetic distances are considered to be evidence that two groups have been less isolated from one another than have other pairs of groups. Differences in population sizes, however, strongly influence estimates of genetic distance.

5. Patterns of Variation for Different Segments of the Human Genome Evolve Independently

At each segment of the human genome, variation is introduced through mutation and subsequently modified through natural selection, random genetic drift, and gene flow between groups. The current pattern of variation of each such segment tells a story. Closely linked segments (those near one another on a chromosome that undergoes recombination, or on the same non-recombining chromosome) are likely to tell a similar story, at least going back a short time. In general, however, segments of the human genome evolve independently and therefore have different histories. The most recent common ancestor (MRCA) of all contemporary representatives of a particular DNA segment (e.g., the non-recombining portion of the Y chromosome) existed in a particular place at a particular point in time. The MRCA of another, independently evolving DNA segment (e.g., the mitochondrial genome) is likely to have existed in quite a different place, at a different point in time.

In the absence of natural selection, the pattern of variation at a particular segment reflects population sizes and movements of people. Even if each DNA

segment evolves independently and therefore has its own history, population size and migration influence the variation at all segments of the biparentally inherited parts of the genome similarly. Human evolutionary geneticists, therefore, combine the information from multiple DNA segments to draw conclusions regarding population processes. This consideration of multiple gene histories increases the robustness of conclusions (Mountain 1998).

An important consideration is the rate of mutation at each genomic location. DNA repair mechanisms that influence the rate of mutation differ between nuclear and mitochondrial DNA. Furthermore, DNA segments that include repeated elements (e.g., CACACACACACA) tend to mutate more rapidly than do nonrepetitive segments. Indeed, even within a short DNA segment of one hundred nucleotides, mutation rates may vary by an order of magnitude. The rate of mutation strongly influences the time depth over which a DNA segment is informative. A slowly mutating segment, for instance, will be uninformative over a short timescale. Mutation rates are very difficult to measure or estimate, primarily because mutation typically is rare (usually far fewer than 1 mutation per 10,000 nucleotide sites per generation).

With respect to gene flow, it is important to distinguish between interspecific and intraspecific questions regarding history and evolution. For questions regarding multiple species (e.g., the Hominoids), it is typically assumed that gene flow between species is essentially nonexistent. One consequence of a lack of exchange between groups is a greater chance of observing fixed genetic differences. That is, all individuals of one group are of one genetic type, while all individuals of another group are of a different type. Intraspecific questions must be approached differently in that gene flow may have occurred between groups below the species level. The current paucity of fixed differences between human populations reflects, among other factors, a great deal of genetic exchange between populations. There is a great deal of debate regarding the precise nature of genetic exchange between human populations. Migration models typically include a constant rate of gene flow over time. While such a model is likely to apply to human history at some level, more appropriate models include changing rates of gene flow and the possibility of rare but influential major migration events (Fix 1999).

When natural selection has a stronger influence on genetic variation than random genetic drift, the pattern of variation within and among groups is affected. Natural selection has clearly played a strong role in generating human genetic variation, especially in constraining the observable variation. While an environmental factor such as average temperature may influence variation at a number of loci simultaneously, and in a similar direction, other factors may influence only one or two loci. This provides one

way of distinguishing between a model of natural selection and a model involving population processes alone; the latter are expected to influence all loci, while environmental factors influence only a subset. Incorporating natural selection in addition to population processes into evolutionary models greatly increases the total number of models under consideration, so that demonstrating the impact of positive natural selection (i.e., new alleles arise within an individual and provide that individual with a selective advantage over other individuals) is a serious challenge. One thoroughly investigated case of natural selection influencing human genetic variation is that of the rise of lactose tolerance in adults (Durham 1991).

6. Development of the Field

As currently practiced, research in human evolutionary genetics builds upon over a century of the development of a theoretical framework, of technical tools (data analysis and biological), and of collections of data. Prior to 1900 there was essentially no research in the realm of human evolutionary genetics. The latter half of the nineteenth century was critical, however, in that C. Darwin, G. Mendel, and their contemporaries provided the basis for a theoretical framework with an emphasis on variation and natural selection. This continued into the twentieth century with the classic work of G. H. Hardy and W. Weinberg in 1908 (Provine 1971). Independently, they described the simple but elegant relationship between the frequencies of an allele and the frequencies of resulting genotypes (in a population, with individuals mating at random with respect to genotype).

6.1 Development of Tools and Theory: 1900–50

The first half of the twentieth century saw the development of some of the essential elements in the field—technical tools and population genetic theory. Serological (blood typing) techniques, developed by 1900, made possible the genotyping of the ABO locus and led to a decrease in the risk associated with blood transfusions. Initial studies of the ABO locus focused on the variation among United States' soldiers. The resulting data are recognized as the first population data for what are now called 'classical genetic markers': Blood groups, and protein and enzyme polymorphisms. The discovery of blood groups and protein polymorphisms continued slowly during this period.

Other dramatic developments relevant to the field of human evolutionary genetics prior to 1950 lay in the area of population genetic theory, 'the quantitative synthesis of Mendelism and Darwinism' (Provine 1971, p. 129). As early as the 1920s Sewall Wright, Ronald

A. Fisher, and J. B. S. Haldane were publishing in the field. Fisher's *The Genetical Theory of Natural Selection* (1930), Wright's 'Evolution in Mendelian Populations' (1931) and Haldane's *The Causes of Evolution* (1932) are part of a proliferation of theoretical work of this period. They were joined by Theodosius Dobzhansky, whose *Genetics and the Origins of Species* (1937) was just one of several major contributions. Together, these authors outlined the theory behind the roles of natural selection, mutation, migration, and random genetic drift in the genetic evolution of populations (see Provine 1971, for an accessible history of the origins of theoretical population genetics). These theoreticians continued contributing well into the second half of the century.

By 1950, a number of classical genetic markers had been studied across human populations (Cavalli-Sforza and Bodmer 1971). Variation within the Rh genetic system, discovered in 1939, was considered in light of population genetic theory, as researchers attempted to explain the existence of alleles that are clearly deleterious to individuals under certain conditions. While studies often focused on the health impact of genetic variation, there was also an anthropological component: Researchers sought to measure and explain the genetic differences among human groups. The elevated frequency of Rh—individuals among Basques, for instance, was initially seen as evidence that they were descendants of the earliest modern humans in Europe. In 1948, Linus Pauling and colleagues described the role of the different forms of hemoglobin in sickle cell anemia, developing a revolutionary technique called electrophoresis to distinguish the different forms. At about the same time, J. B. S. Haldane put forth a hypothesis to explain the high frequency of another hemoglobin disorder, alpha thalassemia. He suggested that the blood cells of individuals heterozygous (having two different alleles) at this locus were more resistant to attack by malarial protozoa. This hypothesis was subsequently extended to sickle-cell disease. The story of the advantage of the sickle cell allele in malarial regions remains one of the most convincing examples of the population genetic phenomenon known as heterozygote advantage. For a succinct summary of the research see Durham (1991).

6.2 Development of the Field: 1950–75

The third quarter of the twentieth century saw rapid development in several dimensions of human evolutionary genetics. In the mid-1950s, Mourant published his first *Distributions of Human Blood Groups* (1954) with tables of population frequencies for nine blood groups and the sickle cell trait. In the mid 1960s, Harris, Hubby, and R. Lewontin (see Lewontin 1974, for references) discovered that a large fraction of

proteins show variation within a species. This discovery led to a dramatic increase in the number of informative polymorphic systems typed across human populations: by the early 1970s, Cavalli-Sforza and Bodmer (1971) list 27 such systems. In 1972 Lewontin (1972) published a seminal article concluding that most human genetic variation is found within local human populations. Remarkably, today's more sophisticated techniques and larger data sets consistently confirm this original finding.

A number of research groups carried out detailed studies of nonWestern populations of particular interest to anthropologists at the time. J. Neel and colleagues focused on South American populations, particularly the Yanomama of southern Venezuela, beginning in the 1960s. They were surprised by the extent of genetic differentiation between Yanomama villages, attributing this to the small size of groups founding the villages. For details of this and other regional studies conducted during the 1960s, see Crawford and Workman (1973).

During this period new numerical methods, phylogenetic methods, and computer simulation were introduced and were immediately applied to the burgeoning set of data on classical genetic markers. In the early 1960s, L. Cavalli-Sforza and A. Edwards developed several methods for generating trees from population genetic data, and applied these to data for five blood groups for 15 human populations from around the world. While their initial results are not entirely consistent with results from larger, subsequent data sets, many of the features of that initial tree can be found in recent analyses as well. One surprising finding was that genetic and anthropometric traits (physical dimensions) gave very different pictures of differentiation among human groups. While anthropometric traits indicated a similarity between Africans and Australians, genetic traits (blood groups) indicated then, and still do, that these geographically distant populations are genetically distant as well.

Meanwhile theoretical population geneticists, particularly P. A. P. Moran and his students, as well as M. Kimura, continued to extend and develop models relevant to the field of human evolutionary genetics (Ewens 1979). In the context of new understanding about DNA and mutation, Kimura (1969) argued that most mutations, or variants, found in living individuals provide neither a survival/reproductive disadvantage nor advantage to the individuals whose genomes include these mutations. Instead, he claimed, these mutations are neutral, or nearly so. If most variation is neutral, then patterns of variation are to be interpreted in terms of population history rather than in terms of natural selection. During this period, statistical methods were developed to test the null hypothesis of no natural selection. The subsequent finding that the null hypothesis was difficult to reject (Lewontin 1974) had a direct impact on the field of human evolutionary genetics. It reinforced the idea

held by some geneticists that random genetic drift (random changes in allele frequencies over time) plays an important role in microevolution (see Kimura 1983, for discussion of controversy). While drift was certainly recognized as a powerful evolutionary force prior to Kimura's work, his 'Neutral Theory' of molecular evolution combined with the results of tests for natural selection led to a pendulum swing, with increased emphasis on drift (Weiss and Buchanan 2000). This was particularly the case for recently derived variation such as that found within the human species.

A number of research developments in biology, such as the discovery of the structure of DNA by J. Watson and F. Crick in 1953, had a significant impact on the field only much later, after DNA sequencing became sufficiently routine for population studies. The development of the clarifying concepts of cladistics by W. Hennig during the 1950s and 1960s had a much more immediate impact within systematics than within human evolutionary genetics. Nonetheless, Hennig's work has clearly had an influence in recent years, as phylogenetic methods using parsimony criteria have been applied to human DNA sequence data. In 1960, L. Pauling and E. Zuckerkandl introduced the concept of mutations occurring regularly enough at a particular nucleotide site to be considered 'clock-like.' To the extent that evolution at the molecular level is clock-like, researchers can infer dates, at least relative dates, given the number of mutational differences between DNA sequences.

6.3 Development of the Field: 1975–90

By the early 1980s, new techniques for assessing genetic variation supplemented serological and protein typing. Restriction enzymes were used to cut DNA segments at very precise locations. The sizes of the resulting DNA fragments varied depending on the underlying DNA sequence of an individual. Gel electrophoresis, introduced much earlier (1948), enabled researchers to compare and estimate fragment sizes. This approach led to an explosion in the number of known human polymorphisms during the 1980s. These were initially ascertained for medical genetic research (in the search for genes associated with disease), but were immediately recognized as valuable for human evolutionary genetics research. While thousands of polymorphisms were discovered, only a few hundred were actually studied in a geographically broad set of human populations.

During the 1980s DNA sequencing began replacing restriction analysis as the tool of choice within human evolutionary genetics. The initial focus was on the most highly variable regions of the mitochondrial genome, with its large number of copies per cell. The comparison of mtDNA sequences from numerous individuals throughout the world brought public

attention to the field. The public was captivated by the notion of 'the mother of us all,' as the mtDNA ancestor was labeled. Many people were also surprised by the interpretations of the patterns of mtDNA variation: That the common ancestors of living members of the human species lived relatively recently (less than 200,000 years ago rather than 1–2 million years ago (Klein 1999)) and that these ancestors probably lived in Africa.

7. Human Evolutionary Genetics at the Turn of the Millennium: Population Processes and Events

7.1 Human Genome Project

The rapid development of human evolutionary genetics over the final decade of the twentieth century owes a great deal to the Human Genome Project. Much of the technology developed under the auspices of that project (e.g., improved speed and accuracy of DNA sequencing methods) is currently applied in the field. In 1991, a group of human evolutionary geneticists led by L. Cavalli-Sforza proposed that the Human Genome Project be extended to include consideration of variation among individuals. Almost immediately, the proposal encountered criticism from groups concerned that geneticists might exploit individuals contributing DNA samples. Nonetheless, a subsequent Human Genome Project plan did include consideration of variation across individuals as a major goal. New methods of detecting variation among individuals were developed. The result is hundreds of thousands of nucleotide sites known to be polymorphic in humans across the entire human genome. Very few, however, have been studied in a broad set of human populations.

7.2 DNA of Ancient Peoples

Under some conditions (e.g., low temperature, aridity) DNA persists relatively intact within bones and teeth for thousands, if not hundreds of thousands, of years (Wayne et al. 1999). During the 1980s human evolutionary geneticists began to take advantage of this preservation of genetic information of the past. DNA from the bones of Neanderthal specimens, for example, has been compared with that of living humans in order to address the question of whether Neanderthals and anatomically modern humans interbred on a large scale within the last 100,000 years. The DNA from early anatomically modern humans promises to provide even further insight into human evolution during this time period (Relethford 2001). With respect to more recent human evolution, ancient DNA of Native Americans living prior to 500 years

ago has demonstrated no evidence of a dramatic loss of genetic variation in the intervening period, despite a dramatic reduction in the number of Native Americans (O'Rourke et al. 2000).

7.3 DNA of Living Peoples: Inferences Across a Range of Time Scales

While the DNA of ancient peoples has great potential to provide insight into human evolution, the DNA of extant peoples has been studied much more extensively (Cavalli-Sforza et al. 1994). The DNA of living people retains the signatures of events and processes at various time depths ranging from hundreds of thousands of years down to tens of years. Autosomal, Y chromosomal, and mtDNA of living people has shed light, for instance, on the peopling of Europe over the last 40,000 years. Researchers have estimated the relative contributions of Paleolithic and Neolithic Europeans to the present-day European gene pool. Independently evolving segments of the human genome are consistent in supporting the hypothesis of a relatively large contribution by Paleolithic Europeans to today's European gene pool, while Neolithic peoples from the Near East appear to have contributed roughly a quarter of that gene pool (Lell and Wallace 2000).

Two examples illustrate the potential for making inferences on a shorter time scale (a few thousand years at most). The social structure of the Hindu caste system, for instance, has been investigated using a combination of Y chromosome and mtDNA. Human evolutionary geneticists have compared the relative rates of male and female movement between castes, concluding that the caste system is driven by the social mobility of women (Bamshad et al. 1998). As a second example, oral traditions have been examined in light of DNA variation. According to oral history, the Lemba, a Bantu-speaking people of southern Africa, descend from Jews who moved from the Middle East to Yemen and then to southern Africa within the last 2,700 years. The finding that Lemba Y chromosomes are similar to those common amongst Jews and rare amongst the Lemba's African neighbors supports this tradition (Owens and King 1999).

7.4 New Theoretical Tools

While much of the population genetic theory and many of the data analysis tools relevant to human evolutionary genetics were developed prior to the 1990s, valuable new tools have been introduced more recently. Among these are Markov Chain Monte Carlo (MCMC) methods (see, for example, Griffiths and Tavaré 1995), and coalescent-based approaches (Kingman 1982). MCMC methods make possible

estimation of effective population size, population growth rate, migration rate, and mutation rate. The coalescent approach to inference involves considering the history, or genealogy, of a sample of present-day genes. The approach has been most valuable in efforts to date events of gene histories such as the age of the MRCA of all human Y chromosomes (Thomson et al. 2000). Most applications of the genealogical approach have been in the context of lengthy, non-recombining segments of the genome (i.e., mtDNA and the non-recombining portion of the Y chromosome). A major goal in the field today is to apply the same approach to a large number of segments of the autosomes and X chromosome.

8. Human Evolutionary Genetics at the Turn of the Millennium: Phenotypic Evolution

For an unknown number of segments of the human genome, current variation among human beings or among primates reflects the impact of natural selection. In these cases, while population processes and events have played a role, differences in fitness between individuals with different alleles have had greater influence. The most readily detected evidence of selection is in the context of disease, as exemplified above in the case of sickle cell anemia, malaria, and the hemoglobin genes. The field of human evolutionary genetics overlaps with medical genetics in that researchers in both fields seek to understand the origins of any genetic variation that appears to influence disease. The history of the gene associated with cystic fibrosis, for example, has been explored extensively.

A current focus of human evolutionary genetics is phenotypic traits other than disease, such as locomotion, dentition, and skin color (Weiss and Buchanan 2000). While little is known about the genetic diversity underlying skin color, variation at one gene, the melanocortin-stimulating hormone receptor gene, is known to be associated with red hair and fair skin that burns easily. Genetic diversity at this locus is significantly higher than the average nucleotide diversity in human populations and is likely to be an adaptive response to variation in environmental factors such as the availability of sunlight (Owens and King 1999).

Recently, researchers within human evolutionary genetics have begun focusing their attention upon the question, 'What makes us human?' Primate evolutionary geneticist Morris Goodman (1999) and others have argued for a 'Human Genome Evolution Project.' The goal is to identify features unique to the human genome that influence such traits as cognition, the form of the hind limb, and susceptibility to disease. There are approximately 40 million nucleotide differences between the human and chimpanzee genomes amongst the roughly 2×10^9 base pairs of non-

repetitive DNA. While most of these DNA level differences appear to confer no difference at the phenotypic level, a subset clearly does. The challenge is to identify that subset.

Known differences between humans and chimpanzees include chromosomal inversions and fusions, gene duplications, and repetitive element insertions (Varki 2000). Changes in the regulation of gene expression, however, are expected to be a more common source of phenotypic change (Varki 2000, Weiss and Buchanan 2000). Goodman and colleagues have described one example, wherein γ globin genes went from being expressed only within the embryo to being expressed primarily during the fetal stage. This change correlates with an evolutionary trend towards prolonged fetal life in the anthropoid primates (Goodman 1999). The search for other such changes is likely to occupy human evolutionary geneticists for the foreseeable future.

9. The Human Evolution Puzzle: Genetics Provides Some of the Pieces

The study of human evolution takes place within a broad range of academic disciplines. These include the fields of paleoanthropology and archaeology (Klein 1999), as well as such seemingly disparate fields as linguistics, developmental biology (Weiss and Buchanan 2000) and paleoclimatology (Hewitt 2000). Only within the context of findings in these fields can the study of genetic variation provide insight into some of the many details that constitute the evolutionary history of the human species. Human evolutionary geneticists face an exciting challenge as they attempt to integrate their findings regarding human evolution with those of counterparts in these linked disciplines.

See also: Cultural Evolution: Overview; Evolution, History of; Evolutionary Selection, Levels of: Group versus Individual; Genotype and Phenotype; Human Genome Diversity and its Value; Human Genome Project: German Perspective; Human Genome Project: Japanese Perspective; Population Dynamics: Theory of Stable Populations; Primates, Evolution of

Bibliography

- Bamshad M J, Watkins W S, Dixon M E, Jorde L B, Rao B B, Naidu J M, Ravi Prasad B V, Rasanayagam A, Hammer M F 1998 Female gene flow stratifies Hindu castes. *Nature* **395**: 651–52
- Cavalli-Sforza L L, Bodmer W 1971 *The Genetics of Human Populations*. W H Freeman, San Francisco, CA
- Cavalli-Sforza L L, Piazza A, Menozzi P 1994 *History and Geography of Human Genes*. Princeton University Press, Princeton, NJ

- Crawford M H, Workman P L (eds.) 1973 *Methods and Theories of Anthropological Genetics*. University of New Mexico Press, Albuquerque, NM
- Dobzhansky T 1937 *Genetics and the Origins of Species*. Columbia University Press, New York
- Durham W H 1991 *Coevolution*. Stanford University Press, Stanford, CA
- Ewens W J 1979 *Mathematical Population Genetics*. Springer-Verlag, New York
- Fisher R A 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, UK
- Fix A 1999 *Migration and Colonization in Human Microevolution*. Cambridge University Press, Cambridge, UK
- Goodman M 1999 The genomic record of humankind's evolutionary roots. *American Journal of Human Genetics* **64**: 31–9
- Griffiths R C, Tavaré S 1994 Simulating probability-distributions in the coalescent. *Theoretical Population Biology* **46**: 131–59
- Haldane J B S 1932 *The Causes of Evolution*. Harper & Row, New York
- Hartl D L, Clark A G 2000 *Principles of Population Genetics*. Sinauer Associates, Sunderland, MA
- Hewitt G 2000 The genetic legacy of the quaternary ice ages. *Nature* **405**: 907–13
- Kimura M 1969 The rate of molecular evolution considered from the standpoint of population genetics. *Proceedings of the National Academy of Sciences, USA* **63**: 1181–8
- Kimura M 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK
- Kingman J F C 1982 The coalescent. *Stochastic Processes and their Applications* **13**: 235–48
- Klein R G 1999 *The Human Career: Human Biological and Cultural Origins*. University of Chicago Press, Chicago, IL
- Lell J T, Wallace D C 2000 The peopling of Europe from maternal and paternal perspectives. *American Journal of Human Genetics* **67**: 1376–81
- Lewontin R 1972 The apportionment of human diversity. *Evolutionary Biology* **6**: 381–98
- Lewontin R 1974 *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York
- Mountain J 1998 Molecular evolution and modern human origins. *Evolutionary Anthropology* **7**: 21–37
- Mourant A E 1954 *The Distribution of Human Blood Groups*. Blackwell, Oxford, UK
- O'Rourke D H, Hayes M G, Carlyle S W 2000 Spatial and temporal stability of mtDNA haplogroup frequencies in native North America. *Human Biology* **72**: 15–34
- Owens K, King M-C 2000 Genomic views of human history. *Science* **286**: 451–3
- Provine W B 1971 *The Origins of Theoretical Population Genetics*. University of Chicago Press, Chicago, IL
- Relethford J H 2001 Ancient DNA and the origin of modern humans. *Proceedings of the National Academy of Sciences, USA* **98**: 390–1
- Simpson G G 1944 *Tempo and Mode of Evolution*. Columbia University Press, New York
- Thomson R, Pritchard J K, Shen P, Oefner P, Feldman M W 2000 Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. *Proceedings of the National Academy of Sciences, USA* **97**: 7360–5
- Varki A 2000 A chimpanzee genome project is a biomedical imperative. *Genome Research* **10**: 1065–70
- Wayne R K, Leonard J A, Cooper A 1999 Full of sound and fury: The recent history of ancient DNA. *Annual Reviews of Ecology and Systematics* **30**: 457–77
- Weiss K M, Buchanan A V 2000 Rediscovering Darwin after a Darwinian century. *Evolutionary Anthropology* **9**: 187–200
- Wright S 1931 Evolution in Mendelian Populations. *Genetics* **16**: 97–159

J. L. Mountain

Human Genome Diversity and its Value

1. Introduction

The era of modern genetics began with the rediscovery of Mendel's laws in 1900. However, long before this rediscovery, Charles Darwin (1859) recognized that variation in characteristics among individuals provided the basis of origin of species. Evolutionary biologists and population geneticists have long been interested in studying such variations, especially those that have genetic bases, and their origins and maintenance. Genomic diversity of a population is the raw material for its evolution by natural selection. Genomic diversity of a population can be measured in several ways. The most commonly used measure is the probability that two genes at any particular position of the chromosomes chosen randomly from the population are different. Of course, it is not practically possible to choose genes, since at any position of the chromosomes, two genes remain in combination in an individual. The above probability is equivalent to the probability that if an individual is drawn at random, this individual will possess a combination of two different genes at the chromosomal location under consideration. The different gene variants at any chromosomal location (locus) are called alleles. The combination of two alleles at a locus in an individual is called the genotype of the individual. When the genotype comprises the same two identical alleles, it is said to be homozygous; otherwise, heterozygous. The higher this probability, the greater is the genomic diversity. When all individuals of a population have identical genes, this probability attains the value of zero indicating that there is no genomic diversity in the population at that locus.

1.1 The Beginning of Human Genome Diversity Studies

The fact that individuals can differ in inherited traits was first clearly established with the identification of the ABO blood group system by Landsteiner in 1900. Soon after World War I, the Hirschfelds (1919) first discovered that human populations differ in frequencies of ABO blood group types. This finding marked the beginning of human genome diversity

studies. Subsequently, the development of biochemical- and molecular-genetic techniques have provided greater understanding and more refined estimates of human genome diversity.

1.2 The Importance of Genetic Diversity

It is crucial for any population to be genetically diverse. Humans, like many other species, are susceptible to various infections, many life-threatening. There is an increasing realization from recent research that genetic factors are involved in conferring susceptibility/resistance to most such infections. If all individuals in a population were to be genetically the same, then in the event of an outbreak of a new infection, it is possible for the entire population to be easily wiped out by the infectious agent if no individual carries genes conferring resistance to this agent.

2. Human Genomic Diversity at Various Levels

The development of starch gel electrophoresis by Oliver Smithies in 1955, permitted genomic diversity to be estimated at the level of proteins. Electrophoretic surveys of a variety of serum proteins and red-cell enzymes indicated extensive polymorphism among humans (Harris 1966). He showed that about 30 percent of human proteins are polymorphic, however, the genomic diversity estimated on the basis of such polymorphisms averaged between 10 percent and 20 percent. What this means is that at loci that determine various proteins and enzymes, there is more than one allele present in roughly 30 percent of such loci, however, at these loci, only about 10–20 percent of individuals are heterozygous on the average. Since electrophoresis can only detect such genomic variation that result in electrical charge change, and since only about 25 percent of genomic changes result in charge changes of the corresponding protein molecules (Nei 1975), this technique can only detect a small fraction of genomic changes. Hence, estimates based on data of such polymorphisms necessarily provide an underestimate of the true genomic diversity. There are, however, large differences in estimated diversities using different human proteins. These differences are, in part, due to differences in the degree of functional constraint of various proteins (Nei 1987). Not all proteins are functionally equally important. Changes in genes determining such proteins that are functionally important are generally more detrimental to the individual than changes in genes determining functionally less important proteins. Random changes, known as mutations, in genes create new alleles and enhance genomic diversity. Since mutations in genes of functionally important proteins are more detrimental, the new alleles created by such mutations

are quickly removed from the population by natural selection. Hence, genomic diversity is related to functional constraint.

The development of the Southern blotting technique in 1975, allowed polymorphisms to be studied directly at the DNA level. The first DNA variants using this technique were detected in the late 1970s as restriction fragment length polymorphisms (RFLPs). Such length polymorphisms are due to cleavage or noncleavage of DNA at particular sites, commonly caused by single nucleotide polymorphisms (SNPs), that are detectable using specific restriction endonucleases. RFLP data also showed significant genomic diversity in human populations. Nucleotide sequencing, however, is the most direct method to investigate genomic changes. While systematic nucleotide sequencing studies involving large segments of the human genome and a large number of individuals are just beginning, from the limited data that are already available it has been estimated that the average level of diversity per nucleotide varies between 0.0001–0.005 (see Chakravarti 1999 for a summary of recent results). Thus, in every 10,000 nucleotide positions along the human chromosomes, between 1 and 50 positions are estimated to differ between any two copies of that chromosome. This remarkable 50-fold difference in nucleotide diversity is again in part due to the fact not all regions of the genome are functionally equally important. Accumulation of mutations is tolerated in the functionally less important parts of the genome, but strongly selected against in the more important parts of the genome.

3. The Value of Human Genome Diversity in Understanding Human Evolution

With the passage of time, a population splits into subpopulations because of various cultural, economic, demographic, and ecological reasons. For example, after a period of demographic expansion of a population, subgroups of individuals may be compelled to move to another locality because of depletion of resources in the original habitat. Subsequent to this move, the original group and the newly founded subgroup may cease to have any genetic contact that is realized through exchange of mates. Generally, such subgroups are founded by a limited number of individuals, which impacts upon the genomic diversity of the subpopulation. When several such subpopulations are formed and remain isolated from one another so that individuals of one subpopulation do not exchange mates with other subpopulations, genomic diversification of the subpopulations is hastened. (Fusion of populations also occurs, but is possibly much less common than fission, especially in expanding populations.) Admixture, or the exchange of genes, increases the genomic similarity between two

populations. The primary barriers to admixture are cultural and linguistic differences and geographical distance. Thus, one expects genes, cultures, and languages to evolve in tandem. Indeed, genomic diversity depends on demographic size and history of a population. Recent advances in molecular and statistical techniques have shown, for example, that pairwise comparisons of sequence differences among people within a population can reveal past demographic patterns of the population (Slatkin and Hudson 1991, Rogers and Harpending 1992, Thomson et al. 2000).

3.1 Some Important Evolutionary Inferences

The above facts have been used to make important inferences on human evolution. By examination of worldwide human genetic variation, Cavalli-Sforza et al. (1994) have found a major concordance between language similarities and genetic affinities. Their studies have also revealed that the spread of agriculture was by human migration, rather than through the spread of customs and knowledge. The most interesting, and perhaps the most controversial, inference derived from genome diversity data has been that all anatomically modern humans evolved in a single geographical region, Africa. If one considers a segment of the genome on which the possibility of any strong effect of natural selection may be negligible, and if by comparing the nucleotide sequences of this segment in two individuals, one finds them to differ by a certain number of nucleotides, then it is possible to easily estimate the date of their common ancestor if the mutation rate (the rate at which nucleotide changes take place per year) is known and can be assumed to be constant over time. Based on nucleotide sequence data and several other lines of genetic evidence, it has been inferred that (a) modern humans are a relatively young population, (b) Africans are genetically the most diverse group of humans, (c) the largest genetic divergence among Old World populations is between Africans and non-Africans, and (d) modern humans arose in Africa between 200,000 and 100,000 years ago and dispersed throughout the Old World to replace archaic human species completely (Stoneking 1993, Nei 1995), although some geneticists (Ayala 1995) do not believe in complete replacement or even in the single-origin scenario for anatomically modern humans (Wolpoff 1999, Relethford 1995).

3.2 Human Genome Diversity Studies Can Answer Some Central Biological Questions

The nature and extent of genome diversity depend on various evolutionary forces, such as, mutation, natural selection, random genetic drift, admixture, and recombination. The relative importance of the various evolutionary forces in determining and maintaining

diversities within and between human populations remains poorly understood. However, from past studies of human populations with diverse structures, it is clear that random genetic drift (genetic events dictated primarily by chance that result from the small size of local groups) and the flow of genes from one population to another through migration or mate exchange are responsible for much of the patterning within human populations. No reliable estimate exists of the fraction of the human genome which is under the direct effect of natural selection. Put differently, how much of the observed variation in the human genome has a direct bearing on the fitness of individuals is largely unknown.

With the clearer understanding of the features of the human genome that is now emerging, thanks to the Human Genome Project, it will soon become possible to answer some of these central biological questions. For this, it will be necessary to carefully partition the genome into different equivalence classes of segments based on particular features, such as function, tissue specificity, gene density, distance from chromosome centromere or telomeres, etc., and then estimating within- and between-population diversities separately for each of these classes.

4. Human Genome Diversity and Molecular Medicine

The practice of medicine today focuses on 'treatment after disease onset'; the promise of molecular medicine is 'prevention before disease onset.' The hope for this promise is that genetic bases for most diseases will be discovered as a result of the spectacular advances in molecular genetics. Genes for diseases will be mapped on chromosomes; mutations precipitating clinical conditions will be identified. These discoveries will permit delineation of genotypes that are predisposed to specific diseases. Knowledge of the genotype of an individual will allow prediction of risk to disease. Preventive measures, that will not only include lifestyle modification but also genotype-specific therapy, will be undertaken for high-risk individuals.

4.1 Genome Diversity Studies Can Lead to Gene Discovery

Genomic diversity data on human populations can be used successfully in localizing on chromosomes the genes that cause inherited disorders, identifying predisposing/susceptibility factors to various diseases, and estimating relative roles of genetic and environmental factors in disease causation. In a large population, alleles at a pair of loci generally combine randomly into possibly genetic types. Thus, in a large population, there is usually no association between loci, unless they are physically very close on the same chromosome. However, in small isolated populations

or in recently admixed populations, strong associations between alleles at different loci are expected. Put differently, in such cases genetic loci behave non-independently in a statistical sense. This phenomenon is known as linkage disequilibrium. Disequilibrium decays every generation at a rate dependent on how far apart the genes are on the chromosome, which is measured by the rate of recombination between the genes. By examining patterns of linkage disequilibrium between diseases and genomic markers, especially in small isolated populations, it is possible to map loci controlling diseases. For example, in recently admixed populations, disequilibrium can be observed over chromosomal areas as large as 5–10cM, which is about 2–5 percent of the estimated linkage length of an average human chromosome (Morton 1991). These estimates, in turn, help identify mutations that cause various genetic disorders. Thus, genome diversity studies can lead to gene discovery.

Variations at the DNA level among individuals that occur, on an average, once in every 500–1000 base pairs (bp) along the 3 billion bp human genome can result in a wide range of susceptibility or resistance to diseases. Polymorphisms in different regions of the genome can interact among themselves and/or with the environment in conferring susceptibility/resistance. By DNA sequencing of appropriate genomic regions (e.g., those that are involved in detoxification pathway, signal transduction, etc.) in individuals exposed to various levels of pathogens and environmental toxins and correlating the information thus obtained with clinical outcomes, it is potentially possible to identify specific alleles as well as environmental exposures that precipitate disease. Genome diversity research will enable more precise identification of the interaction of genomic polymorphisms with environmental agents in disease causation. Such research can lead to implementation of public health programs for protecting susceptible populations and for targeting genetic screening to groups at greater risk to disease. Indeed, the focus is shifting from gene discovery to gene diversity.

4.2 Risk Assessment via Genome Diversity

One of the primary values of genome diversity research is that it will help define genetic polymorphisms and gene products that can predict development, progression, and severity of disease. While for rare diseases, single mutations with large effects are generally involved in the precipitation of disease, for common diseases, perhaps common variations in several genes are involved. These common variations singly do not significantly enhance the risk to a common disease, but jointly precipitate the disease condition. So the estimation of frequencies and effects of common genetic polymorphisms, particularly single nucleotide polymorphisms (SNPs), in defined populations, are expected to play an integral role in

molecular medicine. These estimates may, one day, enable high-risk patients to be diagnosed at a very early stage of disease and be treated with effective therapies that reverse the disease process or even prevent disease from occurring. Drugs can be targeted specifically to those genotypically defined patients who will respond well to the drug with no adverse side effects.

5. Debunking the Myth of Race

Humankind is a reproductive community, all members of which are connected by ties of mating and parentage. Yet, any two individuals, identical twins excepted, carry different sets of genes. To systematically study this enormous genomic diversity, biologists and anthropologists have described and named complexes of organisms to identify for themselves, and to let others know, what they are talking and writing about. Thus, classifying and giving names to classes began as a convenience-tool of science. Human biologists too have named classes of populations as ‘races.’ For example, Garn (1965) recognized nine ‘geographical races’ and 32 ‘local races’ of man. Most such racial classifications were based on visible characteristics, such as skin color, nose height, etc. Such characters are generally adaptive to environmental conditions. Further, such characters also influence mate choice. Therefore, racial differences that are recognized by such characters are really the products of natural and sexual selection. Darwin (1871) noted ‘Although the existing races of man differ in many respects ... yet if their whole organization be taken into consideration they are found to resemble each other closely in a multitude of points.’ Studies on human genome diversity have gradually made the ineptitude of typological approaches evident to increasing majorities of anthropologists. Yet, racial classifications are often used in sociopolitical debates in which certain races are claimed to be ‘genetically superior’ to others. Past studies on genetic diversity have shown that there is usually much greater intra-populational genetic variation than interpopulational variation (Lewontin 1972, Nei 1975, Barbujani et al. 1997). This means that a pair of individuals from a so-called ‘genetically superior race’ may be genetically more different than when one of them is compared with another from a ‘genetically inferior race.’ This is clearly in contradiction to the notion of genetic ‘superiority.’ Indeed, genome diversity studies have provided further evidence that has helped debunk the myth of races. Through their analyses of data compiled from genetic studies in global populations, Cavalli-Sforza et al. (1994) have shown that frequencies of genes often form geographical gradients, which make the task of drawing lines on a map to divide the regions of different races nearly impossible. Thus, race boundaries are never sharp. Worse still, gradients of frequencies of different genes often criss-cross on a

geographical map. When boundaries of gene frequency gradients are steep, they often correspond with linguistic boundaries or other similar cultural barriers to gene mixing. Thus, observed genetic differences among populations have no relation with genetic 'superiority' or 'inferiority.' Such differences are primarily the results of natural selection, barriers to genetic intermixture and also chance effects (random genetic drift). Human genome diversity studies have been valuable in debunking the myth of race.

See also: Evolution, History of; Evolutionary Social Psychology; Human Evolutionary Genetics; Human Genome Project: German Perspective; Human Genome Project: Japanese Perspective

Bibliography

- Ayala F J 1995 The myth of Eve: Molecular biology and human origins. *Science* **270**: 1930–6
- Barbujani G, Magagni A, Minch E, Cavalli-Sforza L L 1997 An apportionment of human DNA diversity. *Proceedings of the National Academy of Sciences USA* **94**: 4516–9
- Cavalli-Sforza L L, Menozzi P, Piazza A 1994 *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ
- Chakravarti A 1999 Population genetics—making sense out of sequence. *Nature Genetics Supplement* **21**: 56–9
- Darwin C 1859 *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. Murray, London
- Darwin C 1871 *The Descent of Man, and Selection in Relation to Sex*. John Murray, London
- Garn S M 1965 *Human Races*. Charles Thomas, Springfield
- Harris H 1966 Enzyme polymorphisms in man. *Proceedings of the Royal Society, London B* **164**: 298–310
- Hirschfeld L, Hirschfeld H 1919 Serological differences between the blood of different races. *Lancet* **ii**: 675–8
- Lewontin R C 1972 The apportionment of human diversity. *Evolutionary Biology* **6**: 381–98
- Morton N E 1991 Parameters of the human genome. *Proceedings of the National Academy of Sciences USA* **88**: 7474–6
- Nei M 1975 *Molecular Population Genetics and Evolution*. North Holland, Amsterdam
- Nei M 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York
- Nei M 1995 Genetic support for the out-of-Africa theory of human evolution. *Proceedings of the National Academy of Sciences USA* **92**: 6720–2
- Relethford J H 1995 Genetics and modern human origins. *Evolutionary Anthropology* **4**: 53–63
- Rogers A, Harpending H C 1992 Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution* **9**: 552–69
- Slatkin M, Hudson R R 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**: 555–62
- Stoneking M 1993 DNA and human evolution. *Evolutionary Anthropology* **2**: 60–73
- Thomson R, Pritchard J K, Shen P, Oefner P J, Feldman M W 2000 Recent common ancestry of human Y chromosomes:

Evidence from DNA sequence data. *Proceedings of the National Academy of Sciences USA* **97**: 7360–5

Wolpoff M H 1999 The systematics of *Homo*. *Science* **284**: 1774–5

P. P. Majumder

Human Genome Project: German Perspective

1. The Human Genome Project

The Human Genome Project differs from any previous biological or medical project in size and cost. Its ambitious goal is the deciphering (sequencing) of all 3 billion building blocks of our genetic make-up—the so-called DNA—by 2005, the identification of all genes encoded in that DNA, and the understanding of the role of these genes in health and disease. The knowledge of these genes and their function is crucial for basic biological research as well as for the improvement of prevention, diagnostics, and therapy of disease. It is the prerequisite for a targeted design of pharmaceuticals and for novel approaches like gene therapy. This knowledge poses hope to millions of affected people and contains also an immense economic potential.

The Human Genome Project was started in the USA in 1990, James Watson, the codiscoverer of the DNA structure being its first co-ordinator. It was clear from the beginning, due to the estimated cost of US \$ 3 billion and the immense amount of work involved, that the Human Genome Project had to include many countries. The Human Genome Organization (HUGO), an independent international organization of genome scientists, was established to co-ordinate the duties. The US Human Genome Project started off with considerable public funding (US \$ 87 million in 1990) and was soon followed by the UK and France. Between 1991 and 1996 France contributed a comprehensive genetic map of the human genome. It is noteworthy that this work was predominantly financed by private money from a patients' association. The British Wellcome Trust for its part set up the world's largest sequencing facility, the Sanger Centre. Smaller initiatives later emerged, for instance in Japan and Canada. No such activity, however, was seen in Germany until 1995. By 2001 a nearly complete 'working draft' of the human genome has been presented by the publicly funded Human Genome Project, including a significant German contribution. The complete sequence will be available in public databases some time ahead of schedule, probably by 2003. In 1998 an emerging competition to the Human Genome Project by private companies, namely by

Craig Venter's Celera, had speeded up the deciphering of the human genetic code tremendously. But the knowledge of our comprehensive genetic make-up is not considered as a benefit for mankind by everybody. Profound ethical issues are raised by the possibilities inherent in this knowledge. This had been realized by the founders of the project and therefore, about 3 percent of the budget was dedicated to exploring the ethical, legal, and social implications of the Human Genome Project. In most countries participating in the Human Genome Project, an extensive discussion took place about the opportunities and risks attached to these novel technologies. In Germany a parliamentary committee was established to elucidate the topic. But the discussion was much more polarized in Germany than in other countries.

2. Cutting Edge Technology is Considered a Necessity for Germany's Industrialized Society

Germany is a strongly industrialized country. Due to its lack of raw materials, its economy is heavily dependent on export-oriented manufacturing industries. To reach and maintain its high standard of living, and to survive on the world market, Germany is in permanent need of cutting-edge technologies. Traditionally Germany's strength lay in the engineering and chemical industries. Other fields like agriculture and coal mining have undergone deep structural crises and are nowadays only viable on the basis of high subsidies. This is particularly true for rural areas or areas that have been focused on one industrial branch for generations. German labor law and the strong position of German trade unions protect employees from unemployment and provide broad social protection, but on the other hand slow down structural changes. Revolutionary technical developments can give a well-developed industry an opportunity to solve structural problems and to gain new markets. During the 1970s and 1980s, progress in microelectronics and computer technology constituted such a revolution. But neither industry nor academia realized this development in time so that it was subsequently missed in Germany. During this time, future prospects for highly qualified young scientists became increasingly worse. Germany had to face a severe 'brain drain,' predominantly to the United States. The structural problems dramatically deteriorated following German reunification—the entry of the Five New *Länder* (states) of the former German Democratic Republic to the Federal Republic of Germany, which became known as 'German Disease' in the 1990s. The economic recovery—the so-called *Wirtschaftswunder*—of West Germany after World War II, often promised to East Germans after reunification by West German politicians, was not easy to repeat. The resulting increasing economic pressure led to a resumption of the debate about the

pros and cons of biotechnology in general. After the end of the Cold War, other ethically controversial economic sectors such as the arms industry became less significant. Up to this time there had been no co-ordinated funding for biotechnology research, let alone any participation in the Human Genome Project, although the scientific basis was established and many technical developments pointed the way ahead. Though among a few scientists discussions on the need for genome-wide analysis had started in Germany around 1980, leading up to a proposal to sequence the entire genome of the bacteria *Escherichia coli* or even human chromosome 21 at the European Molecular Biology Laboratory in Heidelberg where much of the early technology had been developed, both the German scientific community, as well as funding agencies were at this point by and large not able to recognize the immense potential of genome-wide analysis.

While a large number of new powerful biotechnology companies emerged in the United States during the 1980s, recombinant pharmaceuticals had to be imported to Germany. The first version of the German Genetic Engineering Law was very restrictive and bureaucratic. It took the Frankfurt-based company Hoechst (now Aventis) years to get the approval to produce recombinant human insulin in their brand-new facilities. As a consequence, most large German pharmaceutical companies transferred their research departments to locations in the United States. Politicians realized very late the immense economic damage caused by this development. To take countermeasures two large co-ordinated research programs were started by the Federal Ministry of Research—the German Human Genome Project (DHGP) in 1995 and the BioRegio Competition in 1996.

3. Genomic Research Against the Background of German History

To fully understand the assessment of and the reservation about genetics in German society, it is helpful to take a closer look not only at economic aspects but also at intellectual history and the National Socialist period.

3.1 The Idea of Eugenics

At the beginning of the twentieth century various political and intellectual movements in search of a new order existed in Germany. As in many other countries at that time many of these were characterized by strong nationalism and/or antisemitism. Some of them had their roots in German romanticism or naturalism: movements like the *Wandervögel*—a rambler's association and precursor of a new German youth movement expressing a strong love of nature.

This milieu, in combination with vague Darwinian ideas, ushered in the idea of improvement of the human race and prevention of its degeneration under the overwhelming influences of a modern and industrialized world: the idea of eugenics originally introduced by Galton in the 1870s. Many of these movements were infiltrated by National Socialists and forced to take the party line during the *Gleichschaltung* after the Nazis had seized power in 1933. Fragments of these simplistic ideas were soon adopted by the Nazis. When the philosopher Nietzsche talked about the 'overman,' a superior human, it was interpreted that an inferior human was the conclusive consequence. Eugenics was not a purely German affair. Many countries, including the United States, had a strong eugenics movement during that time. In none of these countries, however, was eugenics used as a justification for profound discrimination, leading to the Nuremberg Laws of 1935 and eventually the extermination of some 6 million Jews by 1945.

3.2 The Role of German Physicians and Geneticists During the National Socialist Period

The role of German physicians and geneticists during the Nazi period is complex and exceeds by far the scope of this article. It should be sufficient to illustrate their influence by some examples. After the rediscovery of Mendel's Rules many scientists were worried that mental diseases like schizophrenia or idiocy, or the mixing of the white 'race' with individuals considered to be inferior (e.g., members of ethnic groups from the former colonies) would lead to a decline of European culture. Already during the first two decades of the twentieth century, the superiority of the white race to other humans and the destruction of so-called 'life unworthy to be lived' was discussed openly even in international scientific circles. Eugen Fischer, a scientist at the University of Freiburg, published a book in 1913 in which he addressed 'the problem' of mixed marriages between white men and Hottentot women and considered the progeny of such marriages as nonviable without protection by a superior race. Many of the racist ideas of Fischer and others were ingested in the emerging National Socialist ideology at that time. In 1927 Fischer became the director of the newly founded Kaiser Wilhelm Institute (KWI) for Anthropology, Human Heredity, and Eugenics in Berlin; in 1939 he was elected president of the University of Berlin. In this latter position he was responsible for signing the dismissal notices of his Jewish colleagues. A motivation for most of the scientists was an almost paranoid anxiety of the decline of their own culture and 'race.' In many cases, personal advantages that derived from the dismissal of their Jewish colleagues might have encouraged this development. Scientists supporting the National Socialist idea of racial hygiene like Fischer, Rüdin of the KWI of Psychiatry, Hallervorden of the KWI of Brain Research, or von

Verschuer, Fischer's successor as the director of the KWI for Anthropology, were welcomed by the leaders of the Nazi party. Most of the scientists involved, however, considered themselves as being apolitical. From their point of view they supported the system for the sake of science. Eugen Fischer wrote in a German newspaper: 'It is a rare and special good fortune for theoretical science to flourish at a time when the prevailing ideology welcomes it and its findings can immediately serve the policy and the state' (*Deutsche Allgemeine Zeitung*, March 28, 1943). Since most of these scientists did not have an official post in the Nazi administration and were not actively involved in the killing of people, only a few had to flee from Germany after the end of World War II. Josef Mengele, known as the 'Angel of Death' of Auschwitz, managed to escape to Brazil where he lived unmolested probably until 1979. Those who had deliberately worked with human experimental material like eyes, brains, etc., gained from concentration camps or psychiatric hospitals or had asked for this material from their co-workers, mostly avoided prosecution.

3.3 Germany Dealing with its Difficult Past

After World War II the faith of society in a detached and neutral science was profoundly disturbed. Genetics came under special suspicion because its advocates had provided 'scientific' basis for the National Socialist racial hatred. Scientific terminology had been adopted and occupied by the Nazis. As the Nobel Prize laureate Christiane Nüsslein-Volhard puts it quoting the term 'selection,' 'Scientists think of Darwin's fundamental theory of the origin of species while German society even today is reminded of the gas chambers of Auschwitz.' The reasons for this are diverse and many. Most geneticists did not differentiate themselves properly from the National Socialists. The borderline between fanatic followers and murderers like Josef Mengele, opportunists who profited from the system like Eugen Fischer and Otmar von Verschuer (who was Mengele's superior in Berlin), and those who had remained in their positions with tacit permission, remained hazy. After 1945 no genuine reappraisal of the role of physicians and geneticists during the Nazi period occurred. Directly after the war, Germany was too busy working on its economic recovery to ask questions about guilt. Many members of the intellectual elite had either emigrated or had been killed during the Nazi period. But many key proponents of the genocide reappeared in important academic positions. Otmar von Verschuer became professor of genetics at the University of Münster where in 1954 he even rose to the position of Dean of the Medical Faculty. The Kaiser Wilhelm Society was renamed Max Planck Society (MPS), and Adolf Butenandt, the Nobel Prize laureate of 1939 who was sometimes criticized for not differentiating himself clearly from the Nazi regime, became its first president.

Many of the experiments that were committed in Auschwitz and elsewhere were financially supported by the Kaiser Wilhelm Society or the Deutsche Forschungsgemeinschaft (DFG), the national German Research Council. It was, however, not before the late 1990s that DFG and MPS initiated an independent review of their own role during the NS period.

Because many natural sciences had fallen into disrepute after 1945, public discussion was completely dominated by intellectuals of the social sciences. This was in sharp contrast to the period from the turn of the century until 1933 when, for instance, physicists like Einstein, Heisenberg, and Planck were actively involved in the social and political debate, and disputes among artists, philosophers, theologians, and scientists were common. After 1945, genetics had not become common knowledge in Germany even in intellectual circles. Instead, a vague fear of genetic reductionism and determinism remained. This led Robert N. Proctor, a US historian of science, to state: 'The legacy of Nazism influences the attitudes many Germans show towards genetics today and is probably responsible for the absence of a prominent science-oriented intellectual elite. Where are the Goulds, Hawkings and Dawkins of Germany?' The polarization between social scientists and natural scientists who were mainly seen as technocrats still persists. The seizure of power by man over himself created the fear that humanity could be the first victim.

3.4 The German Human Genome Project

During the 1980s the tremendous successes of genetic engineering boosted the biotechnology industry in the USA and provided the prerequisite for the idea of the Human Genome Project. While in the US, companies like Amgen turned over more than US \$ 1 billion per year, the production of recombinant pharmaceuticals was still hampered in Germany. At the same time, recombinant pharmaceuticals and vaccines had to be imported into Germany. One of the first dedicated centers for genome research in Germany, the *Genzentrum* in Munich had bulletproof windows installed in fear of an attack. The international Human Genome Project was started in 1990 without any German participation. But increasing economic pressure forced Germany to a rethinking of genome research. After a thorough investigation of the issue by a parliamentary commission, two programs of the Christian Democratic-Liberal coalition started in the mid-nineties. The BioRegio Competition aimed at exploiting the synergy of self-organization of whole regions as centers of excellence for biotechnology and was unexpectedly successful. The German Human Genome Project (DHGP), integrated into the international effort to decipher the human genome, was thematically much more focused. It was strongly

industry-oriented and a special infrastructure was to guarantee the effective transfer of the knowledge gained to industry. Initial efforts to raise half of the budget from the German pharmaceutical industry failed. Subsequently, the whole budget of DM 40 million per year was granted by the Federal Ministry of Education and Research. To safeguard the interests of industry, a Patent and Licensing Agency (PLA) was established whose task was to screen the results of the projects granted within the DHGP and to guarantee a patent application if applicable. The patent is granted to the inventor at no cost. In turn, the Industry Association gets the right of first refusal on the patent.

The agreement also gave the Industry Association the right to screen the data produced in the German DNA sequencing centers—a clear infringement of the so-called Bermuda Principles—an agreement between all academic sequencing centers worldwide making all data immediately publicly available without any restrictions. The Industry Association waived their claims only after massive pressure from the international scientific community.

The DHGP has had a profound effect on the German biotechnological industry. Contrary to the expectation that the big pharmaceutical companies would exploit the scientific results, licensing of patents was sparse. Instead, many researchers preferred to start their own companies, a tendency up to then only rarely found in German academia. The success of the DHGP initiative led to Germany exceeding the United Kingdom in the number of biotech companies by the year 2000.

Initial concerns among scientists that funding of genome research would be cut after the change of government to a coalition of Social Democrats and the Green Party were groundless. Not only has the DHGP been maintained, but also a plant genome project—initiated by the prior administration—as well as a microorganism genome project were started, assisted by additional projects on bioinformatics and proteomics. In 2000, a long-expected concept document was released by the Federal Ministry of Education and Research. This paper draws up a long-term perspective for genomic research in Germany and safeguards the existence of central infrastructure units (resource centers) for genomic research. At the same time, boosted by the announcement of the draft version of the human genome, a tremendous increase in funding was approved that eventually brought Germany in line with genomic funding in countries like Japan, UK, and France.

The economic success as well as the prospects of new diagnostic and therapeutic possibilities has very much changed the German attitude towards human genetics. Medical genetics is now much more accepted than it was a few years ago. In contrast, due to concerns about health as well as to the aggressive marketing strategies of certain agrochemical biotech companies, plant genetic engineering has met with

deep disapproval in Germany. However, the fatalistic attitude in Germany towards fundamental natural sciences has somewhat changed. The former lack of discussion in society that permitted simplistic arguments to gain predominance has been overcome. The majority of the German intellectual elite has realized that an informed discussion within society is needed to solve the problems stimulated by genome sciences.

See also: Ethical Issues in the 'New' Genetics; Eugenics, History of; Eugenics in Europe; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Screening for Disease-related Characteristics; Genetic Testing and Counseling: Educational and Professional Aspects; Human Evolutionary Genetics; Human Genome Diversity and its Value; Human Genome Project: Japanese Perspective

Bibliography

- Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Morissette J, Weissenbach J 1996 A comprehensive genetic map of the Human Genome based on 5,264 microsatellites. *Nature* **380**: 152–4
- Ernst & Young 2000 *Gründerzeit; Zweiter Deutscher Biotechnologie-Report 2000*. Ernst & Young, Stuttgart, Germany
- Hammerstein N 1999 *Die Deutsche Forschungsgemeinschaft in der Weimarer Republik und im Dritten Reich; Wissenschaftspolitik in Republik und Diktatur*. Beck, Munich, Germany
- International Human Genome Sequencing Consortium 2001 Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921
- Kaufmann D (ed.) 2000 *Geschichte der Kaiser-Wilhelm-Gesellschaft im Nationalsozialismus; Bestandsaufnahme und Perspektiven der Forschung*. Wallstein-Verlag, Göttingen, Germany
- Klee E 1997 *Auschwitz, die NS-Medizin und ihre Opfer*. Fischer Verlag, Frankfurt, Germany
- Laubichler M D 1999 Frankenstein in the land of Dichter and Denker. *Science* **286**: 1859–60
- Müller-Hill B 1998 *Murderous Science: Elimination by Scientific Selection of Jews, Gypsies, and Others in Germany, 1933–1945*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
- Proctor R N 2000 *Adolf Butenandt (1903–1995). Nobelpreisträger, Nationalsozialist und MPG-Präsident; Ein erster Blick in den Nachlaß. Ergebnisse 2*. Präsidentenkommission (Max Planck Society) Geschichte der Kaiser-Wilhelm-Gesellschaft im Nationalsozialismus, Berlin
- Sachse C, Massin B 2000 *Biowissenschaftliche Forschung in Kaiser-Wilhelm-Instituten und die Verbrechen des NS-Regimes; Informationen über den gegenwärtigen Wissensstand. Ergebnisse 3*. Präsidentenkommission (Max Planck Society) Geschichte der Kaiser-Wilhelm-Gesellschaft im Nationalsozialismus, Berlin
- Schiermeier Q 1997 German truce on genome data access. *Nature* **388**: 409
- Schmuhl H-W 2000 *Hirnforschung und Krankenmord; Das Kaiser-Wilhelm-Institut für Hirnforschung 1937–1945. Ergebnisse 1*. Präsidentenkommission (Max Planck Society) Geschichte der Kaiser-Wilhelm-Gesellschaft im Nationalsozialismus, Berlin

- Unterhuber R 1997 Germany gets warning on access to sequence data. *Nature* **387**: 111
- 1997 Germany may relax genome data policy. *Science* **276**: 1323
- 2000 Human Genome Project makes its mark in Germany. *Nature* **406**: 821

J. Maurer and H. Lehrach

Human Genome Project: Japanese Perspective

The Japanese Human Genome Project (JHGP) was initiated in 1988. The project experienced some difficulties in the initial stages, such as scarcity of resources and the perception that genome research was not hard science. As a result the JHGP was unable to follow the specific aims of sequencing DNA of the human genome. This article explains how the JHGP adopted strategies to overcome these problems and how it became an effective contributor to the international human genome project.

1. History of the Human Genome Project in Japan

The drive toward an international approach to the Human Genome Project began in 1985. In 1988, as a response to this development and the progress of the human genome project activities in the United States and Europe (see *Human Genome Project: German Perspective*), the Japanese human genome project (JHGP) conducted a feasibility study. The study was supported by a general research grant from the Ministry of Education, Science, Sports and Culture (ESSC: MONBUSHO). As a result of this study, the research council of MONBUSHO proposed the formulation of a special human genome research grant and a five-year plan for the promotion of genomic research.

The genome research supported by this special grant had a project-like characteristic, and because of this was confronted by criticism from various sources. There was a strong belief among certain groups that this kind of project was not 'hard research' but rather mere routine work. Moreover, difficulties were experienced in relation to the scarcity of technicians skillful in this area who could support the genome researchers and also in the reluctance shown by the government and private institutions to invest heavily in a single project. As a result of these difficulties

Table 1

Mass DNA data submission to DDBJ (submitted and distributed January 1998–December 1998)

Organism	Accession number (number of entries)	Submitter (institution)
BCT (bacteria)		
<i>Escherichia coli</i>	ABO11548-ABO11549 (2)	Makino, K. (Osaka University)
<i>Unidentified alpha proteobacterium</i>	ABO11682-ABO11791 (110)	Shikano, S. (Tohoku University)
<i>Unidentified nitrogen-fixing bacteria</i>	ABO11841-ABO11964 (124)	Ohkuma, M. (RIKEN) ^a
<i>Bacillus stearothermophilus</i>	ABO20064-ABO26165 (102)	Matsuura, T. (Osaka University)
<i>Plasmid colIb-p9</i>	ABO20234-ABO20235 (2)	
	ABO21078 (1)	Sampei, G. (The University of Electro-communication)
HUM (human)		
<i>Homo sapiens</i>	ABO14511-ABO14610 (100)	Ohara, O. (Kazusa DNA Research Institute)
	ABO18254-ABO18353 (100)	
	ABO20626-ABO20725 (100)	
<i>Homo sapiens</i>	ABO19437-ABO19441 (5)	Matsuda, F. (Kyoto University Graduate School of Medicine)
<i>Homo sapiens</i>	ABO20858-ABO20878 (21)	Hirakawa, M. (JST) ^b
HTG (high throughput genomic sequencing)		
<i>Arabidopsis thaliana</i>	ABO10068-ABO10077 (10)	Nakamura, Y. (Kazusa DNA Research Institute)
	ABO10692-ABO10700 (9)	
	ABO11474-ABO11485 (12)	
	ABO12239-ABO12248 (10)	
	ABO13387-ABO13396 (10)	
	ABO15468-ABO15479 (12)	
	ABO16870-ABO16893 (24)	
	ABO17059-ABO17071 (13)	
	ABO18107-ABO18121 (15)	
	ABO19223-ABO19236 (14)	
	ABO20742-ABO20755 (14)	
VRL (viruses)		
<i>Hepatitis G virus</i>	ABO10155-ABO10202 (48)	Tanaka, Y. (Nagoya city University Medical School)
<i>Human immunodeficiency virus type 1</i>	ABO14775-ABO14874 (100)	Sato, H. (National Institute of Infectious Diseases)

a The Institute of Physical and Chemical Research Japan. b STA, Japan.

JHGP could not follow the specific aim of sequencing the DNA of the human genome.

Instead, JHGP adopted the strategy of developing excellent genome-related research. By this strategy, JHGP successfully contributed to the construction of human genome physical maps, the development of full-length cDNA construction techniques, and the making of cDNA expression maps, leading to an enhancement of human genome research in Japan. However, JHGP was far behind the United States and Europe, particularly in the field of genetic disease analysis based on human genome mapping. More specifically, as most of the major human genome projects were still in their infancy, JHGP was behind in terms of planning rather than the research itself.

Although, for Japan, this predicament was partly due to some researchers failing to recognize the

importance of the human genome project, the main reason was the failure to establish a single human genome research center coordinated by the relevant ministries and institutions. For instance, JHGP was planned independently by two ministries and a government agency. These bodies included the MONBUSHO, comprising many University researchers, the Ministry of Health and Welfare (MHW) supervising medical activities, and the Science and Technology Agency (STA) supporting various large-scale projects. As a result of the lack of collaboration between these government organizations, talented researchers, useful materials, and research money were unnecessarily dispersed. Moreover, because of the slow response of bioindustries to the human genome project, government agencies were unable to formulate a collaborative liaison with them in Japan.

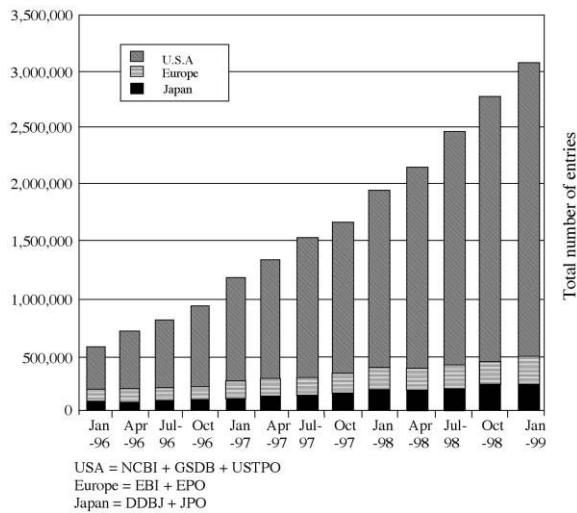


Figure 1

Contribution of the three geographical areas to the DDBJ/EMBL/Genbank International Nucleotides Database (January 1996–January 1999)

These problems were essentially based upon the political and social infrastructures within Japan; however, in spite of these difficult situations, several success stories emerged from JHGP. For example, seven research groups in Japan successfully contributed to the completion of the *Bacillus subtilis* (*B. subtilis*) genome sequence by determining the nucleotide sequences of its shared region of 1.3 Mb. The complete genome sequence of *Synechocystis* sp. was determined by the Kazusa DNA Research Institute, which was established by Chiba prefecture and industries. The Ministry of Agriculture, Forestry and Fisheries of Japan (MAFF) launched the rice cDNA project, which is currently leading international research on the rice genome. In 1998, the *Escherichia coli* (*E. coli*) genome project team in Japan successfully completed three quarters of the *E. coli* genome sequence, although the American team completed the whole *E. coli* sequence in the same year.

Genome Informatics was initiated from scratch because there were no major research activities which could act as a focal point, except the DNA databank of Japan (DDBJ) within the National Institute of Genetics (NIG). In particular, the recruitment of talented researchers, education and training of research students, installment of a computer network for genome research, and the establishment of a DNA database had to be done within a short period of time. Although these were successfully dealt with to some extent, there were still some problems to be solved such as intimate collaboration among the different institutions in charge of various bioinformatics activities.

2. The Present Situation of the Human Genome Project in Japan

As genome research advances internationally, there has been increasing recognition that genome research constitutes the necessary foundation for the future development of biology, medical research, and biotechnology. In particular, the necessity of establishing a large-scale genome-sequencing center has been accepted widely. However, the realization of this plan has been difficult because of budget limitations.

Since the basic law of science and technology development was passed in 1997, the budget limitations have, to some extent, disappeared under the national goal that Japan should become a country of creative science and technology. Because of this basic law, the government invested approximately \$US 200 million in the genome research project in the 1998 fiscal year alone. Presently, various activities in JHGP are conducted under the initiative of governmental organizations and private industries. MONBUSHO successfully established the Center for Information Biology in NIG to support DDBJ and the newly established Human Genome Research Center in Tokyo University. The Ministry of International Trade and Industry (MITI) and the MHW coordinated a promotion business venture of genome-related enterprise and the STA created the genome science center in Yokohama. Moreover, large research grants were made to many genome research groups through various funding mechanisms. However, not all of these activities are along the lines of well-planned strategies of genome research.

The general focus of JHGP is toward the determination of large-scale genome sequences and a functional analysis of the genes that will be discovered by sequence determination. JHGP is placing more emphasis on the areas of genomic structure, function, and information. As for genome sequence determination, Kitazato University, Japan Marine Science and Technology Center, and STA and the Product Evaluation Center under MITI are pushing the genome sequencing of microorganisms by producing 122 Mb each year. In 1995, the Japanese Science and Technology Corporation initiated the human genome-sequencing project. The subgenomic segments such as HLA, immunoglobulin variable, Down's syndrome, and oncogene-related regions, have been determined for about 15 Mb by Kitazato University, Cancer Institute, Keio University and Tokai University.

In total, the sequence production in Japan is about 5 Mb per year, which is one order of magnitude smaller than that of the Sanger Center in the United Kingdom and the Sequence Center at Washington University in the United States. Therefore, JHGP has great expectations of the role of the newly established Genome Science Center, which is planned to produce 30–50 Mb per year. Kazusa Research Institute is producing several Mb genome sequences of *Arabidopsis* per year.

Moreover, the genome project team supported by MAFF has a ten-year plan of sequence determination of the complete rice genome of 400Mb, producing 20Mb per year for the initial several years. In summary, the sequencing capacity of genomes in Japan has so far been 15–20Mb per year. Judging from recent increases of the number of genome research groups focusing on bacteria, fungi, lower organisms, insects, and fish and the enhancement of genome research on humans and the mouse, 200Mb sequence production per year is expected in the near future.

JHGP has a slight edge over other countries in the field of functional genomics. In particular, the Osaka University genome team was the first to analyze a cDNA expression profile of human genetic data (called 'body mapping') and other genome teams have quickly developed a full-length cDNA determination technique. Although these activities were not directly connected to a bioindustry as they are in the United States, MITI recently pushed the utilization of these technical advances to industry. Moreover, human genome diversity has become recognized as important for disease gene analysis and functional analysis. In particular, the so-called 'complex diseases' such as multi component diseases are becoming important targets of the human genome diversity project (see *Human Genome Diversity and its Value*). This project will require a lot of genetic data concerning the Japanese population, in particular a large amount of polymorphism data from several thousand people (see *Race: Genetic Aspects*).

Several preliminary research projects have already started in Japan. The cDNA project of *Caenorhabditis elegans* (*C. elegans*) genome has been conducted by NIG in cooperation with the Sanger Center and Washington University. The *B. subtilis* and *E. coli* genome teams have already begun functional analysis of all genes contained in their respective organisms. Many researchers are participating in these genome projects by conducting bioinformatics research and constructing databases. Thus, JHGP has resumed its original goals in a more integrated and coordinated way.

3. The Future Challenge of the Human Genome Project in Japan

In Japan, the aim of the human genome project was redirected to not only sequence the human genome but also to include the functional analysis of genes as well as the elucidation of protein tertiary structures. Although this redirection has successfully eased various criticisms, the main focus of the human genome project has been diffused to a considerable extent, leading to an ambiguity of the Japanese contribution to the international effort of human genome sequencing. Given the expectation that human genome sequencing

has been achieved mainly by the United States and Europe, the Japanese human genome project will be successful if it orients its major effort towards the promotion of functional genomics, comparative genomics, evolutionary genomics, and genomic diversity analysis, along with the development of DNA chips or microarrays.

Conducting functional genomics requires analytic technologies that are capable of dealing with a large number of genes and proteins. Toward meeting these requirements, genome researchers in Japan have started contributing to the development of differential display, DNA chips and the identification of variant DNA. In the near future, these efforts should be connected to advancement of industrial technologies to produce unique and significant methodologies useful to functional genomics.

For genome informatic research, DDBJ at CIB/NIG, the Human Genome Research Center at Tokyo University, the Chemistry Institute at Kyoto University, and other research groups are making efforts to incorporate information science into genome research, in the belief that 'genome science is virtually information science.' There is no doubt that informatic aspects are extremely important to genome research. In particular, the elucidation of gene networks will be crucial for understanding the development and evolution of organisms, and in this sense, bioinformatics will play a crucial role. Another aim of JHGP is the allocation of talented bioinformatics and the education of students who understand both biological and information sciences. This will be an important component of the future activities of the JHPG.

See also: Bioethics: Examples from the Life Sciences; Bioethics: Philosophical Aspects; Brain, Evolution of; Evolution, History of; Genes and Culture, Coevolution of; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Studies of Behavior: Methodology; Genetic Testing and Counseling: Educational and Professional Aspects; Human Evolutionary Genetics; Human Genome Project: German Perspective

Bibliography

- Bellgard M, Itoh T, Watanabe H, Imanashi T, Gojobori T 1999 Dynamic evolution of genomes and the concept of genome space. *Annals of the New York Academy of Science* **870**: 293–300
- Gojobori T 1999 Evolution. In: Creighton T E (ed.) *The Encyclopedia of Molecular Biology*. Wiley, New York
- Gojobori T 1999 DNA sequence analysis. In: *Encyclopedia Life Sciences*. Macmillan Reference Ltd
- Normile D 1997 Japan centers in on genome work. *Science* **278**: 1700–2

Sugawara H, Miyazaki S, Gojobori T, Tateno Y 1999 DNA Data Bank of Japan dealing with large-scale data submission. *Nucleic Acids Research* 27: 25–8

T. Gojobori

Human Resource Management, Psychology of

Human resource management (HRM), as practiced in work organizations, includes five broad activities that together constitute the HRM system: staffing, retention, development, adjustment, and managing change (Cascio 1998). *Staffing* comprises the activities of (a) identifying work requirements within an organization, (b) determining the numbers of people and the skills mix necessary to do the work, and (c) recruiting, selecting, and promoting qualified candidates. *Retention* comprises the activities of (a) rewarding employees for performing their jobs effectively, (b) ensuring harmonious working relations between employees and managers, and (c) maintaining a safe, healthy work environment. *Development* is a function whose objective is to preserve and enhance employees' competence in their jobs by improving their knowledge, skills, abilities, and other characteristics; HR specialists use the term 'competencies' to refer to these items. *Adjustment* comprises activities intended to maintain compliance with the organization's HR policies (e.g., compensation, discipline), and business strategies (e.g., cost leadership). *Managing change* is an ongoing process whose objective is to enhance the ability of an organization to anticipate and respond to developments in its external and internal environments, and to enable employees at all levels to cope with the changes.

1. Responsibilities and Objectives

These activities can be carried out at the individual, work team, or larger organizational unit (e.g., department) level. Sometimes they are initiated by the organization (e.g., recruitment efforts or management development programs), and sometimes they are initiated by the individual or work team (e.g., voluntary retirement, safety improvements). It might appear that the activities of staffing, retention, development, adjustment, and managing change are the special responsibilities of the HR department. However, these responsibilities also lie within the core of every manager's job throughout any organization—and because line managers have authority (the organizationally granted right to influence the actions and behavior of the workers they manage), they have considerable impact on the ways workers actually behave. This implies two things: (a) a broad objective of HRM is to optimize the performance and prod-

activity of all workers in an organization, and (b) a special objective of the HR department is to help line managers manage those workers more effectively. The HR department accomplishes this special objective through policy initiation and formulation, advice, service, and control in close communication and coordination with line managers.

Generally speaking, HR specialists provide the technical expertise in each area, while line managers (or, in some cases, self-directed work teams) use this expertise in order to manage people effectively. In a small business, however, line managers are responsible for both the technical and managerial aspects of HRM.

For example, in the area of retention, line managers are responsible for treating employees fairly, resolving conflicts, promoting teamwork, and providing pay increases based on merit. To help line managers do these things effectively, the HR department must devise a compensation and benefits system that employees will perceive as attractive and fair, establish merit increase guidelines that will apply across departments, and provide training and consultation to line managers on all employee relations issues, such as conflict resolution and team building.

2. The Evolution of Human Resource Management

To appreciate where the field of HRM is going, it is important to appreciate where it has come from. As it is practiced in most developed countries around the world, modern HRM has emerged from nine inter-related sources.

(a) Rapid technological change that increased the specialization of labor associated with the industrial revolution.

(b) The emergence of free collective bargaining, with constraints established for unions and for employers.

(c) The scientific management movement.

(d) Early industrial psychology.

(e) Governmental personnel practices growing out of the establishment of the Civil Service Commission in the USA.

(f) The emergence of personnel specialists and the grouping of these specialists into personnel departments.

(g) The human relations movement.

(h) The behavioral sciences.

(i) Social legislation and court decisions.

2.1 The Industrial Revolution

Three characteristics of the industrial revolution were the development of machinery, the linking of human power to the machines, and the establishment of

factories in which a large number of people were employed. The result was a tremendous increase in job specialization as well as in the amount of goods that workers could produce. 'Division of labor' became a rallying cry of this revolution.

Clearly, the industrial revolution greatly accelerated the development of business and commerce. Owners and entrepreneurs generally did quite well for themselves, but the average citizen fared poorly in comparison to today's workers in terms of purchasing power and working conditions. Labor was considered a commodity to be bought and sold, and the prevailing political philosophy of *laissez-faire* resulted in little action by governments to protect the lot of workers (French 1994).

2.2 The Emergence of Free Collective Bargaining

Because of the way they were abused, it was inevitable that workers would organize to protect themselves and to improve their lot in life. From the perspective of workers, the industrial revolution fostered specialization, along with the need for workers within each specialization to organize themselves against its abuses. Trade unions, also called 'labor unions,' spread rapidly, and so did the incidence of strikes (see also *Labor Unions*). For example, in 1886 employees at the McCormick Reaper Works in Chicago went on strike for an eight-hour work day (Cohen 1960). In the United States, courts generally adopted an antiunion stance until 1935. In that year the National Labor Relations Act was passed. It focused largely on labor's right to organize. It became an unfair labor practice for an employer to coerce or to restrain employees in the exercise of their rights, to dominate or interfere with a labor organization, or to refuse to bargain collectively with a legal representative of the employees. Subsequent legislation refined, broadened, and set legal limits on the scope of management and union activities. From the perspective of labor, the development of free collective bargaining and the USA labor movement created the need for what is now recognized as effective human resource management.

2.3 Scientific Management

From the perspective of management, the scientific management movement also created the need for effective HRM. Frederick Winslow Taylor was the prophet of scientific management, and the stopwatch was his 'bible' (Bell 1972). Taylor began his experiments in the steel industry in 1885. His objective was the systematic analysis and breakdown of work into its smallest mechanical elements, and then their rearrangement into the most efficient combination. In addition to the scientific study of the task itself ('time and motion' study), Taylor argued that individuals selected to do the work should be as perfectly matched

physically and mentally to the demands of the job as possible, and that overqualified individuals should be excluded.

Employees should be trained carefully by supervisors (whose own work was also divided into specialties) to ensure that they perform the work exactly as specified by prior scientific analysis. In no case, however, should employees ever be called upon to work at a pace that would be detrimental to their health.

Finally, to provide an incentive for the employee to follow the detailed procedures specified (and closely supervised by line supervisors on the shop floor), Taylor felt that employees should receive an addition of between 30 and 100 percent of their ordinary wages when the task was done right and within the time limits specified. The first piece-rate incentive system had arrived.

Taylor also was interested in the social aspects of work, although he saw little good emerging from the social interaction within work groups. He felt that work groups fostered a level of individual efficiency equal to the level of the least productive worker in the group. In other words, he believed that the efficiency of the group would not be any greater than the efficiency of the least productive member. Thus Taylor tried to institute a kind of social physics into the workplace. While many of his ideas are controversial, there is no doubt that application of the principles of scientific management resulted in much higher productivity than would otherwise have been possible.

2.4 Early Industrial Psychology

In 1913, Hugo Munsterberg's book, *Psychology and Industrial Efficiency*, described experiments in selecting streetcar operators, ship's officers, and telephone switchboard operators. His contributions to human resource (or personnel) management are noteworthy for his emphasis on the analysis of jobs in terms of (a) the abilities required to do them, and (b) the development of testing devices—aptitude as well as work sample tests—that were able to forecast subsequent job performance (Munsterberg 1913).

Munsterberg's 1917 book, *Business Psychology*, included the results of questionnaires sent to hundreds of laborers. The questionnaires requested anonymous data on attitudes toward their work, their involvement in the labor movement, and their leisure activities.

Both Munsterberg and Taylor were concerned with how workers were selected for jobs, as well as with the individual worker's feelings and aspirations. Although they differed on several issues (notably, on how social structure affected productivity), Munsterberg's writings clearly suggest that he felt that the two emerging fields of personnel selection and social psychology (the study of how people affect and are affected by one another) were complementary (Landy 1992, Moskowitz 1977).

Other developments at about the same time included the use of rating sheets for interviewers, advances in checking references given by workers, and in statistical methods for estimating validity (the extent to which selection devices accurately forecast job performance). World War I accelerated the development of intelligence tests, such as the Army Alpha and the Army Beta, so that each individual could be matched more effectively with job requirements. Other kinds of psychological measures also appeared during and after World War I, such as measures of aptitude, interest, and personality.

2.5 The US Civil Service Commission

The Pendleton Act of 1883 established the US Civil Service Commission (today known as the US Office of Personnel Management). A forerunner in progressive personnel policies, the act provided that competitive examinations be administered as a basis for employment in the public service. It also provided a measure of employment security for those selected, prohibited discharge for refusing to engage in political activity, encouraged a nonpartisan approach to appointments, and mandated that a commissioner administer the act (Van Riper 1958). Perhaps the major impact of this act was to foster employment and promotion policies in the federal government based on merit. Over the years, these policies have influenced personnel practices in state and local governments as well as in private industry.

2.6 Private Industry's Approach to Personnel Management

Historians consider 1912 the approximate date of the emergence of the modern personnel department (Eilbert 1959). The term 'personnel,' with its connotation of managing people in organizations, began to appear about 1909. Between 1900 and 1920, while advances were being made in scientific management, industrial psychology, and the federal civil service, a number of personnel specialists emerged in companies such as B. F. Goodrich, National Cash Register, and Standard Oil of California. The specialists managed such areas as employment, employee welfare (financial, housing, medical, and educational), wage setting, safety, training, and health. This kind of specialization formed the basis for the organization of the modern human resources department.

2.7 The Human Relations Movement

In Chicago, in 1923, the Hawthorne Works of the Western Electric Company provided the setting for

one of the most famous behavioral research efforts of all time. The purpose of the research was to study the effects of illumination on worker productivity. In one experiment, production went up when the illumination increased. In another experiment, production went up when the illumination decreased. After three years of continually inconsistent results, the researchers concluded that in a situation involving people, it was impossible to change one condition (the lighting) without affecting other variables at the same time. Employee motivation and work group morale seemed to be the really crucial factors.

The results of subsequent experiments at the Hawthorne Works by Professors Elton Mayo, Fritz Roethlisberger, and T. North Whitehead of the Harvard Business School indicated that productivity was directly related to the degree of group teamwork and cooperation. The level of teamwork and cooperation, in turn, seemed to be related to the interest of the supervisor and the researchers in the work group, the lack of coercive approaches to productivity improvement, and the participation afforded the workers in changes affecting them (Pennock 1930, Roethlisberger and Dickson 1939). In short, the researchers came to view the organization of workers as a *social system*, in contrast to Taylor's view of that organization as a *technical-economic system*.

The conviction that group behavior and workers' feelings were associated with morale and productivity characterized much of the research and theorizing in the human relations movement for the next two decades. In practice, unfortunately, many managers seemed to use human relations for the short-term purpose of manipulating workers to increase output, rather than for the long-term goal of satisfying worker needs while meeting organizational needs. Failure to find evidence that these programs made a difference in workers' satisfaction or that happy workers were productive workers helped kill this approach to HRM (Brayfield and Crockett 1955).

2.8 The Behavioral Sciences

The behavioral science approach to managing people is an outgrowth of the human relations studies, although it embraces a wider base of academic and applied disciplines and is concerned with a wider range of problems. The term *behavioral sciences* refers to the social and biological sciences concerned with the study of human behavior. Much of the knowledge of HRM and many of its practical applications have come from such behavioral science disciplines as the following:

(a) *Industrial/organizational psychology*: the study of the behavior of people at work (see *Industrial–Organizational Psychology: Science and Practice*).

(b) *Social psychology*: the study of how people affect and are affected by one another.

(c) *Organization theory*: basic philosophies about why organizations exist, how they function, how they should be designed, and how they can be effective.

(d) *Organizational behavior*: the study of the causes of individual and group behavior, and of how this knowledge can be used to make people more productive and satisfied in organizational settings.

(e) *Sociology*: the study of society, social institutions, and social relationships.

We now know that the ways people behave in organizations cannot be explained simply by human relationships. The organization itself, through its unique 'culture,' molds, constrains, and modifies human performance. The way the organization is structured, the authority attached to different positions, and job and technology requirements clearly affect behavior. Even though the present understanding of the determinants and effects of behavior is incomplete, there is a better sense of the ways in which separate influences interact with each other to affect individuals.

Another area that has exerted a considerable influence on HRM theory and practice is *general systems theory*. Systems theory has developed from such fields as biology, physics, and cybernetics (the study of mechanical-electrical communications systems and devices). It has been applied and elaborated in every behavioral science field.

2.9 Social Legislation and Court Decisions

Governments in most developed countries around the world have passed social legislation and courts have issued decisions on unfair discrimination in employment. While some countries are more progressive on these issues than others, the laws and court decisions frame national policy on equal employment opportunity (EEO). In the USA, discrimination in employment is prohibited on the basis of race, religion, national origin, gender, age, and disability. Aggrieved individuals have the right to pursue their claims in court. At the level of the organization, HRM specialists work with top managers to develop and implement EEO policies to ensure that individuals are treated fairly in all aspects of the employment relationship.

3. Growth Stages of Human Resource Management

The foundation for modern HRM rests on the factors described earlier. Beyond that, HRM as practiced in the USA developed in four stages.

3.1 File Maintenance

The first stage may be called the 'file maintenance' stage, for it typified HRM activities through the mid-

1960s and the degree of emphasis on employee concerns. 'Personnel' was the responsibility of a special department. Its responsibilities included screening applicants, conducting orientation for new employees, collecting and storing personal data on each employee (e.g., date of birth, years of company service, education), planning the company picnic, and circulating memos 'whose impertinence was exceeded only by their irrelevance' (Meyer 1976).

3.2 Government Accountability

The second growth stage of HRM, the 'government accountability' stage, began soon after the Civil Rights Act of 1964 was passed. Laws governing antidiscrimination, pensions, and health and safety were enacted, federal agencies to enforce the laws were created, and court rulings affecting almost every aspect of the employment relationship were issued. All of these events accelerated the rise and importance of the HRM function. Class action lawsuits and the large financial settlements for the winning suits illustrated the costs of mismanaging people. Managers outside the HRM function began to take notice because top management let them know that ineptitude in this area would not be tolerated. Staying out of federal court became a top organizational priority. These trends also signaled the need for specialized competence in each aspect of HRM. Thus specialties began to appear in compensation and benefits, affirmative action (the promotion of minority concerns in all aspects of employment), labor relations, and training and development.

3.3 Organizational Accountability

In the late 1970s and 1980s, when many firms were struggling simply to survive, a combination of economic and political factors (high interest rates, growing international competition, shrinking US productivity growth) led to a demand for greater accountability in dollar terms of all the functional areas of business. HRM activities were no exception. Thus this third stage may be termed 'organizational accountability.' Although methods of assessing the costs and benefits of HR programs are available, they are not widely known (Cascio 2000). In addition, social trends, such as more women in the workforce, more minorities, immigrants, older workers, and poorly educated workers, accelerated the demands for improving the quality of work life, for managing cultural and ethnic diversity, and for continual training and retraining.

3.4 Strategic Partnership

In the 1990s and in the new millennium, HRM has evolved to a fourth growth stage as it assumes a central

role in the struggle of both large and small firms to gain and sustain a competitive advantage in the worldwide marketplace. This stage may be termed 'strategic partnership.' Top management looks to the HR department, as it does to line managers, to control costs, to attract and retain top talent, to enhance competitiveness, and to add value to the firm in everything it does.

4. Learning From the Past, Looking Toward the Future

As noted earlier, the field of human resource management is composed of five broad activities that together constitute the HRM system: staffing, retention, development, adjustment, and managing change. The field as it is known today has emerged from nine interrelated sources: the industrial revolution, the emergence of free collective bargaining, scientific management, early industrial psychology, governmental personnel practices growing out of the establishment of the Civil Service Commission in the USA, the emergence of personnel specialists and the grouping of these specialists into personnel departments, the human relations movement, the behavioral sciences, and social legislation and court decisions.

Since the 1950s, HRM has evolved through four growth stages: file maintenance, government accountability, organizational accountability, and strategic partnership. However, to be business partners who make genuine contributions to the success of an organization, HR professionals need to be proactive, not reactive. They must anticipate changes in the social, legal, political, and economic environments, and then facilitate effective organizational responses to them. Some of these issues include finding and keeping top talent in tight labor markets; capitalizing on the information revolution in ways that will benefit organizations, such as Internet-based recruitment and the development of interactive, company-based intranets to facilitate employee inquiries and inputs; managing the performance of geographically dispersed employees and teams (virtual teams); and providing training and development opportunities on demand. Indeed, this is the new paradigm of work: any time, anywhere, in real space or in cyberspace. HRM will contribute to the bottom-line success of organizations and to the betterment of employee welfare by developing proactive responses to the workplace challenges wrought by this new paradigm.

See also: Industrial–Organizational Psychology: Science and Practice; Job Analysis and Work Roles, Psychology of; Job Design and Evaluation: Organizational Aspects; Labor Unions; Performance Eval-

uation in Work Settings; Psychological Climate in the Work Setting; Strategic Human Resources Management; Well-being and Burnout in the Workplace, Psychology of

Bibliography

- Bell D 1972 Three technologies: Size, measurement, hierarchy. In: Davis L E, Taylor J C (eds.) *Design of Jobs*. Penguin, London
- Brayfield A H, Crockett W H 1955 Employee attitudes and employee performance. *Psychological Bulletin* **52**: 396–424
- Cascio W F 1998 *Managing Human Resources: Productivity, Quality of Work Life, Profits*, 5th edn. Irwin/McGraw-Hill, Burr Ridge, IL
- Cascio W F 2000 *Costing Human Resources: The Financial Impact of Behavior in Organizations*, 5th edn. Southwestern, Cincinnati, OH
- Cohen S 1960 *Labor in the United States*. Charles E. Merrill, Columbus, OH
- Eilbert H 1959 The development of personnel management in the United States. *Business History Review* **33**: 345–64
- French W L 1994 *Human Resources Management*, 3rd edn. Houghton Mifflin, Boston, MA
- Landy F J 1992 Hugo Munsterberg: Victim or visionary? *Journal of Applied Psychology* **77**: 787–802
- Meyer H E 1976 Personnel directors are the new corporate heroes. *Fortune* **93**: 84–8
- Moskowitz M J 1977 Hugo Munsterberg: A study in the history of applied psychology. *American Psychologist* **32**: 824–42
- Munsterberg H 1913 *Psychology and Industrial Efficiency*. Houghton Mifflin, Boston, MA
- Pennock G A 1930 Industrial research at Hawthorne and experimental investigation of rest periods, working conditions, and other influences. *Personnel Journal* **8**: 296–309
- Roethlisberger F J, Dickson W J 1939 *Management and the Worker*. Harvard University Press, Cambridge, MA
- Van Riper P P 1958 *History of the United States Civil Service*. Row-Peterson, Evanston, IL

W. F. Cascio

Human Rights, Anthropology of

Human rights constitute a field of increasing anthropological concern, not only for its substance but also for its inherent tensions. Since 1948, when the Universal Declaration of Human Rights (UDHR) was internationally accepted, anthropology has moved from an explicit opposition to the idea of a universal standard of human rights, over a more routinely skeptical attitude to a growing positive interest of both an empirical and a theoretical nature. This trend is related both to changing notions of anthropology and to the historical process of globalization. (See

Downing and Kushner 1988, and Messer 1993, and Hastrup 2001a, 2001b for overviews of the relationship between anthropology and human rights.)

1. Precursors

Although human rights, in the sense used here, were not on the international agenda before 1948, it was a latent idea for centuries, especially as far as civil rights were concerned. The American 'Bill of Rights' (1776) and the French declaration on '*Les droits de l'homme et de citoyens*' (1789) are often cited as ancestors of the UDHR (along with others), and from there the roots of human rights are traced back to classical Athens or early Christianity. In the nineteenth century, when a wave of humanism swept over Europe, one of the results was a new idea of an essential 'humanity,' deriving from seventeenth- and eighteenth-century treatises in political philosophy as well as from sixteenth-century renaissance humanism. Political philosophy in many respects was a forerunner of anthropology that was to take shape in the late nineteenth century. The idea of a universal humanity was a precondition of the nascent discipline, destined to give the proposition about a shared humanity empirical substance by way of qualifying cultural differences as variations upon the common theme of humanity.

In the early ages of anthropology the conjunction of the two concerns of equal rights and local difference is passionately testified to by Franz Boas, who, in his letter-diary from Baffin Island (1883–4), writes about his sincere wish to live in the USA:

I should much prefer to live in America in order to be able to further those ideas for which I live. ... What I want to live and die for, is equal rights for all, equal possibilities to learn and work for poor and rich alike! Don't you believe that to have done even the smallest bit for this, is more than all of science taken together? (quoted in Cole 1983, p. 37)

We know that later Boas was to fight fervently against racism, and on the whole most anthropologists see their discipline as genuinely respectful of all peoples. This collective sentiment makes no simple entry into the discussion of a universal standard of human rights, however, as this is conceptualized and articulated in those declarations, conventions, charters, and protocols that have been passed in the decades following the UDHR.

2. Anthropological Reactions to the Universal Declaration of Human Rights

Human rights in the modern sense were born in 1948 with the adoption of the UDHR. In the drafting

process, advice and comments were sought from various bodies, including the American Anthropological Association (AAA). A statement was produced by Melville Herskovits (1947) in the name of the AAA which carefully considers the idea of a universal standard and notes its local cultural implications, each of which is firmly anchored in contemporary anthropological knowledge. Noting that human rights appertain to the individual, the argument proceeds through the following successive conclusions:

(a) 'The individual realizes his personality through his culture, hence respect for individual differences entail a respect for cultural differences' (p. 541);

(b) 'Respect for differences between cultures is validated by the scientific fact that no technique of qualitatively evaluating cultures has been discovered' (p. 542);

(c) 'Standards and values are relative to the culture from which they derive so that any attempt to formulate postulates that grow out of the beliefs or moral codes of one culture must to that extent detract from the applicability of any Declaration of Human Rights to mankind as a whole' (p. 542).

This is clearly a stand against what was perceived as a Western bias in the UDHR, and it leads to the final conclusion: 'Only when a statement of the right of men to live in terms of their own traditions is incorporated into the proposed Declaration, then, can the next step of defining the rights and duties of human groups as regards each other be set upon the firm foundation of the present-day scientific knowledge of Man' (p. 543).

What transpires is a profound skepticism, grounded in the state of the anthropological art, including the general professional sentiment that culture is the sole legitimate source of moral values. Herskovits expresses it thus: 'Even where political systems exist that deny citizens the right of participation in their government, or seek to conquer weaker peoples, underlying cultural values may be called on to bring peoples of such states to a realization of the consequences of the acts of their governments, and thus to enforce a brake upon discrimination and conquest' (p. 543).

As pointed out by Richard Wilson (1997, p. 2), this is an attempt to play anthropology's trump card: its unique ability to elicit and assert any culture's underlying values. In addition it could be said that it also features the comparative knowledge inherent in anthropology, a knowledge both of cultural difference and of its translatability into valid, and more or less shared, human standards of good governance and respectable political action. In spite of the disclaimer, which was militantly underscored by Julian Steward (1948, p. 251), when he said 'we are prepared to take a stand against the values in our own culture which underlie ... such imperialism,' the UDHR was passed in 1948.

Together, the two attitudes exemplified by Boas and Herskovits, both of which are understandable and laudable, take us right to the core of anthropology's

comparative consciousness and its ambiguous implications. Since 1948, the discussion of the concept of human rights in anthropology has centered on the unsolved question of universalism or relativism (Dembour 1996). Another target has been the cultural translation of human rights concepts, allegedly deriving from a Western, and Christian, liberal democratic system, into 'like' cultural concepts within other moral and political systems. This is neither an easy task (as shown by e.g., Wilson 1996, and Asad 1997), nor is it an obvious way to go (Hastrup in press).

In actual practice the anthropological concern with human rights has taken various forms since 1948, and has been named accordingly as advocacy, activism, antiracism, and anti-imperialism. It is characteristic, though, that most of these trends have continued the relativistic attitude embedded in the AAA statement (and the implied skepticism toward human rights), while also advocating equal rights on the basis of a vision of a shared humanity. At the beginning of the twenty-first century, when human rights are appealed to by the international community under the banner of a 'culture of human rights' with increasing force, anthropologists have also taken a renewed interest in this culture.

3. The Culture of Human Rights

The intense talk of human rights since the early 1990s is not necessarily a symptom of a high moral profile worldwide, rather it is a symptom of the malaises of the times and of the will to confront them by way of new international laws. The choice of cure reflects a general process of legalization of culture in a post-modern era fraught with ontological insecurity, if we are to believe Anthony Giddens (1991). Where, earlier on, local cultures provided people with a sense of belonging and identity, the present globalization and fragmentation of identities have allegedly left people at a loss. They then look toward the law as the glue of community, whether local or global; the law is the new means of inclusion or exclusion *par excellence*.

Whether it is possible to single out causal links in this complex process or not, we do experience an increasingly strongly articulated 'culture of human rights.' We encounter this notion in most recent statements made by the UN high commissioner for human rights, Mary Robinson, as well as by academics of various kinds (e.g., Rorty 1993). From an anthropological perspective it is a peculiar culture in the sense that it is declared rather than lived, and that it is future-oriented rather than based in tradition. As such it parts company from those cultures that were the traditional object of anthropology, but so does, indeed, anthropology itself, becoming increasingly concerned with connections between the local and the global, and their mutual implications. This provides a

theoretically significant conjuncture between anthropology and the human rights discourse.

The culture of human rights is distinct also in its explicit acknowledgement of multiple sources. Whereas, traditionally, nations and other self-declared ethnic groups have referred to a particular genesis, an ultimate ancestor, or a culture-hero, the human rights culture draws upon all kinds of ancestry. As Mary Robinson (1998) puts it:

Today the Universal Declaration of Human Rights stands as ... one of the great documents in world history. The *travaux préparatoires* are there to remind us that the authors sought to reflect in their work the differing cultural traditions in the world. The result is a distillation of many of the values inherent in the world's major legal systems and religious beliefs including the Buddhist, Christian, Hindu, Islamic, and Jewish traditions.

Not surprisingly, after the demise of colonialism, and with the increasing critique of Eurocentrism and the hegemony of Western values, the search for roots has changed from the attempt at establishing a one-stranded genealogy to suggesting multiple sources. Mary Robinson's statement may be empirically doubtful, but it goes to show that there is a will to shift whatever Western legacy one might like to trace in the human rights instruments to a truly global commitment.

The declaration of a global culture of human rights is thus closely related to the historical process of globalization. In one sense, the notion of globalization simply states the obvious; due to increasingly rapid information and money exchange systems, we experience a time-space compression, that accounts for a new global consciousness. This global consciousness has important implications, one apparently being the appeal to a transcendent 'culture of human rights.'

To invoke a culture is also to invoke a shared language; as far as the human rights culture is concerned, the language is decidedly legal. As 'declarations' the international charters are exemplars both of language and of law. Whether hard or soft law, it is inevitably an expression of particular ideas about society, while not necessarily embracing its total reality. Far from being simply rules for social behavior, law is part of a distinctive manner of imagining the real (Geertz 1983, p. 173).

In the case of the international discourse on human rights, there is a concept of the modern nation-state as the responsible agent vis-à-vis the citizens, and a vision of an international community that transcends the more narrow interests of the states. Clearly, this community is no less 'imagined' than the nation (cf. Anderson 1983). It is imagined in terms of precisely those international charters that are its main expression. As a literally *declared* culture, its language comes to carry a much heavier burden of representation than any 'natural' language, being just one among other media for storing cultural knowledge.

Any culture in some ways rests on representation; for it to be conceivable as a whole, it must be demonstrated as such. In a 'legal culture,' such as the international culture of human rights, the representation becomes of a particularly reductive kind, because the experiences, by being represented as if on permanent trial in a courtroom, become subjected to an unavoidable process of reduction, or as Clifford Geertz has it about law in general, a 'skeletonization of fact, the reduction of it to the genre capacities of the law note' (Geertz 1983, p. 172). Legal thought and vocabularies, then, are constructive of social realities rather than merely reflective of them; this also applies to the legal vocabulary of human rights, and anthropologists naturally take an increasing interest in this particular social construction of a (global) community.

According to some anthropologists, globalization and human rights conjoin in a social transformation of unprecedented magnitude. Ellen Messer, for one, writes like this:

Since the close of World War II, the United Nations has been assembling declarations, legislation, and enforcement mechanisms to promote human rights. Both the ongoing efforts to establish a global community and to base membership on a universal, but evolving standard of values constitute perhaps the greatest social transformation of this century. (Messer 1996, p. 1)

This transformation is part of a larger trend that has often been seen as the end of modernity, implying an end also to ontological security. This could be one of the reasons why human rights have become the new standard of hope to many. The historical changes are matched by new scholarly efforts. We have witnessed important changes in the conception of the anthropological object, that can be summed up under the rubrics of a de-essentialization of the object (culture, society) on the one hand, and a humanization of the subject, seen now to be an active agent of history, on the other. This facilitates a truly anthropological investigation into issues of human rights.

Intellectually, one of the greatest advances of the twentieth century was a willingness to learn from history and anthropology that humans are extremely malleable (e.g., Rorty 1993, p. 115). Thanks to comparative knowledge of humanity in time and space we have realized the flexibility and the self-shaping character of humans, and given up on the idea of a naturally rational animal.

The historical and intellectual climate is thus very much in favor of achieving a 'human rights culture.' While the dominant trends in present day anthropology—the stress upon globalization, on process, and on agency among other things—make a new approach to universal human rights possible, there is no theoretical reason why we should not at the same time maintain a distinct, more or less bounded object

of analysis: the local worlds of people. This has fueled a new interest in what are known as cultural rights.

4. Cultural Rights

In view of the traditional anthropological ambivalence toward human rights, generally based in view of culture as an inviolable and inherently valuable entity, the notion of cultural rights might ease the tension. This is not self-evident, however.

In the early instruments, such as the UDHR and the 1996 International Covenant on Social, Economic, and Cultural Rights (ICESCR), 'cultural rights' are individual rights to 'enjoy culture,' in compliance with the general emphasis on individual rights and the obligation of the state to facilitate them. In addition to this focus on the individual's cultural rights, a more challenging (politically and theoretically) perspective on cultural rights is opening up with increasing force, however. This is a notion of collective cultural rights, in relation to which the individual enjoyment of culture gives way to, or is at least supplemented by, a right of a particular group to protect and develop their culture. It was notable already in the *International Covenant on Civil and Political Rights* (ICCPR 1966, Article 27): 'In those States in which ethnic, religious, or linguistic minorities exist, persons belonging to such minorities shall not be denied the right, in community with the other members of their group, to enjoy their own culture, to profess and practice their own religion, or to use their own language.' It is surprising, perhaps, that the protection of minorities is placed within the civil and political rights, and not the social, economic, and cultural rights; if nothing else, this confirms the lack of clarity with which the concept of culture is hampered in the international instruments.

In spite of the weak beginnings and the unclear focus within the field of cultural rights, there is an increasing interest in them at the beginning of the twenty-first century. It is related to the historical fact that most conflicts today are so-called ethnic conflicts, at closer inspection often rooted in economic discrimination or scarce resources. Ethnic difference often covers other inequalities, at least when they are expressed in conflict. The ethnic conflicts take place within states, that are the traditional duty-holders in the human rights system, and they raise endless new questions of how to define those peoples, minorities, or indigenous populations that are the new beneficiaries of collective rights. In spite of the definitional problems, the UN system proceeds toward an increasing proliferation of instruments dealing with collective rights.

One such instrument is the UN *Declaration on the Rights of Persons Belonging to National or Ethnic, Religious or Linguistic Minorities* (DM) of 1992. As a Declaration it is not legally binding in the strict sense of the term, but state signatories nevertheless commit

themselves to the spirit of the Declaration. In many ways the DM can be seen as an extension of Article 27 of the ICCPR quoted above:

States shall protect the existence and the national or ethnic, cultural, religious and linguistic identity of minorities within their respective territories and shall encourage conditions for the promotion of that identity. (DM, Article 1)

Persons belonging to national or ethnic, religious and linguistic minorities ... have the right to enjoy their own culture, to profess and practice their own religion, and to use their own language, in private and in public, freely and without interference or any form of discrimination. (DM, Article 2)

From an anthropological perspective, what seems problematic is the recurrent emphasis on people's own culture as a natural property. The embedded concept of culture is an (outdated) essentialist one that goes counter to more recent anthropological notions of culture as a more or less temporary framework of understanding.

In the 1993 *Draft Declaration on Indigenous Peoples Rights* (IPR) the notion of culture as essence is further cemented; to give just one example:

Indigenous peoples have the right to practice and revitalize their cultural traditions and customs. This includes the right to maintain, protect, and develop the past, present, and future manifestations of their cultures, such as archeological and historical sites, artifacts, designs, ceremonies, technologies, and visual and performing arts and literature, as well as their right to the restitution of cultural, intellectual, religious, and spiritual property taken without their free and informed consent or in violation of their laws, traditions, and customs. (IPR, Article 12)

While one cannot deny that the indigenous peoples have suffered the more sinister effects of colonialism, it is unlikely that such collective rights that are propounded in the IPR will undo history. It is still only a draft declaration but seen in conjunction with the previously cited instrument there is evidence of a gradual transformation of cultural rights from individual to collective rights, that may potentially threaten the very foundation of human rights: their claim to universality. It seems, then, that at a time when anthropologists have left behind the notion of culture that prompted the AAA to take a stand against the idea of universal human rights, the very same notion hits back from within the system of human rights itself, potentially undermining it.

5. New Anthropological Avenues

The brief discussion of cultural rights shows that there is a need to reconsider the issue of human rights in the light of recent theoretical developments. If the notion of culture as essence is obsolete and cultural relativism generally self-refuting, human rights can no longer be

invalidated as a matter of course. The comparative perspective of anthropology itself is based on an idea of a shared humanity and a cross-cultural comprehensibility that tallies with the idea of equality and dignity in all humans. One of the things that anthropology has helped realize is the imaginability of 'others' as somehow like 'us' in spite of diversity.

Ethnographies of violence and suffering are still needed to add a dimension of experience to the (legal) language of rights, and thus contribute to a shared understanding of those sufferings that a universally accepted standard of human rights may help alleviate. Not only ethnography, but also our regained confidence in anthropological theory will contribute to the achievement of human rights, because anthropology's theoretical language can embrace cultural differences within a single (scholarly) discourse. Theory expresses what is not otherwise said; as such it is a social practice and contributes to the making of the world—not unlike the representation of society in legal terms, as discussed above. This imaginative contribution makes room for the renewed rapprochement between anthropology and human rights that we are witnessing.

See also: Boas, Franz (1858–1942); Cultural Rights and Culture Defense: Cultural Concerns; Human Rights, History of; Human Rights in Intercultural Discourse: Cultural Concerns; Human Rights: Political Aspects; Offenses against the Laws of Humanity: International Action

Bibliography

- Anderson B R O'G 1983 *Imagined Communities. Reflections on the Origin and Spread of Nationalism*. Verso, London
- Asad T 1997 On torture, or cruel, inhuman and degrading treatment. In: Wilson R A (ed.) *Human Rights, Culture and Context. Anthropological Perspectives*. Pluto Press, London, pp. 111–33
- Cole D 1983 'The value of a person lies in his *Herzensbildung*': Franz Boas' Baffin Island letter-diary, 1883–1884. In: Stocking G W Jr (ed.) *Observers Observed: Essays on Ethnographic Fieldwork*. University of Wisconsin Press, Madison, WI, pp. 13–52
- Dembour M-B 1996 Human rights talk and anthropological ambivalence. The particular contexts of universal claims. In: Harris O (ed.) *Inside and Outside the Law*. Routledge, London, pp. 19–40
- Downing T, Kushner G (eds.) 1988 *Human Rights and Anthropology* (Cultural Survival Report 24). Cultural Survival, Cambridge, MA
- Geertz C 1983 Local knowledge: Fact and law in comparative perspective. In: *Local Knowledge*. Basic Books, New York, pp. 167–234
- Giddens A 1991 *Identity and Modernity*. Polity Press, Cambridge, UK
- Hastrup K (ed.) 2001a *Human Rights on Common Grounds. The Quest for Universality*. Kluwer Law International, The Hague
- Hastrup K (ed.) 2001b *Legal Cultures and Human Rights. The Challenge of Diversity*. Kluwer Law International, The Hague

- Hastrup K in press Anthropology's comparative consciousness: The case of human rights. In: Fox R, Gingrich A (eds.) *Comparison in Anthropology*. Routledge, London
- Herskovits M 1947 AAA statement on human rights. *American Anthropologist* **49**: 539–43
- Messer E 1993 Anthropology and human rights. *Annual Review of Anthropology* **22**: 221–49
- Messer E 1996 Anthropology, human rights, and social transformation. In: Moran E F (ed.) *Transforming Societies, Transforming Anthropology*. University of Michigan Press, Ann Arbor, MI
- Robinson M 1998 *The Universal Declaration of Human Rights: A Living Document*. United Nations Press Release 27 January 1998 (UNHCHR)
- Rorty R 1993 Human rights, rationality, and sentimentality. In: Shute S, Hurley S (eds.) *On Human Rights. The Oxford Amnesty Lectures 1993*. Basic Books, New York, pp. 111–34
- Steward J 1948 Comments on the Statement of Human Rights. *American Anthropologist* **50**: 351–2
- Taylor C 1985a *Human Agency and Language. Philosophical Papers I*. Cambridge University Press, Cambridge, UK
- Wilson R A 1996 The Sizwe will not go away. The Truth and Reconciliation Commission, human rights and nation-building in South Africa. *African Studies* **55**(2): 1–20
- Wilson R A 1997 Human rights, culture and context: An introduction. In: Wilson R A (ed.) *Human Rights, Culture and Context. Anthropological Perspectives*. Pluto Press, London, pp. 1–27

K. Hastrup

Human Rights, History of

The expression 'History of Human Rights' has two different meanings. It can be understood (as it usually is) as the history of the idea and the concept of human rights, or (less frequently) as the historical occurrence and study of human rights and their abuses. Therefore, the basic debates about the idea and concept of human rights are first clarified here from a historical perspective. Then, a range of methodological and ethical problems in the historical study of human rights are presented.

1. History of the Idea and Concept of Human Rights

1.1 Moral vs. Positive Rights

In 1689–90, just after the Glorious Revolution of 1688, the English philosopher John Locke (1632–1704) wrote his *Two Treatises on Government*, which contained the first classical formulation of human rights (then still called 'natural rights'). In retrospect, this work stood at the juncture of a set of amazingly diverse ideas, some of which went back to antiquity and some of which will survive the twenty-first century. Locke's point of departure was the doctrine of natural law as it had developed since antiquity and had been

transformed, in certain respects almost beyond recognition, by medieval Christianity and early-modern philosophers. According to Locke, the doctrine maintained that nature endowed all humans with certain basic moral entitlements (such as the right to life, liberty, and property), to which the man-made laws of society, both positive and customary, had to conform.

These natural rights were thought to possess five distinct characteristics: they were discoverable by reason, itself a faculty of human nature; they existed before humans entered society; they applied to all humans by virtue of their being human; they were inalienable; and they were restricted only by the recognition of the rights of others. As such, these rights limited the power of the society—generally in its predominant historical-political form, the state—and its laws, and respect for them was the first test for the legitimacy of political power. Ancient ideas on rationalism, equality, and the restrained exercise of power formed the substratum for this natural law doctrine, which proved to be very influential. Despite its claim of eternity, however, many of the doctrinal components, such as the exact catalog of moral principles or the source of authority for these principles, turned out to be rather flexible.

Although temporarily in eclipse because of the ascendancy of the sixteenth-century absolutist state, the natural law doctrine would be fostered by three developments initiated in that very period. With the Renaissance came an emphasis on individualism and secularism, the outcome of the religious struggles following the Reformation taught the necessity of tolerance, and the seventeenth-century scientific revolution marked the triumph of rationalism. These developments reversed the philosophical priorities of the hierarchical medieval societies: natural law gradually became associated with rights rather than with duties and with individuals rather than with groups.

The perception of the source of these natural rights also changed. Whereas Christianity had maintained that God was the ultimate source and political absolutism had invested the monarch with divine rights, the consent of the governed came to be seen as the final authority. Individuals gave this consent in a social contract, which presupposed that they possessed rights before entering society. Governments owed their origin to that contract, by which free and equal citizens agreed to entrust to a neutral sovereign body some of their rights, to accept its authority and rules as long as it protected them, and to rebel against it if it did not. This formed the basis for the rule of law and the separation of, and balance between, the legislative, executive, and judicial functions of government.

Although previous theories (such as Hugo Grotius') paved the way and later ones expressed similar ideas, it was Locke's political theory which was to prosper, especially outside England. In the age of Enlightenment it exerted profound influence upon the politicians behind the great revolutions at the end of the

eighteenth century. In the United States and France, natural rights, mainly of a political and civic character, were incorporated into the declarations and constitutions of the revolutionaries. In 1789 natural rights transformed into the 'rights of man' (*droits de l'homme*). This marked the beginning of a process of codification of rights within the constitutions of most states and in international law that still has not ended.

From this brief historical overview, it is clear that, although many aspects of the human rights conception have an ancient pedigree, other crucial ones have not. The concept was the result of a unique historical combination of a series of events and philosophical thought reflecting upon them, and developed relatively late in Western history. The natural law doctrine itself knew a curious fate: with unprovable tenets and with the possibility of abusive claims 'in the name of nature,' it became overshadowed by the powerful state-centered thinking of the nineteenth century (as it had been in the sixteenth), but it survived. It was rehabilitated in a modest form after the two world wars. The twentieth-century approach did not derive human rights from nature any longer, but from (historically determined) rational ideas on human dignity.

1.2 Absolute vs. Relative Rights

The debate on moral vs. positive rights was continued by liberal, conservative, and socialist critics of human rights. Typical of this criticism were the works of Jeremy Bentham (1748–1832), Edmund Burke (1729–97) and Karl Marx (1818–83). These critics argued against the abstract and absolute character of human rights. They maintained that rights had to be related to the society in which they were to be exercised. The liberal utilitarian Bentham believed only in the force of positive legislation. For him, natural rights were imaginary, 'nonsense upon stilts.' He feared that they were powerful rhetoric in the hands of rulers and a substitute for effective legislation.

Unlike Bentham, the conservative Burke did not deny the existence of natural rights. However, he feared them for reasons contrary to Bentham's: in the hands of the common people, natural rights would stimulate revolutionary sentiments and cause social upheaval leading to terror and to the destruction of the social fabric. Rights (and the duties linked with them) ought to be derived from the principles and values emanating from the particular traditions of a society.

For the socialist Marx the human rights doctrine was not radical enough. It left the inequalities generated by capitalism untouched (particularly private property) and therefore served the interests of the ruling bourgeois class. Because of these inequalities, the workers were unable to enjoy their rights fully, which therefore remained largely formal. As a class,

the workers had a duty to create the conditions for social change by claiming collective social and economic rights with which to overturn the economic system. The new, egalitarian society would be one without a state, laws, or rights.

These thinkers did not stand alone in their criticism, but their arguments gained a classic character and inspired powerful currents. Bentham's ideas gave impetus to the tradition of legal positivism. Ideas similar to Burke's would feed the romanticism and historicism of the nineteenth century to the extent that they too emphasized cultural identity and diversity. Marx's concerns heightened awareness of the need for social legislation and influenced the 1917 Russian Revolution.

1.3 Men vs. All Humans and Individuals vs. Groups as Rights-holders

Clearly, until long after the era of revolutions, the concept of 'rights of man' covered only very restricted categories of humans, usually male tax-paying citizens. The large majority of women, slaves, and foreigners were excluded. The struggle to include these categories in the ranks of rights-holders would lead to the final transformation of the concept into 'human rights.' As early as 1791, Olympe de Gouges struggled for women's rights. With the movement against slavery, another group followed. Categories of individuals deemed particularly vulnerable (women, children, aliens, refugees, prisoners, the disabled) were eventually protected by special regulations.

Likewise, the difficult problem of how to realize group rights was introduced at an early stage in the human rights debate. Groups deemed vulnerable (minorities, colonized peoples, indigenous peoples) increasingly received special attention. Specific treaties for national, ethnic, or religious minorities were drafted long before general treaties came into effect. The idea of self-determination appealed to peoples under colonial rule or comparable alien subjugation and proved to be a crucial concept. It was a dangerous concept too, because the narrowing of self-determination to only one of its possible interpretations—secession—was potentially explosive and because consensus was lacking as to whom the rights-bearer ('a people') exactly was. Self-determination was promoted to a group right only in the 1966 covenants (see below).

From an acute attention to making the concept 'human' all-embracing, two further questions on the extension of rights can be derived. First, do past and future generations have human rights? As to past generations, one could think of the right to a decent burial or the right to be treated with respect in historical works. As to future generations, the preservation of the cultural heritage of humanity and of the

natural environment are chief concerns as is the relationship between human rights and the human genome. As to future generations in relation to past generations, the accountability of successor governments for human rights abuses committed by their precursors and the obligation to investigate them are matters of legitimate debate. Second, what consequences does the anthropocentric character of human rights have, as it excludes other sentient beings and is often based on restricted and exploitative conceptions of nature?

1.4 Domestic vs. International Jurisdiction Over Human Rights

Since 1800, two tendencies, one international, the other national, partly opposite, partly complementary, have dominated the debate over who was to control the protection of human rights. A process of internationalization started when human rights were incorporated into the constitutions of growing numbers of countries. At the same time, human rights became part of the agenda of intergovernmental meetings, leading to the humanitarian interventions of the nineteenth century (continuing a tradition dating from the seventeenth century) and the gradual incorporation of human rights into international law. That human rights abuses often created conflicts that threatened international peace was a lesson learned slowly. Treaties were drafted to protect selected vulnerable categories such as slaves, war prisoners, and minorities. The League of Nations, founded in 1919, did not formally recognize human rights but added pioneering treaties for minorities, refugees, and mandate territories.

The decisive historical moment for internationalization, however, came during World War II, when the link between human rights violations and war (hence the link between human rights protection and international peace and security) became clearly visible and, shortly after, when the Nazi and Japanese atrocities led to worldwide condemnation embodied in the international war crimes tribunals of Nuremberg and Tokyo. Although the United Nations only paid secondary attention to human rights in their 1945 Charter, this did not prevent work on an International Bill of Human Rights, consisting of a morally compelling but not legally binding Universal Declaration of Human Rights, adopted in 1948, and two International Covenants, one on economic, social, and cultural rights, and one on civil and political rights, both approved in 1966 and entered into force in 1976.

In the efforts to transform human rights into positive rights by making them enforceable in positive law, usually three stages are distinguished: standard-setting (codification of standards), ratification (making standards binding), and implementation

(enforcing standards via sanctions). Seminal steps for enforcing standards were the elaboration of complaint mechanisms for state and nonstate actors (mainly since the 1970s), the appointment of a High Commissioner for Human Rights (1993), and the adoption of the Statute of an International Criminal Court with universal jurisdiction to prosecute alleged perpetrators (including heads of state and government) of genocide, crimes against humanity, war crimes committed in international and internal conflicts, and crimes of aggression (1998). Important preparatory work for these new devices was done by a coalition of non-governmental organizations, the representatives of civil society. The elaboration of human rights institutions and mechanisms on a world scale was paralleled, and often preceded, by similar developments on a regional scale in Europe, America, Africa, and the Arab world.

In the midst of this evolution, the state was the key player. The cherished principles of the Westphalian world order, established in 1648, were national sovereignty and interstate equality. Their corollary, domestic jurisdiction, excluded interference from abroad in the internal affairs of the state. The vigorous doctrine of nationalism and successive waves of decolonization since the late eighteenth century strengthened these principles. Paradoxically, recognition of national sovereignty made possible progress in international law with its core principle of interference in the internal affairs of states. In the first place, states slowly realized that when they adopted human rights as the basis of the rule of law, they implicitly created a duty to protect them everywhere, including beyond their borders. In the second place, it was painfully obvious that the states themselves were the main perpetrators of human rights abuses. Therefore, states reluctantly but voluntarily began to accept an international monitoring element that restricted their sovereign domestic power. The number of humanitarian interventions increased drastically in the 1990s, but their enabling conditions, forms (diplomatic, political, economic, or military), potential abuses, and effects were hotly disputed.

1.5 First- vs. Second- vs. Third-generation Human Rights

As the success and the criticism of the human rights doctrine expanded the list of human rights, analysts grouped them into generations. The first generation, tied to the idea of liberty, were the traditional or individual rights of a political and civic nature. They were characterized as negative rights ('freedoms from') because they favored governmental abstention. The second generation, tied to the idea of equality, were the modern or collective rights of a social, economic, and cultural nature. They were characterized as

positive rights ('rights to') because they favored governmental intervention. The third generation, tied to the idea of fraternity, were modern or collective rights of a global nature. They were characterized as solidarity rights and usually included the rights to development, human security, peace, and self-determination. The rights of indigenous peoples and the right to a life-sustaining environment are frequently added, but these are also listed as fourth-generation rights by some.

Each generation emerged in a particular historical context. The first generation was the result of the late-eighteenth-century revolutions led by the bourgeoisie that abolished feudalism (and colonialism in the case of the United States). Increasing political emancipation and democratic participation in the next two centuries gave further impetus. The second generation developed in the nineteenth century as the socialist tradition with its reformist and revolutionary strategies, and introduced ideas of welfare and social legislation to counter the exploitation of working classes and colonial peoples and to create the necessary conditions for the full exercise of rights. The third generation is usually associated with the Third World demand for a worldwide redistribution of power and wealth; another factor is the tension generated by the staggering increase of interdependency and global problems on the one hand and the need to safeguard cultural identities and diversity on the other. The three generations of rights can be placed on a continuum with justice at the one end and solidarity at the other.

This bewildering diversity of human rights occasioned controversies on their compatibility, priority, and indivisibility. Because all human rights were conceived as inalienable, they sometimes came into mutual conflict. A classic example of a conflict within the same generation of rights is the tension between expression and religion: what is called 'an opinion' by one person, might be labeled 'blasphemy' by the next. But most conflicts occurred between, rather than within, generations. Whereas first-generation rights assumed the character of universally enforceable standards, second- and third-generation rights were rather aspirations whose fulfilment was utterly dependent on the creation of specific conditions. Faced with all these demands, some states, mainly in the First World, gave priority to first-generation rights, while others, mainly in the Second and Third World, struggled, if not for the primacy, then at least for the equivalence of second- and third-generation rights. Acrimonious debates were held on whether political freedom or economic development was to be accorded priority. Those defending the priority of political freedom ('free minds') were accused of perpetuating inequality, those defending the priority of economic development ('full bellies') of perpetuating dictatorial government. In the Cold War era (and beyond), these arguments were used as political propaganda. The hypothesis that political freedom or democracy should

be suspended until certain levels of economic development are attained is contradicted by the bitter experience that in such cases freedom is usually trampled underfoot, while no economic development is attained after all. Unaccountable governments generally exacerbate economic and environmental crises. Moreover, how can a present suspension of freedom be conceived as preparing for its future protection? The reverse hypothesis seems to have firmer ground: with their political rights respected, citizens will participate more fully in the economic development of their society. The correlation between economic development and political freedom seems to be accepted, but under which conditions, for which precise political and economic areas, and in which causal direction the correlation is valid remains largely unresolved.

At the 1993 World Conference on Human Rights in Vienna, the second such conference ever and the first after the Cold War, the indivisibility of human rights was strongly reaffirmed. Even so, seven rights acquired a special status as they were the only ones stipulated by the 1966 Covenant to be nonderogable, even in times of public emergency. They are the rights to life, to be recognized as a person before the law, and to freedom of thought, conscience, and religion; and the prohibitions against torture, slavery, imprisonment for non-fulfilment of contractual obligations, and retroactive application of criminal laws.

1.6 Universality vs. Relativity of Human Rights

Seen from a historical perspective, most major religions and ethical systems have been concerned with human dignity. Opinions differ, however, on whether they were also concerned with human rights (viewed as one specific way to realize human dignity). Developing the human rights concept took two millennia in Western history and was so full of unforeseen links, ruptures, reversals, and coincidences, that it may convincingly be called a unique product of that history (see above). The quasi-universal emphasis on the community in which individuals had mainly duties switched in Europe and its American colonies to an emphasis on individuals and rights. Elements of the human rights concept, such as some natural law principles, were available outside the West, but the combination of factors that led to its formulation was not. As the West conquered the rest of the world, it introduced its dominance and violence. At the same time, its notions of freedom, nationalism, and cultural diversity, developed initially for home consumption, fell on fertile ground overseas. These notions, including human rights, proved able to transcend their particular roots and context. In the hands of colonized peoples, they echoed similar indigenous values and served to justly criticize Western exploitation in the language of the West itself. Thus, whereas human

dignity is a universal value with venerable traditions everywhere, the claim of the universality of human rights has an anthropological basis, not a historical one, and is rooted in the universal appeal of a specific system of thought.

The anthropological debate was no less vivid than the historical debate. During the preparatory work on the 1948 Universal Declaration, leading American anthropologists emphasized the cultural relativity of values and denounced the universalist pretensions of the intended declaration as cultural imperialism. Moreover, the historical critics of the universality of human rights (presented above) were succeeded by neo-Marxists, communitarians, and postmodernists. But the most influential group of critics were the vociferous ruling elites of some non-Western countries, especially in Asia and Africa. They challenged specific formulations of private rights such as religion or marriage, and advocated, instead of universality, the principle of domestic jurisdiction or the primacy of second- and third-generation rights over first-generation rights. Remarkably, the surviving victims of imprisonment, torture, and other abuses in non-Western countries (among them many future elites of these countries) found such criticism unconvincing. These victims argued that denying universal human rights to citizens outside the West was itself a sign of cultural arrogance. A new consensus seemed to emerge when 171 states unanimously reaffirmed the universality claim at the 1993 Vienna Conference and thereby made the claim far more representative than at the time of the 1948 Universal Declaration which was approved by a vote of 48 to zero with eight abstentions. At the same time, care for second- and third-generation rights is growing. In 1998 a draft of a Universal Declaration of Human Responsibilities was submitted to the United Nations General Assembly, as an attempt to achieve a better balance between freedom and responsibility and between justice and solidarity.

2. History and Historiography of Human Rights and Their Abuses

The second interpretation of the expression 'History of Human Rights' points to historical research on human rights. When looking at this area, we are confronted with a paradox. On the one hand, as human rights encompass so many spheres of life, large parts of the historiographical production inevitably deal with concrete aspects of human rights and their abuses (e.g. the history of education, the history of slavery, the history of censorship). Equally, historians endeavored to write histories of emancipation movements or of freedom. Grand theories of history and world histories sometimes had implicit views on human rights, and the more recent they were, the more they covered the whole of humanity in an equitable manner. Indeed,

with little exaggeration one could say that almost all history touches on some present human rights concern. On the other hand, it is hard to find research explicitly designed to study a particular right in historical perspective. Some of the methodological and ethical problems involved in this type of research are outlined below.

2.1 Methodological Problems

In the following example, a general question is analyzed to demonstrate the methodological complexities of human rights research. That question is: 'Do human rights abuses increase or decrease over time?' A historical question because of its time dimension, it is also an ethical question to the extent that the answer gives clues about the moral progress of humanity (philosophers of all times have been in dire need of such clues). Questions as broad as this one should be rendered operational. This may be done by narrowing the 'human rights' of the question to one of its categories, 'civil and political rights,' by selecting from this category one important right, 'the right to life,' and, finally, by selecting from this right its most serious abuse, 'genocide.' Reformulated in operational terms, the question thus becomes: 'Do genocides increase or decrease over time?' This question can be approached logically and historically.

A priori logical considerations distinguish between absolute and relative levels of abuse. In absolute terms, the case of an overall increase of genocides is defensible. The population growth on earth, the increasing sophistication of repressive bureaucracies and technologies, and the anonymity often physically and morally shielding the perpetrator are all contributing factors in making this statistically probable. Our perception, however, may be deficient in two contradictory directions: we may overestimate the present absolute level of cruelty because of our lack of knowledge of past cases; alternatively, we may underestimate the actual level because we are too impressed by the countervailing forces, especially since the recent expanding human rights awareness has led to a higher state of general alert and perhaps to more deterrence. In relative terms, the question is very different: given that an equal amount of cruelty probably leads to greater destruction in modern times, were societies more repressive or more human towards their citizens in the past than they are today? In tackling this question, one ought to apply criteria for measuring the degrees of cruelty, but which ones? The number of victims in relation to the total population? The presence of especially vulnerable categories of victims? The occurrence of particularly cruel types of repression? In addition to this difficulty, there is the danger of anachronistically projecting back contemporary conceptions and values into the past. In the

face of the pain suffered by each individual victim, however, such considerations on the moral economy of death may sound superficial or scornful. A probable absolute increase of cruelty or repression in history, then, does not imply that humanity is now more barbarous than in the past, nor that it is less.

Does a historical approach shed more light on the problem? The 1948 United Nations Genocide Convention called genocide 'an odious scourge' that 'at all periods of history (...) has inflicted great losses on humanity.' Thus the term 'genocide'—coined in 1944 by Raphael Lemkin—seems to cover phenomena in all historical periods. Which phenomena? The Convention provides a definition (a series of 'acts committed with intent to destroy, in whole or in part, a national, ethnical, racial, or religious group') that genocide scholars reluctantly accept or decline all together. These scholars argue that many more group types than the ones from the definition have been genocide victims (notably political and social groups) and debate the boundaries between genocides and comparable crimes. They also report that many researchers suffer from psychological barriers in studying the almost unbelievable cruelty of genocide. The body of genocide research only began to grow after 1945, and with the exception of the relatively well-studied Holocaust, it is still lacking in many instances.

Not only do definitions and psychological resistance raise problems, so do heuristics and epistemology. It is often excruciatingly difficult to know how many died and what exactly happened during each specific genocide. Drawing statistics on the number of victims that are more than 'guesstimates' is a skill in its own right. Rudolf Rummel's elaborate statistics (1994) show that the democide rate (a concept including genocide, politicide, and mass murder but excluding war dead) was almost certainly higher in the twentieth century than in any previous century: he presents a calculated total of 169,202,000 dead for the period 1900–87 (including 38,566,000 genocide victims) vs. a hypothetical total of 625,716,000 dead for the period from the thirtieth century BCE to 1900. However, these figures do not allow conclusions about a possible increase of democide, and by implication of genocide, because Rummel derived his hypothetical total of pre-twentieth-century democide from the twentieth-century democide rate and the population for each century since the thirtieth century BCE. Linked to the statistical problem is the source problem: few sources normally survive, at least for the earlier genocides, in history. The sources were either not kept, hidden, lost, or destroyed. When they are available, they come from either perpetrators, victims, or bystanders. When there are sources from two or all parties, they frequently conflict with each other. Finally, their authenticity and reliability have to be tested. In short, the usual techniques of historical criticism have to be applied with unusual sharpness. Even when sources allow the historian to distill a picture of the killings, proof of the

intention of the perpetrator to destroy a group as such (an essential component of all genocide definitions) is difficult. Openly stated motives are exceptional and notoriously unreliable and, therefore, intent generally has to be ascribed. This is precarious, especially where the genocide is unanticipated.

Once we have reliable descriptions of the events, the question remains whether they catch the essence of the horror. The limits of representing the unrepresentable have been much debated in recent decades. The analysis and interpretation of the data pose their own problems. First, genocide is a complex interdisciplinary phenomenon with historical, psychological, sociological, political, legal, philosophical, and religious aspects. Second, the systemic and human factors conditioning genocides and the extent to which they are unique and comparable are open to much controversy. Third, it is important to know whether the nature of genocide has changed over time. Among the dozen or so typologies of genocide, some address this problem, however inconclusively. Frank Chalk and Kurt Jonassohn, for example, distinguish between retributive genocide (to eliminate a real or potential threat), despotic genocide (to spread terror among real or potential enemies), developmental genocide (to acquire economic wealth), and ideological genocide (to implement an ideology). They then classify 22 cases from history under the most relevant type. The distribution of cases reveals that the fourth type is a predominantly twentieth-century phenomenon with few historical antecedents. Even if this suggests that the nature of genocide has (partially) changed, the question remains why. With all its methodological difficulties summarily depicted above, the historical approach is obviously far from conclusively answering the initial question either.

2.2 *Ethical Problems*

The historical study of human rights and their abuses also involves ethical problems. Some of these are related to the researching historians, others to the human beings they study. Most are not unique for this type of research, but they present themselves cogently here. Human rights research can be dangerous, especially when the historians are working on contemporary material in their own society or on cases of the past which suggest easy analogies with the present or endanger the legitimacy of present rulers. This may lead to the persecution of the historians and entail reflection on their professional ethics: what are the rights and obligations of historians working under such restrictive conditions and what should their colleagues working under freer conditions do?

What rights do the human beings under scrutiny have, apart from the right to be treated with respect in historical works (see above)? With regard to the perpetrators of crimes, the question is whether moral

judgments of their acts are possible, and if possible, permissible, and if permissible, obligatory. Opinions of historians widely differ on these points. Few believe that they are judges for eternity, but nevertheless most will admit that there is something like a right to historical truth, for which it is never too late. They would probably subscribe to the professional obligation not to deny or forget the fate of the dead and surviving victims. This justified claim, however, can be at odds with impartial research. It is a task with many risks. Without the passion of the survivors, historians may 'normalize' the cruel abuses of the past by inserting them into the stream of history. They may omit crucial findings for fear of breathing new divisive fever into the collective memory. But still, the distinctive professional contribution of historians to the protection of human rights, if any, is to see that the dead do not die twice.

See also: Cultural Property and Ancestral Remains, Repatriation of; Ethnic Cleansing, History of; Ethnic Conflicts; Genocide: Anthropological Aspects; Genocide: Historical Aspects; Holocaust, The; Human Rights, Anthropology of; Human Rights in Intercultural Discourse: Cultural Concerns; Human Rights: Political Aspects; War Crimes Tribunals

Bibliography

- American Anthropological Association Executive Board 1947 Statement on human rights. *American Anthropologist* 539–43
- Amnesty International 1961–*Report*. Amnesty International, London
- Andreopoulos G J (ed.) 1994 *Genocide: Conceptual and Historical Dimensions*. University of Pennsylvania Press, Philadelphia, PA
- Castberg F 1968 Natural law and human rights: an ideal-historical survey. In: Eide A, Schou A (eds.) *International Protection of Human Rights*. Almqvist & Wiksell, Stockholm
- Chalk F, Jonassohn K 1988 The history and sociology of genocidal killings. In: Charny I W (ed.) *Genocide: A Critical Bibliographic Review*. Mansell, London
- Chalk F, Jonassohn K 1990 *The History and Sociology of Genocide: Analyses and Case Studies*. Yale University Press, New Haven, CT
- Donnelly J 1982 Human rights and human dignity: an analytic critique of non-Western conceptions of human rights. *The American Political Science Review* 76 (whole issue)
- Hill Ch R 1969 Some philosophical problems about rights. In: Hill Ch R (ed.) *Rights and Wrongs: Some Essays on Human Rights*. Penguin, Harmondsworth, UK
- Human Rights Watch 1991–*World Report*. Human Rights Watch, New York, Washington, London, Brussels
- Johnson M G 1998 A Magna Carta for mankind: writing the Universal Declaration of Human Rights. In: Johnson M G, Symonides J (eds.) *The Universal Declaration of Human Rights: A History of its Creation and Implementation, 1948–1998*. UNESCO, Paris
- Jones P 1994 *Rights*. Macmillan, Basingstoke, UK
- Jongman A J (ed.) 1996 *Contemporary Genocides: Causes, Cases, Consequences*. PLOOM, Leiden, The Netherlands

- Kuitenbrouwer M, Leenders M. (eds.) 1996 *Geschiedenis van de Mensenrechten: Bouwstenen Voor een Interdisciplinaire Benadering* [History of Human Rights: Materials for an Interdisciplinary Approach]. Verloren, Hilversum
- Langley W E 1999 *Encyclopedia of Human Rights Issues Since 1945*. Greenwood Press, Westport, CT
- Laqueur W, Rubin B (eds.) 1990 *The Human Rights Reader*. New American Library, New York
- Lauterpacht H 1950 *International Law and Human Rights*. Stevens, London
- Maris C W 1992 Een Natuurlijke Historie van de Mensenrechten [A natural history of human rights]. In: Cliteur P B, Deelder A, Rutgers M R (eds.) *Mensenrechten* [Human Rights]. Gouda Quint, Arnhem, The Netherlands
- Rosenbaum A S (ed.) 1980 *The Philosophy of Human Rights: International Perspectives*. Aldwych Press, London
- Rummel R J 1994 *Death by Government*. Transaction Publishers, New Brunswick/London
- Rüsen J 1993 Human rights from the perspective of a universal history. In: Schmale W (ed.) *Human Rights and Cultural Diversity: Europe, Arabic-Islamic World, Africa, China*. Keip Verlag, Goldbach, Germany
- Schweib E 1968 Human rights. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences VI*. Macmillan, New York
- Szabo I 1978 Fondements historiques et développement des droits de l'homme. In: Vasak K (ed.) *Les Dimensions Internationales des Droits de l'Homme*. UNESCO, Paris
- United Nations Development Program 1992 Political freedom and human development. *Human Development Report 1992*. Oxford University Press, New York and Oxford, UK
- Weston B H 1994 Human rights. In: McHenry R, Hori Y C (eds.) *The New Encyclopaedia Britannica XX*. New Encyclopaedia Britannica, Chicago

A. De Baets

Human Rights in Intercultural Discourse: Cultural Concerns

Though human rights are generally recommended, there are still some strong criticisms. One criticism refers to human rights, somewhat derogatorily, as the civil religion of modernity and, by this, takes them to be only a historical phenomenon of a particular culture. Other critics fear that the spread of human rights is nothing but a new and subtle version of Western colonialism or imperialism, namely a supremacy of a its legal culture, which forces values on other cultures. But actually, human rights belong to a universal legal morality, which is common to all mankind although it was thought of in the most radical manner by Western modernity.

Within Western legal culture, human rights are taken as a matter of course. But as far as their foundations are concerned, there are difficulties which are partly of a political, partly of an intercultural, and partly of an anthropological nature. A specific philosophical justification, a transcendental foundation of human rights might help to overcome

these difficulties. This foundation, at the same time, might permit one to show the universal validity of human rights beyond liberal legal cultures.

1. Human Rights and Basic Rights

The first, political difficulties arise when one considers the possibility that a declaration of human rights might simply remain a declaration. For instance, the People's Republic of China used tanks against demonstrating students, although it recognizes freedom of procession and demonstration as stated in Article 87 of its constitution. The answer to this may simply be compliance with positive law. In order to give human rights an actual power in the world, instead of remaining just word on the page, they have to become solidly embedded in human practice. To this purpose human rights have to make their way into the sometimes written, sometimes unwritten constitution, thereby doing both: obliging the branches of government to human rights and restricting their power, among others rendering possible proceedings before the court of law against any abuse of governmental power and of the nonrecognition of rights. Wherever this occurs, human rights, which were previously just a component of universal legal morality, become elements of positive law. They are transformed from mere human rights into basic rights of a particular legal community.

Not only does this differentiation of human rights and basic rights direct itself against simple declarations, it also draws attention to those origins of basic rights which lie before and beyond the state. In particular, history's first declarations are aware of this elementary meaning. In this respect, the French *Déclaration* (1789) follows the Virginia Bill of Rights (1776). Their beginning establishes valid rights at a level which lies before the state. Subsequently, the basis for legitimation is laid for the exercise of political power; Virginia's Article 2 and France's Article 3 state that the people is the origin of all government. It is not until the third step that the function of the state is touched upon; here the limitation of the exercise of political power as well as the positive tasks of the state is mentioned. In that sense, from a philosophical point of view, we have to reverse Benjamin Barber's title: not 'Democracy first, then Rights,' but 'Human Rights first, then Democracy.'

This sequence contains my first thesis: human rights are rights of all human beings raised against all other human beings. From the standpoint of legitimation the claims in question are primarily raised by all men against one another and they are at the same time granted to one another. With regards to human rights, people recognize themselves as beings with equal rights. Positive law and the branches of government are legitimized in relation to human rights only at a second and, more specifically, subsidiary level: they are incumbent upon transforming innate rights in the

process of the establishment of a constitution and further legislation into positive law. The scholar of law says, 'rights are guaranteed'; with this he means a secondary accomplishment that is strongly differentiated from a primary accord. In German, we distinguish *gewähren* which is done by men themselves, and *gewährleisten* which is the task of the branches of government: the legislator, the executive power, and the courts.

2. An Intercultural Legal Discourse

The second difficulty with human rights is constituent of the fact that we expect the recognition of human rights from all cultures, but that the discourse on human rights only takes place within one of them, namely the Western culture. For this reason there is a risk that human rights, instead of constituting the core of a universal legal morality, will give away their essence and will thereby decline to the level of an 'export-article of Western culture.' In this sense, the objection of 'ethnocentric thinking' can be raised, for a legal institution like human rights, Western culture claims to be valid for the whole of mankind.

Although itself a product of the West, philosophy lends its voice to non-Western cultures. It does not do so in a now current form of a mutation, an act which turns philosophers into a kind of fundamental anthropologist, into theoreticians of a world society who, like Lyotard (1983), Feyerabend (1987), and Rorty (1989), place themselves in the tradition of American anthropology—one thinks, for example, of Herskovits' 'Statement on Human Rights' (1947)—and who, in their fear of destroying the characteristics of non-Western cultures, support a pure relativism. Instead of getting lost in the byways of fundamental anthropology, and its exaggerated relativism, philosophy engages in an intercultural legal discourse.

This discourse begins, in passing, by invalidating the mentioned fear. What Lyotard describes, somewhat drastically, as the 'tyranny of the universal' is based upon a misunderstanding, namely, the confusion of universality with uniformity as though universal legal principles should be equated with the removal of social and cultural differences. Let us take an example: the equal rights of different religions or that of different languages, from the standpoint of human rights, do not undermine the existence of one of them; by contrast, the principle of equality guarantees coexistence of the different parties.

In the next step, our discourse turns the reproach of ethnocentric thinking into a philosophical challenge, in other words, into a question whose answer is still open: how can a regionally developed institution which is not always taken seriously—we recall that the Virginia Bill of Rights was proclaimed in a slave state—lay claim to an absolute universality? Obviously, such a claim is permitted only on a condition which, from the standpoint of legitimation, is self-

evident but which is all too readily overlooked in the case of human rights. Legitimation must be separated from the historical genesis and evolution. My *second thesis* is this: whoever wants to impute the idea of human rights to other cultures must for once divorce himself from their intellectual history, also from their legal and social history, for instance from the question regarding economic prerequisites. If human rights are truly to exist, then they should do so not for the reason that capitalism, the modern form of the economy, requires a global protection of economic rights. Besides, there are well-known practices which contradict the idea of human rights and are nonetheless compatible with a global protection of economic rights, namely, slavery and human bondage, inequality of man and woman, the lack of rights for children and, last but not least, colonization. Furthermore, the legitimation of human rights also needs to be uncoupled from a certain phase of the modern state's development, namely from Reformation and Absolutism. If innate rights truly exist, then the Western experience can have a thoroughly heuristic significance; but in the end, only culturally invariant arguments have the power of justification.

3. A Copernican Revolution of Anthropology

A third difficulty with human rights is expressed in the question: what remains when one considers the separation of legitimation from the historical genesis and abstracts from the particular context of European modernity? The view that the idea of human rights is connected with a certain image of mankind is widespread. To this there is an annexation which not only profiles the matter as far as its content is concerned, but also limits its validity; namely that an image of mankind presents itself as Jewish, another as Christian, a third as Muslim, a fourth as Buddhist, and finally a fifth—a definitive neither/nor!—as atheist. Moreover, the modern European picture of mankind is individualistic and relatively ahistorical, whereas the African image of mankind is determined by a characteristic solidarity of two dimensions: a solidarity with fellowmen and with the ancestors. We can generalize: all specific images of mankind, all views of human nature are therefore culturally dependent.

The consequence for our justification: we even have to rid ourselves of Western images of mankind and evolve a *logos tou anthropou*. In other words, we have to evolve a concept of mankind which, valid for the whole of humanity, is independent of different histories and cultures.

A concept of mankind which is independent of cultural particularities, and therefore interculturally valid, has for a long time been no stranger to Western thought. Since Greek Antiquity, we know two anthropological definitions: men as *zoon logon echon* and as *physei politikon zoon*. Recently, however, skepticism

is the rule. If one lends an ear to philosophers like Rorty (1989, p. 44), then skepticism towards anthropology becomes such a matter of course that one may content oneself with only a short aside: 'fruitless by way of anthropology.' But scepticism is not new. Since Rousseau's concept of *perfectibilité* (*Discours sur l'inégalité*, part 1) and Nietzsche's dictum that man is 'more indeterminate than any other animal' (*'unfest-gestellter als irgendein Tier sonst'*: *On the Genealogy of Morals*, 3rd essay, section 13), philosophy has doubted whether it is possible to fix an essence of man that is independent of history and culture. In the 1920s it was Georg Lukàcs who supports this skepticism which was reinforced later in the 1930s by Max Horkheimer (1935). Their critique was not directed specifically at human rights, but it was partially to blame for the fact that Western Marxism has had such difficulties with this legal institution up to now. Lukàcs (1923, p. 204) sees it as 'the greater danger of every anthropological standpoint,' that it 'causes man to ossify as fixed objectivity and thus pushes dialectics and history to one side.'

The idea of human rights counters skepticism by weakening its claim. According to Aristotle's classical topos, *physei politikon zoon* (*Politics* I, 2), the specifically Greek form of a community, the *polis*, is the only place where man can lead a really meaningful life. Even if Aristotle is right and human potentialities can be fully realized only in a political community, the form of this community remains open. The above mentioned difference between universality and uniformity first becomes important here: human rights favor neither the Greek city state nor the modern form of territorial state nor any other form of political communities.

Anthropological relativization goes even further. In contrast to Aristotle's *Politics*, human rights allow a largely instrumental understanding of the political community. The state does not have to represent what the great neo-Aristotelian philosopher, Hegel, termed '*die Wirklichkeit der sittlichen Idee*' ('the actuality of the ethical Idea': *Philosophy of Right*, para. 257), at least not its full actuality. What a sociologist, a 'pupil' of Luhmann, Willke (1983) terms a 'demystification of the state' remains possible. The idea of human rights admits that the 'political system' establishes no more than premises and framing conditions, whereas true humanity can be sought elsewhere, and then in different ways. (Contemporary philosophy speaks of the difference between the right and the good.)

Where this openness for different ways is lacking and where what counts is a normative concept of man, a concept of the humane man alone, one runs the risk of denying elementary rights to people who are not so human. According to Aristotle, anyone who suffers a fundamental deficiency in terms of rational endowment—in short, anyone who is mentally handicapped—is *physei doulos* and rightly viewed as having lesser rights, in other words, he is taken as a slave. To

prevent this consequence, the idea of human rights is not allowed to favor particular forms of self-realization; with regard to the humane, we need an intentional nonspecification.

The institution of human rights intentionally contains what I shall call a fragmentary anthropology and for that reason alone can be imputed to other cultures. Although *prima facie* a drawback, nonspecification is in fact an advantage: far from being indifferent to more demanding definitions of man, it is, by contrast, open, ensuring that not only different cultures, but, within those cultures, different subcultures and individuals, enjoy the right to be unique, different, and even eccentric.

Nonetheless, this right is not unlimited. Human rights resist a radical pluralization. They alone make possible the very thing that critical intellectuals such as the 'fundamental anthropologists' mentioned earlier are so keen to achieve but which they are unable to situate adequately within their intellectual framework. The plurality of cultures and subcultures can exist only under conditions which are no longer pluralizable. Thus, in the idea of human rights modesty and immodesty enter into a characteristic union. To limit all plurality and to assume that conditions exist which are equally valid throughout history is an immodest aim. On the other hand these conditions make a modest claim, not demanding to present the form of a human life but to be only a condition of a plurality of forms of life and thus abandoning any normative concept of what is humane.

This connection is possible only if we readjust the anthropological perspective; philosophical anthropology has to undergo a conversion which fully deserves the ambitious title of a Copernican revolution. Instead of defining man from the standpoint that has normally been adopted since classical antiquity, that which allows him *eudaimonia*: happiness, self-realization, or a meaningful existence, we dispense with any teleological or normative concept. My *third thesis* is this: human rights are not concerned with perfecting what is humane but with its initial conditions, with elements which deserve the qualification of the relevant rhetoric: as elements inalienable to each person they are his inherent birthright. Here, we find my initial thesis: since we are dealing with conditions of the possibility to be a man, and not a man in a biological specification but as a responsible agent, we can speak of transcendental elements or o(relatively) transcendental interests, relative to (responsible) agency.

Transcendental interests alone leave the task of legitimation underdefined. This much is clear from even a cursory glance at legal history. Since man was regarded by Judaeo-Christianity as having been made in the image of God or as *zoon logon echon* by the Greeks, the West has been familiar with a universal concept of man. Yet, no legal consequences have been drawn from this; both the Jewish-Christian tradition

and the Greek one allow slavery and the inequality of women. A correlative phenomenon, familiar since the thirteenth century, is the sort of bill of rights typified, for example, by the Magna Charta Liberatum and the Golden Bull of Andrew II of Hungary. Although these charters were drawn up in states in which both ideas—'the image of God' and 'rational endowment'—were taken for granted, freedoms were not given to man as man. This deficiency remained—not, it is true, in the legal text, but in legal reality—until the time of the earliest declarations of human rights (we remember the case of a slave state proclaiming a bill of human rights).

Indirectly we stumble across a further difficulty here, which, from the point of view of legitimation, is even greater. Man should have a subjective claim to certain transcendental or anthropological interests. A claim can be made where another person has to meet that claim; subjective rights are bound by their very concept to correlative obligations, human rights to corresponding human obligations.

Three factors deserve emphasis here. First, the claims do not exist in an absolute sense; the right to exist does not mean, for example, that one does not have to die at all. Second, the right which, although only relative is not conditional, refers to ones fellow men and demands that they refrain from using force.

In an intercultural legal discourse, an objection arises at this point which is familiar from the liberalism/socialism debate: particular attention is given to civil rights and liberties, which, on the African side, are accused of favoring the individualist picture of man familiar to the West. To the extent that our fellow men are required to perform a service, not even civil rights are based on any individualistic anthropology. A social nature is assumed in an entirely traditional way, only its manner is no longer traditional. According to Aristotle, the *polis* is natural because man is invested with certain social impulses which ultimately produce the *polis* through a series of certain intermediary stages. The anthropology of human rights does not question these social impulses; it extends its horizons and discovers a complication which is the third factor deserving emphasis.

The complication has two consequences. In the first place, what has hitherto been called simply the 'social being' is given a qualifying rider: it becomes a *positive* sociality providing mutual help, with a *negative* counterpart in the form of a reciprocal threat. From an extended social anthropology—let us say 'cooperation plus conflict'—our legal institution singles out only one aspect, the conflict, namely, a threat from one of our equals, a threat, moreover, which affects the anthropological interests; these are threats of truly existential weight. Since these threats come from our fellow men, each individual is both a potential agent and, at the same time, a potential victim.

The anthropology bound up with being both victim and agent recalls Hobbes' dictum, *homo homini lupus*,

and, like that dictum, sees itself exposed to the reproach that it presents a pessimistic image of man. But, at least as far as the idea of human rights is concerned, this reproach is unjustified. Being modest once more, negative social anthropology is not claimed to be exclusively valid but, once again, to be fragmentary anthropology—not a basic feature of the individual but a danger which cannot be excluded.

The complication caused by threat has a second consequence, inasmuch as the evolution of the *polis* can no longer be regarded as the mere development of pre-existent social impulses, a far greater contribution of ones own is necessary; although he is a social being, man has to produce himself as a social being. Society arises through mutual recognition, which can assume a variety of forms; human rights emphasize only a small part, and that part concerns the elementary level of anthropological interests.

The anthropology of human rights relates to existential threats but refrains from passing judgement on the nature of the conflict, whether that judgement be positive or negative. Therefore, a *fourth thesis* reads as follows: the danger of conflict is accepted as an anthropological fact which is favored in relation to man's cooperational nature only to the extent that it is his conflictual nature alone which makes human rights necessary. Where social relations consist only of mutual help and where help is not only always spontaneous but can be relied on, in a word, where man's positive social nature exclusively is involved, human rights lose their task.

4. Transcendental Interests and Transcendental Reciprocity

It is the fact that elementary interests are at stake that constitutes the seriousness, even existential seriousness of human rights and explains the feelings and the pathos bound up with this institution. An indispensable element—the character of a claim—is not yet clear, however. In many arguments there is a hiatus at this juncture, a legitimate leap which fails to explain why a subjective claim insists on the recognition of transcendental interests. I do not dare to suggest that I can justify this claim in the face of every doubt, but I am making a proposal based on a generally recognized principle.

A moral claim insists on the recognition of a service where the service is rendered on condition that a corresponding service be rendered in return. This generally recognized principle regards the obligations of legal morality (and not of all morality) as a giving that is made on condition that a countergiving is made; it operates with the reciprocity of giving and countergiving, you could say, with an exchange. Anyone who lays claim to those services which are rendered only under the condition of reciprocity now

stands under an *obligation*. Conversely, a (subjective) *right* is involved as long as one actually renders the service which is performed under the condition of a service in return.

In terms of human rights, this general principle must now be applied to the particular case of anthropological interests. For them, there is a correlation between rights and duties where certain interests not only have an anthropological status but also fulfil the condition that they are realized only in and through reciprocity. Here, and only here, in the case of an 'inherent reciprocity,' or an inherent sociality, is one no longer free to say whether to accept an exchange or not. To the extent that, in keeping with their qualification as anthropological, the interests are indispensable and to the extent that they are also bound up with the idea of reciprocity, the indispensability is transferred to the reciprocity; the corresponding exchange is similarly indispensable.

The proposal obtains a moral, albeit undemanding moral element: at the basis of human rights is a modest morality that is withdrawn from the usual controversies. This morality—fairness in exchange—applies in those cases where it is not a question of one person giving and the other taking, but, rather, where there is a reciprocity of giving and taking and where, moreover, the giving and its recompense are more or less equal in value.

Since a moral element is present, the legitimation of human rights obeys the structure 'anthropology plus morals.' In this legitimation the moral part is as good as uncontroversial, but the same cannot be said of the anthropological part. Human rights represent a problem that is very much bound up with legal *ethics*, and yet it is their other aspect, that of legal *anthropology*, which contains the greater difficulties. At least on *their* account what was formerly a great discipline, philosophical anthropology, merits a minor revival.

It is the both-victim-and-agent anthropology which allows us to describe human rights as an exchange which anyone can enter into with anyone else. Each of us exchanges his own potential for committing an act of violence for the interest of not falling victim to another's violence. Characteristic of this exchange is its inherent reciprocity: one can realize one's own interest only by virtue of a service which the others—all the others—perform. In addition, we rediscover a difference to traditional social philosophy. Mutuality consists not of positive services but of a negative reciprocity, in an all-round rejection of violence. Nonetheless, the exchange has a positive significance as such, and not only as consequence, otherwise there would be no reason to agree to it. When the corresponding human *obligations* are fulfilled, human *rights* should be granted. From the reciprocal refusal to kill one's fellow man, to rob them, or to prevent them from practicing their religion issue the right of life and limb, the right to property, and the right to religious freedom.

The question which now arises is whether, for the sake of the positive purpose, each of us should agree to this negative exchange. Does the condition 'innate interest' apply just as much as the condition 'inherent reciprocity'? The question has the character of a preferential choice which is decided on the question of what people prefer. Do they prefer to have the right to kill, but coupled with the danger of being killed themselves, or do they prefer the integrity of life and limb, but at the price of having to accept the corresponding obligation and no longer kill any person? Or, put another way: does the individual prefer to be both the victim and the perpetrator of violence or does he prefer a neither/nor situation?

According to a first attempt of philosophical legitimation—that of Hobbes—every man has an overriding desire: not to die at the hands of a fellow man. The English Civil Wars of that time make it clear enough that many people see religious freedom as more important, while others prefer political or cultural ideas, and a third group prefers honor.

The counterexamples can be subsumed under the heading of self-preservation. What is preserved, however, is not the physical self but a 'moral self.' Is it here, therefore, that the predominant interest is to be found? In the conflict of 'physical vs. moral self,' one can certainly decide in favor of mere survival. If there is really an innate interest, it must be sought on a logically higher level, beyond the conflict of 'physical or moral self.'

A third attempt remains with the idea of an overriding desire, but takes that desire not as something negative but, rather, positive, something, moreover, which is no longer a dominant but an all inclusive interest. The interest that includes all other interests within it we call happiness. As it fulfils all needs and interests, it forms the natural guiding principle of every individual. With so formal a definition, however, it is not possible to answer the question, whether a person would rather perpetrate a violent act and at the same time be its victim or whether he would prefer neither one nor the other. If we introduce a further definition, therefore, and insist on defining the content, the individual is left with the possibility of sacrificing his happiness to a fellow man or to a religious, cultural, or political ideal. As a formal definition, happiness is unrelinquishable, but it forms too weak a premise; conversely, a definition which takes account of content reflects a predominant desire which can be seen in relative terms and, like this, represents too powerful a premise.

Anyone who concludes from this that there is no interest in the corresponding exchange fails to see that many people want nothing more than their own survival, while others want only religious freedom, others again only political freedom, and so on. On each occasion, the corresponding rights are advantageous, but they are not the same rights for each person. In other words, people are happy to vote for

human rights, but they interpret that vote very differently.

We need to look for a fourth answer which preserves the strengths of the first three attempts while overcoming their weaknesses. On the one hand, we need a stronger premise than a merely formal definition of happiness; on the other, the undue assumption of an absolutely overriding desire must be abandoned; what we are looking for is an interest that is both substantial and nonsubstitutable. My *fifth thesis* is as follows: this double condition is met by interests which one can neither abandon in favor of higher interests nor weigh against interests of equal importance. Truly indispensable—and, therefore, anthropological—interests consist of elements which, logically, are on a higher level and which, *faute de mieux*, are called (relatively) transcendental interests. To the extent that, in our abstaining from the use of force, our interests are realizable only in reciprocity, in other words, in an exchange, one can speak—with the same proviso—of a transcendental exchange or a transcendental reciprocity. Human rights refer to conditions of agency based on reciprocity.

The idea of human rights presupposes that there are indeed interests on this higher level and that there is, therefore, an exchange on a logically higher level. Let us take an intentionally uncontroversial example—the protection of life and limb. Even someone not especially attached to life has an interest—conscious or unconscious—in that life, since he can otherwise neither desire anything nor seek to satisfy his desire. Independently of what one strives to achieve or seeks to avoid in terms of life's content and, therefore, a condition of our ability to act, a condition of agency, life is the precondition of a practical desire. It is at this point that the singular nature of human rights emerges most clearly, without the corresponding acts of renunciation, the very thing that distinguishes man—his agency—is not capable of existing within a social perspective. This is true even of those who are prepared to sacrifice their own lives: the martyr wants to decide in what cause he lays down his life, in other words, he acts out of religious or political conviction and not, say, because of some robber's act of violence.

The natural law of the Enlightenment appealed to the principle of self-preservation. If I consider the principle of agency to be more convincing, it is because it is more formal and, therefore, has fewer preconditions attached to it. When I speak of agency, it is not, of course, an ability to act *sans phrase* but simply that part which involves reciprocity.

5. Against a Liberalist Restriction

Man is not merely a living being. More than life and limb are involved in the indispensable conditions behind his agency; in consequence, there is more than

just this one particular human right. We can argue directly on the basis of our agency. Since the concept of action involves intentionality, the conditions for agency also include the conditions of intentionality, in other words, the ability to speak and think. Furthermore, in accordance with the claim that man cannot become a man but within a society of fellow men, agency is bound up with positive social relations and cooperative conditions, thereby opening up an additional field of transcendental interests. (We find the traditional anthropological definition of man: he is *zoon*: an animal, *logon echon*: endowed with reason and language, and *zoon politikon*, a social being.)

My proposal: the majority of human rights can be reconstructed on the basis of these three aspects, which, taken together, make the ability to act or agency, a real possibility, in other words, the living being, the being who is able to talk and to reason, and the cooperational being. The insight contained herein, namely, that there are further transcendental interests, is not of course sufficient to proclaim corresponding human rights. In order to do that, we also need to prove a transcendental reciprocity.

The politics of recent decades have grown sensitive to other aspects, including questions of personal, social, and cultural identity, of economic and social prerequisites and, not least, ecological conditions. Here, too, there are forms of reciprocal threat for which our legitimation model remains applicable. At all events, we are dealing with new forms, with threats which are both indirect and which involve a collective responsibility, so that the legitimation model needs to be developed further. But even here it can be stressed—and this is my *sixth thesis*—that the last word has not yet been spoken on human rights within the classic, liberalist canon.

According to classical liberalism each should know his interests best; but since Nietzsche and Freud this view has prevailed as naive. The concept of transcendental interests reaches far beyond this controversy. Within elementary interests, transcendental refers to a stratum that we have not only independent of attitude but also independent of being man or woman, Christian, Muslim, or nonbeliever. The singular character is due to the fact that one does not want anything else but the following: to be an entity capable of action.

According to Rorty (1989, p. 28), the thought that the belonging to a biological species implies certain rights is a type of metaphysical consolation, which the pragmatist renounces. The preceding legitimation shows on the one hand that it is not the biological belonging that establishes rights, but rather the interaction of transcendental and moral arguments; and on the other hand, that for a metaphysical consolation not only the pragmatist but also the theoretician of human rights can renounce.

A legal order cannot safeguard human rights without governmental powers. My *seventh thesis*—a thesis

which is itself subdivided—argues against anarchism. According to this thesis, a public power is needed first to provide a more detailed definition of human rights; this is the task of the legislator. But even with more detailed definitions there are controversial interpretations, so that an impartial and authoritative interpretation, i.e. an independent judicial system has to settle the relevant controversies. Finally, there is another complication, and once more a complication in the form of a danger: one can vote for the recognition of rights and still attempt to enjoy the advantage of a mutual renunciation of freedom without paying the price of renouncing of one's freedom. It is above all in those instances where countervailings do not occur at the same time as the giving that one can enter into an exchange of freedoms in the manner of a fare dodger. This possibility reveals a further element of negative social anthropology, namely, the danger of parasitically exploiting a universal advantage. Public powers are demanded a third time, therefore being needed here as an implementational force—a kind of 'sword of justice.'

Even if more detailed considerations are needed for a convincing legitimation (see Höffe 1987, Chaps. 13–14), this hypothesis of differentiation appears nonetheless plausible. Looked at from the standpoint of the institution of human rights, the classical—specifically Aristotelian—idea of a political nature develops as two partial natures, one of which is concerned with the law, the other with the state. Man is a being that needs, on the one hand, elementary rights and, on the other, the public definition and implementation of those rights, and in this context, he needs democracy. By consequence: rights first, but in any case, democracy, too. Against an overrating of the state, this aspect is secondary, however, being subsidiary to the primary task, which is that, in order to be a human being among his fellow men, man needs elementary rights and correlatively elementary obligations.

See also: Reproductive Rights in Affluent Nations; Reproductive Rights in Developing Nations

Bibliography

- Feyerabend P K 1987 *Farewell to Reason*. London [dt. *Irrwege der Vernunft*, Frankfurt/Main 1989]
- Höffe O 1987 *Politische Gerechtigkeit. Grundlegung einer kritischen Philosophie von Recht und Staat*. Frankfurt/Main [Political Justice, Cambridge/Oxford 1995]
- Herskovits M J 1947 Statement on human rights. *American Anthropologist* 49: 539–43
- Horkheimer M 1935/1988 *Bemerkungen zur philosophischen Anthropologiein: Gesammelte Schriften* 3. Frankfurt/Main, pp. 249–76
- Lukács G 1923 *Geschichte und Klassenbewußtsein* Malik-verlag, Berlin. [History and Class Consciousness: Studies in Marxist Dialectics. London 1971]

- Lyotard J F 1983 *Le différend*. Editions de Minuit, Paris [*The Differend: Phrases in Dispute*, Manchester 1989]
Rorty R 1989 *Contingency, Irony and Solidarity*. Cambridge University Press, Cambridge and New York
Willke H 1983 *Entzauberung des Staates: Überlegungen zu einer sozialen Steuerungstheorie*. Athenäum, Königstein, Germany

O. Höffe

Human Rights: Political Aspects

Human rights are, literally, the rights that one has simply as a human being. Although the fate of human rights is largely determined by national political struggles and practices, the focus of this article is international, because only in international studies has human rights become a well-established 'mainstream' concern of a social science discipline.

Prior to World War II, human rights was an ignored, even suppressed, issue in international relations. Even the notoriously 'idealistic' Covenant of the League of Nations does not mention human rights. The turning point was the creation of the United Nations, in the shadow of the Holocaust. The UN Charter lists promoting human rights as a principal objective of the Organization. Even more importantly, the Universal Declaration of Human Rights, adopted by the General Assembly on December 10, 1948, established substantive international human-rights norms that have come to be accepted, at least in word, by almost all states and most leading movements of political opposition in all regions of the world.

In the first two decades after the Universal Declaration, however, both superpowers used international human rights largely as an instrument of partisan political abuse. Although decolonization helped to spur relatively nonpartisan international attention to self-determination and racial discrimination, through the early 1970s no prominent state pressed international human-rights concerns except in the presence of strong political interests.

Rights that are given, either by 'benevolent' dictators or foreign powers, usually end up being rescinded. Only when right-holders are able to, and at least periodically do, act to vindicate their rights are they likely to be enjoyed securely. Thus foreign states and international organizations, as well as transnational nongovernmental organizations (NGOs), have increasingly focused their efforts on supporting the development of autonomous, intermediate associations in civil society, in recognition of the fact that international action usually can only be a supplement to national action.

The other major change of the post-Cold War era has been the growing integration of human rights into other areas of the international action. For example,

at the UN during the Cold War, human rights were rigorously segregated from peace and security, and development activities. Today, a human-rights mission is a standard part of peace and security operations. Furthermore, the threats to peace and security to which the UN responds are most likely to arise from systematic human-rights violations.

The penetration of human rights into other domains of international action is far from complete. Most notably, although the United Nations Development Programme stresses the links between human rights and development, human rights have made almost no headway in the work of the International Monetary Fund, which regularly imposes structural adjustment programs that have significant negative impacts on economic and social rights. Nonetheless, at the start of the twenty-first century human rights are increasingly seen as not merely a regular part of international relations but a concern with vital rights enumerated in the Universal Declaration. Thus the Vienna Declaration, adopted by consensus by 171 participating states, in its first operative paragraph 'reaffirms the solemn commitment of all States to fulfill their obligations to promote universal respect for, and observance and protection of, all human rights and fundamental freedoms for all.'

As human-rights values have filtered into the consciousness and practice of states, bureaucrats, citizens, and civil societies across the globe, the bases of national and international political legitimacy have been transformed. The Vienna Declaration, again in its first operative paragraph, proclaims protecting and promoting human rights 'the first responsibility of Governments.' And in sharp contrast to half a century ago, governments today can expect to be reminded, often loudly, of their human-rights responsibilities by foreign states, international and regional organization, the transnational human-rights movement, and, to the extent that they permit it, their own citizens.

The global human rights regime, however, is based on *national* implementation of international human-rights norms. Norm creation has been thoroughly internationalized: the UN Commission on Human Rights drafts treaties and declarations through a process that emphasizes consensus. Implementation, however, lies almost exclusively with sovereign states.

Although the UN Security Council has asserted its enforcement powers in a handful of cases, multilateral human-rights institutions have monitoring rather than enforcement powers. The UN Commission on Human Rights, in addition to its vital work in developing international norms, investigates violations in particular states and substantive issues of broad international concern. The Human Rights Committee (HRC) and the Committee on Economic, Social, and Cultural Rights, which supervise state compliance with the 1966 International Human Rights Covenants, are restricted principally to reviewing periodic reports submitted by states. (The HRC also investigates

individual complaints, but these procedures have been used infrequently—less than 250 decisions in more than 20 years—and its findings have no legal force.) Supervisory committees created under other international human-rights treaties—dealing, most prominently, with racial discrimination, women, torture, and children—are also largely restricted to reviewing and commenting on periodic state reports.

Such reporting schemes have considerable value. In conscientiously preparing a report, states often identify areas where improved performance is possible. At minimum, states know that they are subject to periodic scrutiny by an impartial international body. But these supervisory committees cannot even order, let alone, compel states to alter their practices. Multilateral action thus is most likely to bring change not where violations are most extreme, and thus ‘most needed,’ but in countries where violations are less severe and where governments have a significant internal motivation for change.

The principal exception is genocide, where a system of international enforcement (although not prevention) does seem to be emerging. The Statute of the International Criminal Court, adopted at Rome in 1998, establishes individual criminal responsibility for genocide, crimes against humanity, and war crimes. Furthermore, antigencide norms are increasingly enforced through armed humanitarian intervention (e.g., Bosnia, Rwanda, East Timor). But for all other human-rights violations—that is, all but a tiny (if important) handful of violations—multilateral institutions monitor state practices rather than enforce international norms.

The powers of regional human-rights organizations vary immensely. At one extreme, the European Court of Human Rights provides authoritative judicial enforcement of the 1950 European Convention on Human Rights for the members of the Council of Europe. At the other extreme, there are no regional human-rights institutions for Asia and the Middle East. The Americas lie about halfway in between. Members of the Organization of American States (OAS) are subject to investigation by the Inter-American Commission on Human Rights, but the Inter-American Court of Human Rights has been much less active, and considerably less effective, than its European counterpart. The African (Banjul) Charter of Human and Peoples’ Rights has created only a very weak commission and no court.

In looking at international implementation, we also must recognize that states, for all the talk of globalization, remain the central international human-rights actor, not merely through their control of international and regional organizations but in their bilateral relations. But they still typically apply only the weaker of their available means—praise and protests, symbolic actions, and occasional financial incentives and penalties—in the pursuit of human-rights objectives. No state, weak or strong, ‘nasty’ or ‘nice,’ has been

willing to sacrifice vital national interests, or even very much in the way of economic, security, or political interests, to human rights goals.

But states also lack the authority to use force unilaterally against international human-rights violators. And whether undertaken by states or regional international organizations, sanctions short of direct military intervention, as the embargos against Cuba, Iraq, and Serbia suggest, typically inflict far more harm on innocent victims than the targeted regimes. Furthermore, even where external coercion ends abuses or topples a repressive regime, this usually is only a first step toward establishing a rights-protective regime, as Somalia and Bosnia suggest. Thus even where external actors are willing to use coercive means to pursue their human-rights objectives, it is not clear that they have the capabilities to have a decisive positive impact.

This suggests an unusually important role for transnational human rights NGOs. They have been, and remain, important transmitters of international human-rights norms. They have a single-minded focus on human rights, in sharp contrast to states and international organizations, which must balance a range of often competing interests. And they increasingly understand their role as supporting the local human rights advocates and progressive elements in ‘civil society’ that, along with local governments, hold the key to the future of human rights.

Four events of the mid-1970s signaled the maturing of human rights as an international issue. The 1973 coup in Chile provoked a major and sustained international reaction that for the first time cut across Cold War divisions, with many European states joining the criticism of the US-backed military dictatorship. The Helsinki Final Act of 1975 introduced human rights into the mainstream of East–West relations. The award of the Nobel Peace Prize to Amnesty International in 1977 marked the emergence of transnational NGOs as significant international human rights actors. And when Jimmy Carter became president of the United States in 1977, one of the world’s two leading powers began to raise human rights issues with allies as well as enemies.

Human rights violators continued to meet criticism with vigorous appeals to sovereignty and non-intervention. States rarely went beyond verbal and symbolic action in support of their human rights objectives. The moral power of international and non-governmental organizations to mobilize shame often ran up against shameless governments. Nonetheless, by the early 1990s the international human-rights landscape had been significantly altered.

At the 1993 World Human Rights Conference in Vienna, the Cold War ideological wranglings over the relative priorities of economic, social, civil, and political rights, and parallel North–South debates over the relationship between human rights and development, were largely discredited. Although national

or regional culture continues to be used to justify substantial derivations from international norms, such arguments are generally rejected for the linkages to a wide range of other national and international interests.

We can see a comparable maturing of human rights as an area of scholarly inquiry. Just twenty-five years ago, one could count on one's fingers the number of social scientists who specialized in human rights. Through the end of the Cold war, the scholarly literature on international human rights was largely legal, institutional, and policy oriented, as well as poorly integrated within the leading disciplines. In recent years, human rights has moved into the mainstream of political science and international studies, at least in the English-speaking world. And we are beginning to see a new generation of scholarship that increasingly adopts mainstream disciplinary methods. Thus this encyclopedia appropriately includes an entry on human rights, whereas a comparable project just a quarter century ago almost certainly would not have.

Bibliography

- Alston P (ed.) 1995 *The United Nations and Human Rights: A Critical Appraisal*. Clarendon Press, Oxford, UK
- Cortright D, Lopez G 2000 *The Sanctions Decade: Assessing UN Strategies in the 1990s*. Lynne Rienner, Boulder, CO
- Donnelly J 1989 *Universal Human Rights in Theory and Practice*. Cornell University Press, Ithaca, NY
- Donnelly J 1998 *International Human Rights*, 2nd edn. Westview Press, Boulder, CO
- Dunne T, Wheeler N J (eds.) 1999 *Human Rights in Global Politics*. Cambridge University Press, Cambridge, UK
- Forsythe D P 2000 *Human Rights in International Relations*. Cambridge University Press, Cambridge, UK
- Hannum H 1992 *Guide to International Human Rights Practice*, 2nd edn. University of Pennsylvania Press, Philadelphia, PA
- Risse T, Ropp S C, Sikkink K (eds.) 1999 *The Power of Human Rights: International Norms and Domestic Change*. Cambridge University Press, Cambridge, UK
- Shue H 1980 *Basic Rights: Subsistence, Affluence, and US Foreign Policy*. Princeton University Press, Princeton, NJ
- United Nations Development Programme 2000 *Human Development Report 2000: Human Development and Human Rights*. Oxford University Press, New York
- United Nations Human Rights Website, <http://www.unhchr.ch/hchr.un.htm>
- Wheeler N 2000 *Saving Strangers: Humanitarian Intervention in International Society*. Oxford University Press, Oxford, UK

J. Donnelly

Human Sciences: History and Sociology

Reference to 'the human sciences' is loose and variable. With few exceptions, there is no correspondence between the term and existing disciplines. Indeed, the

term has different connotations for different academics. For some, it refers to what they hope will be a field constituted by debate about what sort of systematic knowledge of human beings is possible, not by a particular body of knowledge. By contrast, others think the field is simply the biology of human beings. Most commonly, the term is one of convenience: 'the human sciences' covers the past and current pursuit of knowledge of human activity, without specifying precisely what is included. The term may cover psychology, linguistics, sociology, anthropology, economics, human geography, and political science, and also parts of the organizational sciences, history, biomedicine, cognitive science, neuroscience, and other areas as well. Nevertheless, it is increasingly common, and significant, for the term to signify a cluster of debates about knowledge and the achievements—or limitations—of existing disciplines.

1. Comparative Meaning and Usage

The difficulty of providing precise definition comes from the subject. 'Man' is not an empirically delimited subject. Theoretical argument about how to characterize the subject, rather than empirical investigation of a given domain, is central. At the same time, empirical research from areas of activity as divided from each other as evolutionary biology and cultural studies have considerable influence. The content of the field is variously said to include the biological human species, the 'Being' of 'Man,' language itself, socially and historically constructed culture, practices of ordering human life, and so on.

When the term refers in an open-ended way to knowledge of humans, usage relates to the '*sciences de l'homme*,' a phrase denoting a conventional division of Francophone academic life covering the psychological and social sciences. There is also overlap between the human sciences and the '*Geisteswissenschaften*,' the sciences concerned with the activity of the human mind or spirit. As German-language usage indicates, the meanings of the word 'science' itself enter argument about the human sciences. In Continental Europe, a disciplined, rational body of knowledge forms a 'science' (theology, for example, may be referred to as a human science), while in the Anglophone world, 'science' is often shorthand for 'natural science.' Except in the usage of some natural scientists, reference to the human sciences follows Continental European usage and includes humanities disciplines as potential 'sciences.' Looked at historically, the human sciences have roots in the eighteenth- and nineteenth-century 'science of man' and 'the moral sciences.' These English-language expressions referred to areas of knowledge thought to be distinctive by virtue of human reason, language, morality, and society.

In the social sciences, it is becoming common for 'the human sciences' to include disciplines that are not subsumed by the natural sciences and that share in

common the goal of understanding human activity and its products. Institutionally, these disciplines often form social science faculties, though they may also spread into natural science faculties (e.g., psychology) or humanities faculties (e.g., history). Reference to the collective term is more than a matter of convenience. Deep and complex issues are at stake. One view holds that debate about the nature of knowledge, or 'science,' of the human subject is intrinsic to being human; 'human science' is therefore an essentially contestable area. The long history of disagreement about the identity, content, and methods of all aspects of the human sciences supports this view. This disagreement continued in the late twentieth century, especially in controversy about the respective contribution of the natural sciences and humanities.

But sometimes scholars have held the view that one particular science provides the foundation for a unified human science. Around 1800, French intellectuals argued on behalf of '*idéologie*,' a science of the origin of action and ideas in sensation and organization. In the USA in the 1950s, there was an attempt to fashion 'behavioral science' as unified human science; more recently, some scientists have promoted evolutionary biology as the basic science of human existence.

2. The Evaluative Connotations of Usage

Earlier encyclopedias of the social sciences contain no reference to the human sciences. Recent usage reflects changing arguments and values in the academic community. First, a number of scholars welcome the interplay of different approaches, and do not think disagreement is a failing of science; according to some, disagreement is intrinsic to knowledge of humans. Second, they argue that the natural sciences do not, and cannot, provide explanations for much of the human sphere. This point is made because natural science knowledge does not address certain topics (e.g., symbolization, and meaning), not because natural science is thought undeveloped.

Third, many scholars hope to establish new relations between the psychological and social sciences and the humanities (that is, disciplines like art history and literary theory). Fourth, an interest in the human sciences often indicates sympathy with the view that history of belief in human realities and their representation is necessary to knowledge. Finally, usage expresses conviction that the existing social science and humanities disciplines are historical, social constructs, which structure knowledge in contingent and possibly restrictive ways. Deeper knowledge may need an overarching conception, signaled by reference to 'the human sciences.'

The spread of the term therefore correlates with the critique of positivist views of the social sciences and of 'behavior' as a unifying concept, and interest in rethinking the relations of history and scientific knowledge and overcoming disciplinary specialization in the

psychological and social sciences. Some observers see correspondence between these values and the spread of a postmodern ethos. The welcome given to open-ended argument, skepticism about the unification of knowledge, and turn to the examination of 'science' using theories and methods developed in other areas of academic culture, does suggest that spreading interest in the category, 'the human sciences,' is part of a wider change.

Yet, there are some scientists for whom the term 'the human sciences' simply describes the extension of the natural sciences to encompass the human sphere. This usage is rejected by most academics. They do not deny that human beings are part of nature (and hence a subject for natural science), but they maintain that human activity is self-created through historical processes of reflection and language (and hence that there are non-natural science forms of objective knowledge). There are major differences of emphasis among those scholars who have no pretension of creating a unified human science. Some, such as the philosopher Charles Taylor, may be called 'humanist.' They have argued for the distinctiveness of the human sciences because of values thought intrinsic to human activity. This orientation has been critically dissected in 'the linguistic turn,' a cluster of arguments about language or representation—rather than 'nature,' 'Being,' or 'Man'—as the subject matter of knowledge. Scholars such as Michel Foucault may be called 'antihumanist,' as their projects deny reference to anything 'essentially human,' and instead consider only the conditions of possibility for knowledge at a particular time. Yet others, such as the philosopher of science, Ian Hacking, find, ways to hold in tension this polarity while opening new views of the human subject.

3. History of the Human Sciences

The modern disciplines of psychology and the social sciences are creations, for the most part, of the late nineteenth and early twentieth centuries. Their establishment is part of the modernizing process, the recreation of human activity as subject to rational discipline, evident at least from the early modern period in Europe. During what historians call 'the Enlightenment,' there was a deeply influential attempt to extend knowledge of laws of nature to include individual and social activity. Scholars worked in a variety of settings, outside as well as within the academy, and according to many intellectual agendas. Giambattista Vico, whose '*scienza nuova*' encompassed myth, poetry, and history as human creations, was a professor of rhetoric at the University of Naples. David Hume wrote 'of human nature' as a literary essayist, while his friend Adam Smith taught about 'the wealth of nations' in courses on moral philosophy at the University of Glasgow. Montesquieu was a French aristocrat whose comparative study of legal systems, government, and custom enhanced notions of

the relativity of social life, while J. G. Herder, a civil servant in Saxe-Weimar, linked the fate of peoples and the fate of culture and language.

There was, however, no reference to 'the human sciences' but rather to 'the science of man.' Historians use the former term to encompass the background for all the intellectual activity divided since the late nineteenth century among the modern disciplines. The philosopher of the social sciences, Peter T. Manicas, for example, claimed that political economy was the first 'scientific' human science. Historians of 'the human sciences' distinguish themselves from historians of psychology, economics, or other disciplines, by stressing the historically relative classification of knowledge and practice.

In the nineteenth century, G. W. F. Hegel, Auguste Comte, Karl Marx, Herbert Spencer, and Charles Darwin each brought human existence within the scope of a unified 'science.' Each attempted to make human existence intelligible through historical forms of explanation. The very generality and scope of their work, however, contributed to the major revisions associated with discipline formation in the social sciences. Speculative historical reconstruction integrated poorly with the disciplined search for precise, empirical knowledge, and with the knowledge generated in the context of social problems.

The outcome was a commitment to differentiating professional and amateur, to institutionalizing research and training within universities as disciplines, and to separating what was held to be science from application. There was debate about the foundations of knowledge and relations between knowledge of nature and knowledge of culture. J. S. Mill, under the heading of 'the moral sciences,' stated the case for explanatory continuity, though he thought that an enormously complex task lay ahead to explain social life by laws of human nature. In the German- and French-language universities, there were arguments, still cited at the end of the twentieth century, about the need for hermeneutic, or interpretative, understanding of human actions; about the distinction between nomothetic (or law-like) and idiographic (or particular) explanations; about structural versus historical explanation (especially in economics and linguistics); and about how values relate to science. The participating scholars—such as Wilhelm Dilthey, Gustav Schmoller, Emile Durkheim, Ferdinand de Saussure, and Max Weber—give the human sciences many of their primary reference points.

In the English-speaking world, especially in the USA, the outcome was the establishment of empirically based disciplines, often accompanied by belief that this followed natural science example. The US model of professional research became dominant almost everywhere after 1945. For many researchers, this carried forward the enlightenment vision of improving human life on the basis of objective knowledge. But only when there were hopes for a more

unified and more effective social science was there discussion of a global project comparable to 'the science of man.' In the interwar years, the Rockefeller Foundation supported a science of personality and society. The Ford Foundation's entry into the social science field in 1952 fostered 'the behavioral sciences.' A commitment to unification was matched at the level of national politics by a perceived 'moral lag,' belief that natural science and technology had outstripped the political and moral capacity to order human affairs.

Reference to 'the human sciences' dates mainly from the 1980s. Weaknesses in both Marxist and empiricist philosophies of social science led to an historical as well as philosophical re-examination of theory. Developments in the humanities and social sciences directed empirical research to the 'representation' of knowledge and power rather than to questions, customary in the empiricist mainstream, about objective truth and material reality. By contrast, some biologists concurrently claimed that their field was uniquely competent to understand human beings. Thus, different groups, with different interests, found it useful to promote the same generic term. The only English-language term available was 'the human sciences,' a term preferred to 'the science of man' in the English-speaking world, as it appeared neutral in relation to gender.

The social circumstances were not propitious for new large-scale institutional arrangements, and the number of scholars who identified themselves as primarily occupied with the human sciences, as opposed to occupied in one of the existing disciplines, remained small. Nevertheless, some organizations, particularly emphasizing history, appeared: a journal, *History of the Human Sciences* (founded 1988); the European Society for the History of the Human Sciences; the *Société Française pour l'Histoire des Sciences de l'Homme*; and a Forum for the History of the Human Sciences within the History of Science Society in North America. When institutions or conferences use the term 'the human sciences,' it tends to be a utilitarian response to local organizational needs, bringing under a common administrative, and sometimes research, umbrella such disciplines as anthropology, applied social science, and social psychology.

4. Theoretical Content

If we take the view that the content of the human sciences is open-ended, the field has the following character. First and foremost, the field does not presuppose what the human subject *is*, but seeks to analyze and understand the historical and epistemic conditions of human activity at any particular time and place. There is therefore a literature about the representation of human action and human nature—

in practice as well as in written, visual, or other sources—and about such basic categories as ‘economy,’ ‘modernity,’ and ‘the self.’ The field, understood in this way, does not presuppose evolutionary biology or, in common language, ‘human nature,’ as the subject of objective knowledge, but rather seeks to understand how what is accepted to be objective knowledge is created through historically or humanly constituted processes. The humanities disciplines have a marked influence, encouraging interest in such topics as symbolic representation, rhetoric, and identity formation, whether of individuals, gender, ethnicity, or culture.

The field overlaps with ‘the history and philosophy of the social sciences.’ But whereas earlier English-language philosophy tended to take for granted the empiricist agenda of social science, and examined, for example, the nature of law-like explanation, realism, and reason and agency, the new field does not. It is more concerned with the constitution of agendas in the first place, and hence examines topics such as the influence of disciplinarity on knowledge, power and representation, and the ‘given’ quality of such basic concepts as ‘society’ and ‘emotion.’ Empirical studies of particulars, rather than abstract theorizing, often suggest the form of analysis.

Recent history and sociology of science (i.e., of natural science) developed independently, for the most part, of questions specific to the psychological and social sciences. But, of course, the sociology of scientific knowledge and technology is itself a major contribution to, and part of the history of, the social sciences. Some scholars, like Kurt Danziger, have extended the sociology of knowledge to the psychological and social sciences. More commonly, however, historians have informally employed the new sociology of science to explain activity in the social sciences in terms of specific factors of time and place—that is, to provide ‘contextual’ explanations.

Much research in the human sciences expresses general shifts, like the increased stress on ‘reflexivity’ and ‘the linguistic turn,’ in intellectual culture. The former term has multiple uses, but in the most general sense it denotes the activity of reflecting—‘reflecting’ in the double sense of making representations of existence and re-creating existence through the process. The most obvious instance is language: language both represents and constitutes the world as it is in human activity. Argument about the impossibility of going ‘beyond’ language to the world, society, or human nature encouraged research focused on language (or other signs, understood as forming a language) as the content of the human sciences. This is ‘the linguistic turn.’ What such claims and shifts imply is subject to intense, and often esoteric, debate. But it is an implication that a field built around such debate is open-ended in its scope and prospects.

Influential discussions derive from Foucault. Foucault, at one stage, claimed that the human

sciences are part of a modern ‘episteme,’ or framework establishing the conditions of truth. This, he held, came into existence around 1800 through practices creating reflection on human beings as subjects, especially in educational and reformatory institutions. He described reflexivity as a special condition of the modern age, and he contrasted this with the manner in which knowledge became possible through language, classification, and political economy in the earlier ‘Classical’ age. While many scholars controvert the specifics of Foucault’s disparate statements, his general exposure of the conditions of knowledge remains significant. His search for the sources of the human sciences in practical arrangements for discipline and government, his exploration of knowledge as power, and his use of discourse bridging the social sciences and humanities, are also conspicuous features of modern research.

Thus, the human sciences exist as an unstable and variable cluster of social science and humanities disciplines and topics, sometimes grouped merely for convenience, but also grouped because shared questions about knowledge and values are at stake. How usage of the term will develop is unclear. But the future will take account of the way the ‘historical turn’ has expanded and rethought historical knowledge about the sources and content of representations of human activity. It will continue to work through the implications of ‘the linguistic turn’ and of representation itself as human activity.

The human sciences add up to a critical reflection on the constitution of disciplines, and indeed of science, in the century and a half since the modern system of legitimating knowledge became established in universities. At the beginning of the twenty-first century, it would be surprising if past prescriptions remained unquestioned. New social forces, such as appear in feminist and postcolonial perspectives, have considerable impact. Different views of how the field is constituted obtain—from an extreme opinion that the human sciences are a branch of biology to the opposite extreme that ignores biology in toto. Within the social sciences, a conception of theory as the unifying and causally explanatory framework for knowledge is diminishing in appeal. Rather, scholars working under the rubric of the human sciences offer empirically researched accounts of matters that, through much of twentieth-century psychology and social science, were taken for granted. If earlier social scientists thought it relatively unproblematic to develop knowledge on the basis of existing categories and concepts, the human sciences open new perspectives by making that unproblematic quality itself the subject of research.

See also: Anthropology, History of; Demography, History of; Developmental Sciences, History of; Economics, History of; History and the Social Sciences; History of Science; History of Science: Constructivist Perspectives; History: Overview; Psychiatry, History

of; Psychology: Historical and Cultural Perspectives; Scientific Disciplines, History of; Sociology, History of

Bibliography

- Blanckaert C, Blondiaux L, Loty L, Renneville M, Richard N (eds.) 1999 *L'histoire des sciences de l'homme: Trajectoire, enjeux et questions vives*. L'Harmattan, Paris
- Danziger K 1990 *Constructing the Subject: Historical Origins of Psychological Research*. Cambridge University Press, Cambridge, UK
- Felt U, Nowotny H, Taschwer K 1995 *Wissenschaftsforschung: eine Einführung*. Campus, Frankfurt, Germany
- Foucault, M 1970 *The Order of Things: An Archeology of the Human Sciences*. Random House, New York (French edition 1966)
- Fox C, Porter R, Wokler R (eds.) 1995 *Inventing Human Science: Eighteenth-century Domains*. University of California Press, Berkeley, CA
- Hacking I 1995 *Rewriting the Soul: Multiple Personality and the Science of Memory*. Princeton University Press, Princeton, NJ
- Manicas P T 1987 *A History and Philosophy of the Social Sciences*. Basil Blackwell, Oxford and New York
- Nelson J S, Megill A, McCloskey D N (eds.) 1987 *The Rhetoric of the Human Sciences: Language and Argument in Scholarship in Public Affairs*. University of Wisconsin Press, Madison, WI
- Porter T M, Ross D (eds.) 2002 *The Cambridge History of Sciences, Volume 7: Modern Social Science*. Cambridge University Press, Cambridge, UK
- Ricoeur P 1981 *Hermeneutics and the Human Sciences: Essays on Language, Action and Interpretation*. Cambridge University Press, Cambridge, UK
- Smith, R 1997 *The Fontana History of the Human Sciences*. Fontana, London; same as *The Norton History of the Human Sciences*. Norton, New York
- Taylor C 1985 Interpretation and the sciences of man. In: *Philosophy and the Human Sciences: Philosophical Papers*. Cambridge University Press, Cambridge, Vol. 2, pp. 15-57 (first published 1971)

R. Smith

Human-Computer Interaction

The increasing contribution of other disciplines is making human-computer interaction focus less on the individual, with the consequence that a great deal can be gained from other fields within the social sciences, and perhaps, provide more insight into human behavior in return.

Human-computer interaction, more widely known by its acronym HCI, is primarily the study of how people use the complex technological artifacts that now pervade their lives. A secondary, but arguably more important focus, is on how these artifacts can be designed to facilitate this use. As an applied discipline, HCI encompasses theoretical and academic re-

searchers from psychological and computer science backgrounds, who may be interested in the area as a domain within which to apply and evaluate their theories of human or system behavior, as well as designers and engineers from industrial and commercial backgrounds, for whom the pay-off is a saleable product, built faster and more economically than their competitors'. In recent years, with the rapid technological development of computing devices and their increasing ubiquity within telecommunications, the field has expanded to include social and organizational researchers, who examine the impact that the technologies have upon human-human interaction. This itself feeds back to the identification of novel applications and the development of further devices. HCI is now an active, multifaceted, and methodologically diverse component of the domain that it studies, both following and directing technical innovation.

1. Origins in Human Factors and Cognitive Psychology

The roots of HCI predate the invention of computers, and the term itself has only been used since the 1970s. At their simplest, computers are just machines, built to help people carry out tasks, and the dominant paradigm for studying HCI has been that of ergonomics or human factors, with study of the *interaction* between the human and computer subjugated to that of the human's *use* of the computer. An earlier name for the field was man-machine interaction, and within the field, the words person and human are still generally replaced by the term 'user.' In the same way that early time and motion researchers would analyze the tasks that a factory or office worker performed, breaking their job down into its constituent elements, measuring the time each part took and assessing their relationships to each other to identify any 'wasted,' redundant, or unnecessary actions, so too have HCI researchers attempted to decompose the user's task, and to identify task steps that are difficult to perform, hard to learn, or confusing.

This emphasis on task analytic decomposition reflects the early predominance within HCI of cognitive psychologists, who brought with them information processing models of human thought. These models were themselves based on a metaphor of the mind as a computer: one that received sensory input, processed it with access to memory and inferential resources, and produced motor output to act upon the world. By far the most influential of these within HCI has been the GOMS model (Card et al. 1983), which stands for Goals, Operators, Methods, and Selection. Based upon an intentionally simplistic conception of a model human processor (MHP), a GOMS analysis would identify the goals that the user has to reach, and the operators or mental actions available to the user to

carry out the task of reaching the goal. Frequently used series of operators become chunked as easily accessible 'methods' that are linked to a subgoal or specific part of the task, and selection routines enable the user to 'decide' which alternative methods or operators to apply if there is more than one possible route to the goal. The MHP provides estimates of the length of time that executing each basic sensory, cognitive, and motor operator takes (these values being empirically derived from cognitive psychology experiments), and so once an accurate GOMS analysis of the task has been produced, it is possible to sum the durations of each component to produce a prediction of the time taken for the whole task to be performed. Subsequent elaboration of the approach to allow the perceptual, cognitive, and motor parts of the architecture to operate in parallel, and to model the processing of concurrent activities using critical path analysis, has had notable success in modeling the efficiency of telephone toll and assistance workstations (Gray et al. 1993).

Despite successes such as this, a limitation on the application of GOMS in practice is that it is intended to model only the optimal performance of a skilled user in ideal circumstances, and so it is best suited to industrial-style settings where a particular task may be executed hundreds or thousands of times a day. An interface design that is a second or two quicker to use in these situations will thus result in a large saving in time, and presumably money, for the organization that is paying for it. GOMS is unable to say anything about another crucial aspect of interface use, that is, how long it might take someone to become an expert, how many errors a nonexpert might make, or how pleasant or unpleasant the experience of executing the operations might be. Indeed, the critical assumption within GOMS that operators can be chunked as methods is not actually addressed by the model itself. It remains the job of the analyst to recognize which sequences of operators can be represented as methods, and how finely these methods need to be related hierarchically. Recognizing that this development of knowledge about the task is the key to learning about an interface, Newell (1990) proposed what he termed a unified cognitive architecture specifically to allow problem solving and learning to be modeled as a process of goal decomposition, such that goals are iteratively broken down into subgoals. By specifying the operation of this architecture within a 'programming language' called Soar, Newell provided a technique for predicting the way that users would integrate their postulated initial knowledge with information derived from the computer interface through their interaction with it. In effect, a Soar analyst writes a program in the form of 'If, ... Then ...' rules, and this is then 'run' to simulate a user's learning behavior.

The GOMS and Soar research communities have had a great deal of influence within the cognitive psychology side of HCI, but as simulation models,

their attention to detailed aspects of the interface is both their strength and weakness. While it allows them to identify small but critical aspects of a design as crucial problems for a user, they both require complete specifications of an interface, and for Soar, the task context and information available to a user. They therefore come into their own for the contrastive evaluation of extant alternatives, but are hard to apply systematically at earlier stages of design, when the question is 'how?' rather than 'which?' The same is arguably true of newer, more detailed techniques such as LICAI, which is based on Kintch's construction-integration model of goal formation in text comprehension (Kitajima and Polson 1997), and EPIC (Kieras and Meyer 1997). EPIC builds on the model human processor by incorporating many theoretical concepts such as the distinction between declarative and procedural knowledge drawn from Anderson's ACT class of models, and limited capacity visual and auditory working memories coordinated by executive processes. However, as Kieras and Meyer themselves note, 'EPIC was not developed primarily for addressing HCI problems but is a larger scientific endeavor to represent important theoretical concepts of human intelligence or abilities' (Kieras and Meyer 1997, p. 394).

2. *Development of Usability Evaluation Methods*

The apparent concentration of cognitive psychologists on the development of cognitive architectures that would ideally explain all of human behavior, as well as the human use of computers, can justifiably be seen as a problem from the point of view of their colleagues who are involved in the day-to-day design of computer interfaces. It was accepted fairly quickly that full GOMS models were too laborious to be practical, and a simpler (but inevitably less reliable) 'keystroke level model' was developed that avoided the need for a full task analysis. Soar models also require a programming and analysis commitment that can seem disproportionate in the context of design, especially when an apparently small modification in the design requires the whole model to be reprogrammed and rerun. It can also be difficult to identify reasons for the simulated behavior that results from an analysis. These cognitive architectures, then, may prove to be more useful in informing their core disciplines than in practical application.

In reaction to these perceptions of too much cost for too little benefit, a number of 'cut down' usability evaluation methods have been developed. All are to some extent derived from well-founded and principled psychological theories, but by encapsulating key components of the modeling techniques within a process of some form, seek to minimize the overhead costs of learning the theory itself, and to provide a positive cost-benefit trade-off by reducing the need for user

trials to be conducted with early prototypes or mock-ups of interfaces (Nielsen and Mack 1994). Cognitive walkthroughs, for example (Polson et al. 1992), provide a checklist by which aspects of the interface can be compared with the user's notional task, step by step, and screen by screen. The steps of the checklist are in fact derived from cognitive complexity theory, and the method focuses on the semantic appropriateness of the verbal labels or commands used throughout the interface, although the practitioners applying it do not need to know this.

Other approaches seek to avoid any role for theory at all, and base themselves in the codification and reuse of practical experience. From the earliest days of HCI, sets of guidelines have been compiled, both within design teams to provide a 'house style' and in a wider sense as a shared set of 'best practices,' so that designers encountering a problem for the first time have access to the solutions found or attempted by others. Some, such as those presented by Gardiner and Christie (1987), have an empirical basis, and specify specific design features such as acceptable screen color combinations and choices for typefaces. Others are less formal, and direct the designer towards desirable attributes or practices such as aiming for consistency, compatibility and coherence. The rapidly evolving nature of the domain conspires against such stored knowledge, however, for anything designed before the mid-1990s is almost certain to be based on expired technology, whose constraints no longer apply. Concrete prescriptions about design solutions become outdated too quickly for their promulgation, while the more abstract (such as 'be consistent') have proven too abstract to be of much real use.

3. *HCI and the Design Process*

This focus upon evaluation reveals a fundamental problem in early notions of the relationship between design and modeling. An implicit assumption of many HCI approaches was that design could be thought of as a sequential or 'waterfall' process, with discrete stages progressing from a requirements analysis, through specification layers, to implementation, with a final evaluation of the finished product. By locating evaluation at this final stage, HCI staff found themselves limited to what became known as a 'fire-fighting' role, being asked to repair problematic designs at a point when any redesign would result in a product falling behind schedule, further worsening the cost implications of their involvement. Organizationally, it is very difficult to convince people that the design they have been committed to for months has some basic usability flaws.

As Landauer (1995) has pointed out in his economic analysis of the computer industry, design flaws have to be identified much earlier in design if their identification is to be helpful and cost effective to repair. Fortunately, few designs actually proceed in

such a strictly formal manner, with a more serendipitous, parallel, and iterative process being more typical. Such cyclical processes allow HCI contributions to be made at more appropriate junctures, and have led to an emphasis upon user-centered design (e.g., Norman and Draper 1986), in which user testing becomes a central aspect of design throughout the process, design itself is directed towards prototyping, and designers are encouraged to 'plan to throw one away,' following Brooks' seminal advice (Brooks 1975).

A very different approach to providing support to designers in these less structured situations comes in the form of a notational decision support tool called QOC, or questions, options, and criteria (MacLean et al. 1991). Aimed at recording the 'design rationale,' QOC represents the design space through which a product has traveled, listing the questions that the design team were trying to answer at various points, the options that were considered, and the criteria that they used to choose one option over another. By making decisions and trade-offs explicit, and recording the reasons that were cited for making them, QOC provides a mechanism both for managing the decision process and for recording it for later review. It has been found to be particularly helpful in providing background information for people who were not part of the decision at the time, either because they only became involved at a later stage or because they were working on some other part of a large project. QOC is important in that it represents a wider concern with design as a process, rather than simply with evaluating its final products.

Involvement in design at a point before the final product has been created means, inevitably, that HCI methods need to provide advice to designers rather than evaluating finished designs. While 'cut-down' usability inspection methods and heuristic evaluations can be conducted on partial designs and mock-ups, it has been argued that their pragmatic separation from theory weakens their value. While they may well identify problems in a design, the fact that they are not associated with an explanatory theoretical underpinning means that the causes of the problems can remain unexplained. The benefit to the designer of a rapid identification of problems is accompanied by the cost of working out why the problem has occurred and what is to be done about it. May and Barnard (1995) recommended that HCI researchers should work towards providing 'supportive evaluation methods' that assist the designers by accounting for usability problems in terms of some theoretical concepts that are helpful in suggesting redesigns.

They referred to a distinction made by Long (1987) between the 'discovery representation' of a scientific theory and its 'application representations.' The discovery representation allows academics and researchers to explore phenomena, such as typical error patterns in HCI, or the way that users learn about interfaces

and solve problems. The application representations, on the other hand, are encapsulations of the theory specifically targeted towards practical application. These may take several different forms, depending upon the intended practitioners and the expected purpose. Unlike cut-down methods, the application representations retain a direct link to the parent theory and, in order to use them, information has to be expressed in theoretically important ways. The advantage for practitioners is that the results of their use is also couched in theoretically relevant forms, and so is linked to an explanatory body of ideas that can be recruited to assist in redesign.

May and Barnard (1995) present the example of a cognitive architecture called interacting cognitive subsystems, or ICS. As a discovery representation, it is unlike GOMS, Soar or EPIC in that it does not seek to simulate cognitive processing, but provides a framework for identifying bottlenecks in the flow of information through a series of cognitive subsystems, each of which is specialized for processing a different level of mental representation (see *Cognitive Science: Overview*). Like the simulation models, these levels include sensory, cognitive, and motor representations. The modeling process involves describing the likely pattern of exchanges between the various levels, the degree of skill-based behavior of relevant aspects of a task, the involvement of the user's recent and longer-term memories, and, crucially, the structure of the user's conceptual task or of perceptual representations of the interface. In this discovery form, the modeling is too complex to be carried out by anyone other than an expert in the theory itself, and the results are couched in terms of the processes and their configuration, rather than specific design consequences.

To support the use of ICS within design, two alternative application representations have been developed: one is a production rule-based 'expert system,' which partly replicates the role of the ICS theorist in constructing a principled model of an interaction; the other is a pencil-and-paper technique for reasoning about the changes over time in the structure of the mental representations encoded at the various ICS subsystems (Barnard and May 1999). Both of these application representations require some familiarity with the core concepts of ICS, but neither requires practitioners to construct a full ICS model themselves. Their value lies in providing problem-solving tools for designers to apply to partially specified or speculative ideas about interfaces, before any commitment has been made.

4. Influx of Novel Approaches

In moving away from a simulation of a user interacting with a computer, ICS is part of a trend away from the concentration upon a dyadic, turn-taking relationship between a single user and a single computer. Classical HCI models can be caricatured as having a human

issuing commands to a computer, waiting for it to carry them out, and then inspecting the screen while the computer patiently waits for them to form some new goals. While this might have been an appropriate model for HCI in the early 1980s, the advent of distributed computing architectures and the massive interconnection of computing systems over the Internet means that it is inadequate. Even modern personal computers will be carrying out multiple tasks, doing background processing almost in-between each keystroke. The humans are not always isolated users pursuing solitary goals, either, but are using computers as part of their everyday activities, often in social or work settings that involve interacting with other humans with the support of their computers. An analysis of interaction that does not allow for such factors has inevitable flaws.

These developments have led to a new involvement within HCI of social and organizational psychologists, as well as researchers from other social sciences such as ethnomethodology, anthropology, and sociology (Dourish and Button 1998). This has led to an influx of novel theoretical ideas, in particular Vygotsky's activity theory (e.g., Nardi 1995) and Suchman's situated action (Suchman 1987). These ideas re-emphasize the relevance of the context within which human actions are formed, and as Carroll (1997) noted, have brought a continental European perspective into what had previously been a largely North American and British endeavor. Instead of the user and their system, activity theory deals with an 'activity system' composed of social aspects such as attitudes, beliefs, and experiences as well as technological factors.

Another development which moves away from the traditional HCI dyad has come from a computer science perspective, where the idea of a system as a single entity has long been outdated. With the introduction of 'object oriented' ideas, computer scientists have moved towards seeing complex systems as built up from a large number of quasi-independent 'interactors' that exchange information and tasks, with no constraint for these interactors to be part of a single machine. A large body of theory has evolved to model such distributed and concurrent activity, and a logical step has been to see a user's cognitive processes as just another set of interactors. The unit of analysis for such 'syndetic models' (Duke et al. 1999) is the overall system composed of human and computer interactors, and their conjoint activity can be modeled using mathematical formalisms to prove literally that certain interactions can or cannot occur, or that certain assumptions must be met for them to be possible.

5. Prospects for HCI as a Discipline

Both the social and the computational contributions to HCI currently fall into the discovery mode of analysis, in Dowell and Long's terms, rather than the

application mode. It is not yet clear whether these novel directions will crystallize into methods or representations that are directly applicable by design practitioners, or whether they will serve more to direct the commercial development of products away from the classical PC in an office towards other niches. They may also represent the evolution of HCI as a discipline in its own right, with its own methods and theories, rather than as an interdisciplinary area colonized by theories developed in cognitive psychology and artificial intelligence. The tension between academic theorists and practical developers will undoubtedly continue to be a major dynamic upon the field, with its predominance of conference papers enabling non-empirical presentations of novel technologies or design tools to share the limelight with hard-nosed experimental evaluations. In recent years, the increasing theoretical quality of the few specialized HCI journals may provide justification for those who have argued for universities to establish HCI departments in their own right, with their own undergraduate courses. On the other hand, the progressive involvement of computing artifacts within our everyday lives is making their use an ordinary part of human behavior. The increasing contribution of other disciplines is making HCI focus less on the individual, with the consequence that a great deal can be gained from other fields within the social sciences and, perhaps, provide more insight into human behavior in return.

See also: Cognitive Theory; SOAR; Computers and Society; Human-Computer Interface; Science and Technology, Social Study of; Computers and Information Technology; User Interface Design

Bibliography

- Barnard P J, May J 1999 Representing cognitive activity in complex tasks. *HCI* **14**: 93–158
- Brooks F P 1975 *The Mythical Man Month; Essays on Software Engineering*. Addison-Wesley, Reading, MA
- Card S K, Moran T P, Newell A 1983 *The Psychology of Human-Computer Interaction*. Erlbaum, Hillsdale, NJ
- Carroll J M 1997 Human-computer interaction: psychology as a science of design. *International Journal of Human Computer Studies* **46**: 501–22
- Dourish P, Button G 1998 On 'technomethodology': foundational relationships between ethnomethodology and system design. *HCI* **13**: 395–432
- Long J B 1987 Cognitive engineering and HCI. In: Warr P (ed.) *Psychology at Work*. Pelican, Harmondsworth, UK, pp. 73–93
- Duke D J, Barnard P J, Duce D A, May J 1999 Syndetic modeling. *HCI* **13**: 337–93
- Gardiner M M, Christie B (eds.) 1987 *Applying Cognitive Psychology to User-Interface Designs*. Wiley, Chichester
- Gray W D, John B E, Atwood M E 1993 Project Ernestine: validating a GOMS analysis for predicting and explaining real-world task performance. *HCI* **8**: 237–309
- Kieras D E, Meyer D E 1997 An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *HCI* **12**: 391–438
- Kitajima M, Polson P G 1997 A comprehension-based model of exploration. *HCI* **12**: 345–89
- Landauer T 1995 *The Trouble with Computers*. MIT Press, Cambridge, MA
- MacLean A, Young R, Bellotti V, Moran T 1991 Questions, options and criteria: elements of design space analysis. *HCI* **6**: 201–50
- May J, Barnard P J 1995 Towards supportive evaluation during design. *Interacting with Computers* **7**: 115–43
- Nardi B A (ed.) 1995 *Context and Consciousness: Activity Theory and Human-Computer Interaction*. MIT Press, Cambridge, MA
- Newell A 1990 *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA
- Nielsen J, Mack R L (eds.) 1994 *Usability Inspection Methods*. Wiley, New York
- Norman D A, Draper S W 1986 *User Centered System Design*. Erlbaum, Hillsdale, NJ
- Polson P G, Lewis C H, Rieman J, Wharton C 1992 Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *International Journal of Man Machine Studies* **36**: 741–73
- Suchman L A 1987 *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press, New York

J. May

Human-Computer Interface

1. Definitions: The Interface as a Carrier of Epistemologies

The computer/human interface refers to the modalities through which people interact with computational technologies. These include different data input devices: you can give the computer information through punched cards or through typing numbers on to a screen. And these include different ways of manipulating software: one can type symbols at a 'command line' or can 'navigate' a data field that appears as a three dimensional space. Specialized applications—software designed for the physically challenged, software designed to simulate airplane flight—have increased the number of available interface technologies. Computer users may employ joysticks, headsticks, trackballs, voice recognition, and touch sensitive screens to communicate with the software that governs the machine's behavior.

The breadth of the topic of computer/human interfaces and how to improve them for usability and ease is too vast to review comprehensively here (Norman 1998, Raskin 2000). This essay examines the most common shifts in interface use over the last half century up to 2000, and centers on the social and psychological implications of interface experience rather than its technical details. Interface design has not only changed the instrumental computer: what the

computer can do for us; it has changed the subjective computer: what the computer does to us, to our ways of relating to others, thinking about ourselves, seeing the world. As the dominant knowledge technology of our era, computational machines offer us images of what it means to know (Penrose 1989, Dennett 1991, Turkle 1995).

Over the last half century to 2000 the trend in interfaces has been from favoring a style of interaction in which the computer is approached as a mechanism towards interactions in which the computer is approached as a behaving and aware organism. Otherwise put, the trend has been from people acting directly on the machine to interacting indirectly with its increasingly elaborate 'self presentation' in which many layers of programming stand between the underlying mechanism and what is presented to the user. The movement has been from interface as an environment for commands (Weinberg 1971) to interface as environment for conversation or theater (Laurel 1990). Whether the computational object presents an interface that demands small or large bodily gestures or particular words or sounds, this direction (increased layering; increased opacity) has had significant psychological and epistemological implications. The computer has moved from being the primary technological model of a culture of calculation towards modeling a way of knowing in a culture of simulation. In a culture of simulation, users are encouraged to take computers 'at interface value.'

1.1 The Recent History of Personal Computing Interfaces

The first personal computers of the 1970s and early 1980s, like the mainframes and minicomputers, required users to know how to issue exact instructions. Being comfortable with their command-line required users to put themselves into the frame of mind of those who had programmed the computer, who had given it its fundamental rules. Those who were most comfortable with this kind of information interface were those who knew something about programming and its intellectual aesthetic (Schneiderman 1987).

Through the late 1970s, the computer interface presented the machine as a giant calculator. The interface carried the message that no matter how complicated a computer might seem, what happened inside it could be mechanically unpacked. The computer interface was a set of command lines that would get you close to those rules that lay beneath the surface of the machine.

The presentation of the machine through its interface captured the essence of the computer in a culture of calculation. This essence corresponded to a modernist or enlightenment model of what it means to understand: it means to have knowledge of what

makes the 'clockwork' tick, and the programmed computer was conceived of very much as mechanical clockwork. When computer users talked about 'transparency' what they meant is that the interface was a window for looking beyond the 'magic' of the machine's extraordinary powers to the mechanism that made them possible.

What was true of the language that surrounded the large mainframes and mini-computers was even more true of the language that surrounded the first personal computers. These machines, first appearing in the late 1970s and early 1980s, required users to know how to issue exact instructions and they were encouraged to optimize their computers' performance by learning assembly language, by going deeper and deeper into the lower levels of a machine's structural organization. By the mid-1980s, increased processing power made it possible to build graphical user interfaces, commonly known by the acronym GUI, which hid the 'bare machine' from its user. The new opaque interfaces—the first popular one on the mass market was the 1984 Macintosh—represented more than a technical change. The Macintosh 'desktop' introduced a way of thinking about the computer that put a premium on the manipulation of a surface simulation. The desktop's interactive objects, its dialogue boxes in which the computer 'spoke' to its user, pointed toward new kinds of experience in which people did not so much command machines as enter into conversations with them (Apple 1985). In personal relationships, we often interact without understanding what is going on within the other person; similarly, when we take things at (inter)face value in the culture of simulation, if a system performs for us, it has all the reality it needs.

In 1980, most computer users who spoke of transparency were referring to a transparency analogous to that of traditional machines, an ability to 'open the hood' and poke around. But when users of the Macintosh talked about its transparency, they were talking about seeing their documents and programs represented by attractive and easy-to-interpret icons. They were referring to an ability to make things work without needing to go below the screen surface. Today, the word 'transparency' has taken on its Macintosh meaning in both computer talk and colloquial language. In a culture of simulation, resonant with a post-enlightenment or postmodern aesthetic, when people say that something is transparent, they mean that they can see how to make it work, not that they know how it works.

2. Simulation and Its Discontents

Suppose you receive this message on a computer screen when playing a simulation computer game: 'Your orgot is being eaten up.' (Turtle 1995). You do not need to know what an orgot is to play the game

although you can infer from the context of the game that it is a kind of organism. Players interact with the interface of the computer game in the same way they interact with the interface of a video game. In computer games, as in video games, one learns from the process of play. You do not first read a rulebook or get your terms straight. The game allows you to manipulate its variables directly; it enables you to avoid commands, but experiment with trial balloons. You are able to act on an intuitive sense of what will work without understanding the rules that underlie the game's behavior. Your game will teach, for example, that you that you must give your 'trilobytes' algae or they will become extinct. Thus, you can keep playing even without any clear idea of what is driving events. Clarity and a grasp of the rules are not essential to the process.

People use their contact with objects to keep in touch with their times, to discover 'what's important.' They use objects to work through powerful cultural images, to help arrange these images into new and clearer patterns. From this point of view, the opaque interfaces of simulation games and simulation tools help to help us think through current challenges to traditional epistemologies where the manifest refers back to the latent, the signifier to the signified. The personal computer culture began with small machines that captured a utopian modernist vision of transparent understanding, but the scintillating, pulsating surfaces of simulation computer games are dynamic, seductive, and elusive, a world without depth, a world of surface. *Players can stay on the surface, taking things at (inter)face value.* The computer interface serves as a carrier object which, in a significant sense, brings postmodernism down to earth.

The culture of simulation presented by the computational interface is our everyday experience of the postmodernist aesthetic. And since it is common to experience the computer as a 'mind machine' this aesthetic influences computational models of mind. For example, 15 years ago the computational models of mind that dominated academic psychology were modernist in spirit: nearly all tried to describe the mind in terms of centralized structures and programmed rules. In contrast, today's models often embrace a postmodern aesthetic of complexity and decentering. Mainstream computer researchers no longer aspire to program intelligence into computers but expect intelligence to emerge from the interactions of small sub-programs. If these emergent simulations are 'opaque,' that is, too complex to be completely analyzed, this is not necessarily a problem. After all, these theorists say, our brains are opaque to us, but this has never prevented them from functioning perfectly well as minds.

Beyond this, the development of the windows metaphor for computer interfaces was a technical innovation motivated by the desire to get people working more efficiently by 'cycling through' different

applications much as time-sharing computers cycle through the computing needs of different people (Microsoft 1995). In practice, however, windows have also become a potent metaphor for thinking about the self as a multiple, distributed, 'time-sharing' system. The self is no longer simply playing different roles in different settings, something that people experience when, for example, one wakes up as a lover, makes breakfast as a mother, and drives to work as a lawyer. The life practice of windows is of a distributed self that exists in many worlds and plays many roles at the same time.

2.1 Conversation, Cuddles, and Creatures

Computers have been thought of as 'artificial brains' for generations. What is new in the current analogies of computers to the brains of behaving systems is the development of interfaces through which we can interact with computers as though they were creatures, either human or animal. An example is the conversational interface used to communicate with software programs known as 'agents' or 'bots.' The computer users types out their side of a conversation in colloquial language; the software agent responds more or less in kind. Of course, such conversations work best if the domain of discussion is highly proscribed, but the experience of such interfaces is quite dramatically that of interacting with an 'other.' Computer agents that literally present 'faces' as their interfaces dramatically increase the degree of anthropomorphism of the computational object. Voice recognition software enhances the conversational effect and the experience of the machine as a being.

Computational creatures are being designed as toys for children. Today, children are asked to interact with these creatures through gesture and touch; in the near future, the interactions will be through language and tone. In the case of the Furby, a furry, cuddly, robotic toy that seems to learn and love, the computational interface presents a creature successfully 'learning' language and being nourished by food; the child is able to respond with gestures expressing affection. In the case of another computational creature, Sony's pet dog robot, AIBO, the computational interface presents a 'trainable' dog whose personality can be shaped by the quality of the user's ministrations, communicated by speech and sound. In the case of My Real Baby, a robotic doll, the interface enables the doll to speak increasingly complex language and show emotional expression through facial musculature while affording the user the possibility of showing nurturance through feeding, caressing, diaper changing, burping, and similarly anthropomorphic 'creature comforts.' Here, perhaps most dramatically, computational design has created an interface that through its emulation of the human explicitly diverts attention from the underlying mechanism and focuses it ex-

clusively on the meaning of the 'baby's' expressions, sounds, and words. Almost no children express curiosity about how the object 'works' in the traditional sense of understanding anything about its 'innards.' They accept its behavior as that of a creature, a creature that while not human, is frequently categorized as 'sort of alive.'

3. Future Trends

Traditional approaches to studying user interfaces have focused on usability, human factors (including level of stress), and psychological studies that have examined sensorimotor adaptation, problems of information overload, ergonomics, and the resonance of interfaces with varying intellectual styles. The projected future of computation and its interfaces suggest that philosophical and even theological concerns might well be added to this mix since the interface is promised to become not only more transparent but more intimately engaged and ultimately fused with its users.

With the computer/human interfaces that researchers project in the near-future the user will experience the interface as a creature (as in increasingly advanced robots) as prosthetics (a pair of glasses that also serve as a computer screen for mobile net browsing), a piece of clothing (with embedded processors that for example use the temperature of the human body as the input to warm or cool the skin), or as a virtual world that people can fully immerse in and navigate by moving their bodies as one would move in the real physical world. Finally, there are proposals for the interface to become part of the self. In this scenario the interface becomes a digital sensor embedded in human flesh. Creatures, prosthetics, immersive environments, and in the end, the interface disappears and people reappear as cyborg.

See also: Cognitive Modeling; Research Logic in Cognitive Science; Cognitive Science: Philosophical Aspects; Computational Imagery; Computers and Society; Discrete State Models of Information Processing; International Migration by Ethnic South and Southeast Asians; Science and Technology, Social Study of; Computers and Information Technology; User Interface Design; Virtual Reality and Spatial Cognition

Bibliography

- Apple Computer 1985 *Inside Macintosh*. Apple Computer, Cupertino, CA, Vol. 1
Dennett D 1991 *Consciousness Explained*. Little, Brown, Boston, MA
Laurel B 1990 *The Art of Human-Computer Interface Design*. Addison-Wesley, Boston, MA
Microsoft 1995 *The Windows Interface Guidelines for Software Design*. Microsoft Press, Redmond

- Norman D 1998 *The Invisible Computer*. MIT Press, Cambridge, MA
Penrose R 1989 *The Emperor's New Mind*. Oxford University Press, London
Raskin J 2000 *The Humane Interface: New Directions for Designing Interactive Systems*. Addison-Wesley, Boston
Schneiderman B 1987 *Designing the User Interface*. Addison-Wesley, Boston
Turkle S 1995 *Life on the Screen: Identity in the Age of the Internet*. Basic Books, New York
Weinberg, G 1971 *The Psychology of Computer Programming*. Van Nostrand Reinhold, New York

S. Turkle

Human-Environment Relationship: Carrying Capacity

Carrying capacity is the maximum population that a specified environment can support indefinitely without degradation of that environment (Ellen 1982, p. 41). Ecological carrying capacity links population of biota to a given habitat in which biophysical elements of the habitat determine the maximum population. In human-environment relationships, however, carrying capacity involves the added dimension of technology and management strategies; as these human-based elements change, so does the carrying capacity (Brush 1975 and Street 1969). Thus, humankind maintains the ability to expand and reduce carrying capacity, and vastly different assessments of the actual trends underway follow from the spatial scale of analysis (i.e., landscape, region or biosphere) and perspective (i.e., environment as resource commodities or as environmental services) taken (cf. Ehrlich et al. 1977 and Simon and Kahn 1984).

Contemporary uses of carrying capacity suggest that the concept was borrowed from the ecological sciences by the human sciences. In reality, however, the ecological sciences took a cue from the ideas of Robert Malthus and subsequent developments of Malthusian principles about human population-resource relationships. The carrying capacity idea can be found in various kinds of works, even though that label was or is not applied. For example, the geographer-geomorphologist John Wesley Powell used the concept in his assessment of the Homestead Act of 1862 (USA), which allowed people to claim 160 acres of public land for farming. He warned that 160 acres were insufficient in most cases because the land in question, that west of the 100th meridian, was arid and much of it lacked access to water. He maintained that these lands, absent the potential for irrigation, had value primarily for livestock and thus 'a farm unit should not be less than 2560 acres' in the arid West of the USA (Darrah 1951, p. 233).

Such examples notwithstanding, the nineteenth century development of the carrying capacity concept

in the human science is often credited to geographical and anthropological research among indigenous peoples in the less developed world, who typically maintained a fairly direct relationship between subsistence and the immediate biophysical environment (Brush 1975 and Street 1969). The British geographer William Allan (1965), for example, used the concept of carrying capacity to evaluate the Native Reserve resettlement process in southern Africa. The essential data required for the calculation were: '(a) a soil or land map based on a suitable form of land classification; (b) an estimate, for each of the land classes employed, of the proportion of land which is practicable for cultivation under the prevailing system of agriculture; (c) a survey of the customary land-usage of the people concerned, with particular reference to the duration of cultivation and subsequent rest on the soils of each land class; and (d) an average for the acreage under cultivation per head of population at any one time' (Allan 1965, p. 8). He found that the carrying capacity of the Native Reserves, registered as persons per unit area, was exceeded at their founding, based on the customary land uses of the inhabitants. Other versions of carrying capacity assessments have been called for and employed for planned occupation throughout the tropical world, including the colonization of the Amazon basin (Fearnside 1986).

As noted, technology and land resource–use strategies play an essential role in the use of carrying capacity. As these change, so does the capacity assessment. The 'Green Revolution' is a case in point. This massive change and transfer of technology and land-management strategies greatly expanded the production and hence 'carrying capacity' of many areas throughout the world by making possible the intensification of agriculture through hybrid seeds, fertilizers, pesticides, herbicides and irrigation schemes. Yet, this revolution and its land-use impacts also expanded the reach of environmental drawdown, from increased downstream flooding and sedimentation to large-scale releases of trace gases to the atmosphere, raising serious questions about the spatial and hierarchical scales at which carrying capacity should be addressed. While earlier studies could concentrate on a village or small area as a well-defined population–resource entity, it is now recognized that, particularly in industrialized countries, such entities no longer conform to such well-defined boundaries. Technology permits a place-based population to substitute for its environment by importing resources and exporting waste. Thus, carrying capacity now involves an assessment of the worldwide 'ecological footprint' of a place-based population (Rees 1992), or the full extent of the use and environmental impacts of that population. Through this lens, what may appear to be a sustainable carrying capacity in an area, registered by arrested or slowed environmental degradation, or even its reversal, may be obtained by transferring elsewhere the environmental consequences of the

area's occupation and the population's consumption. This situation necessitates an examination of carrying capacity on a global as well as a local basis, as the recognition of the 'silent' services provided by environment, such as pollinators for crops, necessitates an examination that includes more than immediate population–resource relationships (Vitousek et al. 1997).

See also: Agricultural Change Theory; Ecological Economics; Ecology, Cultural; Environment and Development; Environmental and Resource Management; Environmental Economics; Environmental Policy; Environmental Policy: Protection and Regulation; Environmental Risk and Hazards; Feminist Political Ecology; Green Revolution; Greening of Technology and Ecotechnology; International and Transboundary Accords, Environmental; IPAT (Impact, Population, Affluence, and Technology); Planning Issues and Sustainable Development; Precautionary Principle; Sustainable Development

Bibliography

- Allan W 1965 *The African Husbandman*. Barnes and Noble, New York
- Brush S B 1975 The concept of carrying capacity for systems of shifting cultivation. *American Anthropologist* 77: 799–811
- Darrah W C 1951 *Powell of the Colorado*. Princeton University Press, Princeton, NJ
- Ellen R 1982 *Environment, Subsistence and System*. Cambridge University Press, Cambridge, UK
- Ehrlich P R, Ehrlich A H, Holdren J P 1977 *Ecoscience: Population, Resources, Environment*. W. H. Freeman, San Francisco, CA
- Fearnside P M 1986 *Human Carrying Capacity of the Brazilian Rain Forest*. Columbia University Press, New York
- Rees W E 1992 Ecological footprints and appropriated carrying capacity: what urban economics leaves out. *Environment and Urbanization* 4: 121–30
- Simon J L, Kahn H (eds.) 1984 *The Resourceful Earth: A Response to Global 2000*. Blackwell, Oxford, UK
- Street J 1969 An evaluation of the concept of carrying capacity. *Professional Geographer* 21: 104–7
- Vitousek P M, Mooney H A, Lubchenco J, Melillo J M 1997 Human domination of earth's ecosystems. *Science* 277: 494–99

M. E. Geores

Human–Environment Relationship: Comparative Case Studies

A comparative case study is a research approach to formulate or assess generalizations that extend across multiple cases. The fundamental intent of a compara-

tive case study—development of knowledge—applies equally in the natural and social sciences, with the latter incorporating qualitative analysis in a discursive mode as well as the quantitative and positivist modes of the sciences. Comparative case studies are particularly important to questions of human–environment interaction, combining case study and comparative traditions and thus extending the depth of the individual case study across several instances in time and/or space.

1. Case Study and Comparative Method

A case study is a rigorous, detailed, spatially bounded, temporally limited description and analysis of an event, process, place, or community intended by the researcher to exemplify a more general insight, principle, or theory. It ‘is an exploration of a “bounded system” or case (or multiple cases) over time through detailed, in-depth data collection involving multiple sources of information rich in context’ (Creswell 1997, p. 61). Case studies of particular communities or places are a common and classical mode of analysis in anthropology, human geography, and sociology.

Comparative analysis, in contrast, explores similarities revealed in different situations or cases sharing some common element(s) while differing in others. Comparative analyses are also common among social sciences, varying in the depth to which components are explicated, from statistical abstractions through vignettes to a level of detail common in individual case studies. Often, a case study has unique elements whose effects on the validity of generalizations may be ambiguous or uncertain. Thus case studies addressed through a conjunction of insights from alternative perspectives will reveal commonalties of particular interest. Comparison between cases can help to delineate such relationships and potentially to clarify ambiguity in a singular case.

1.1 Qualitative and Quantitative Aspects

Comparative analyses may proceed via quantitative analysis of numerical abstractions from cases, through detailed narrative description and qualitative analysis. There is no necessary relationship between qualitative and quantitative approaches in case studies and comparative analyses. Qualitative case studies are frequent in anthropology, some schools of sociology, and contemporary cultural geography. Case studies can also be internally as well as comparatively quantitative, particularly in human and economic geography. Whereas comparative studies are often qualitative, they may be quantitative when there are a large number of cases with a small number of variables, as in social survey research.

1.2 Explanatory Status of Comparative Case Studies

Both case studies and comparative analyses can be *inductive* or *deductive*. However, case studies and qualitative comparative analyses share a common explanatory limitation with models—they are fundamentally *heuristic* rather than *definitive* in that they cannot prove the validity of generalizations beyond providing persuasive evidence in particular contexts (Oreskes et al. 1994). Conversely, a case study illuminating and supporting a general principle does not preclude another case casting doubt. As Leach (1968, p. 344) stated about the comparative method in anthropology,

...comparison...is not, and cannot be, a disguised form of scientific experiment leading to an explanation.’ Moreover, neither misinterpretation of the case nor failure of the case to illustrate a generalization necessarily means that a hypothesis is not valid. However, a comparative case study approach strengthens the heuristic value of individual cases by bringing in-depth understanding in different circumstances, with greater confidence in the conclusions drawn from them. ‘The generality of the case study...is relative to the methodological value of the test process involved in the experiment itself and “holds until proved otherwise” (Hamel 1993, p. 40).

Multiple cases strengthen the confidence of the researcher that the generality is sound.

2. Comparative Case Study Strategies

Researchers investigating human–environment issues can avail themselves of many alternative strategies for developing comparative case studies. Strategies included here are certainly not exhaustive nor mutually exclusive, but they suggest ways in which comparative case studies can be structured creatively for investigation, presentation, and analysis. Often, the power of the comparative analysis is strongly related to the selected strategy. These strategies include research on the basis of community, place, perspective, structure, system, scale, transects, gradients, time, events, analogs, or theory as organizing principles. In practicality, the major organizing principle of a comparative case study often includes others that are subsidiary, but inherent in the nature of the cases selected. For example, selection of case communities along a transect incorporates both community and place perspectives as well, even though the transect is the organizing theme. Examples of each strategy are included in Table 1.

2.1 Community

Comparative case studies in disciplines such as sociology and anthropology often focus on studying

Table 1
Comparative case studies: a literature guide

Community	Bennett J W 1969 <i>Northern Plainsmen: Adaptive Strategy and Agrarian Life</i> . Aldine, Chicago Mortimore M 1999 <i>Working the Sahel: Environment and Society in Northern Nigeria</i> . Routledge, London
Place	Moran E F 1993 <i>Through Amazonian Eyes: The Human Ecology of Amazonian Populations</i> . University of Iowa Press, Iowa City, IA Turner B L, Hyden G, Kates R W 1993 <i>Population Growth and Agricultural Change in Africa</i> . University of Florida Press, Gainesville, FL
Perspective	Knight C G 1974 <i>Ecology and Change: Rural Modernization in an African Community</i> . Academic Press, New York Meinig D W 1979 <i>The Interpretation of Ordinary Landscapes: Geographical Essays</i> . Oxford University Press, New York
Structure	Goldschmidt W R 1966 <i>Comparative Functionalism: An Essay in Anthropological Theory</i> . University of California Press, Berkeley, CA Moran E F 1995 <i>The Comparative Analysis of Human Societies: Toward Common Standards for Data Collection and Reporting</i> . L Rienner, Boulder, CO Turner B L, Brush S B (eds.) 1987 <i>Comparative Farming Systems</i> . Guilford, New York
System–subsystem	DeWalt B R, Pelto P J (eds.) 1985 <i>Micro and Macro Levels of Analysis in Anthropology: Issues in Theory and Research</i> . Westview Press, Boulder, CO Mitchell J K 1974 <i>Community Response to Coastal Erosion: Individual and Collective Adjustments to Hazard on the Atlantic Shore</i> . University of Chicago, Department of Geography, Research Series no. 156
Scale	Wilbanks T J, Kates R W 1999 Global change in local places: how scale matters. <i>Climatic Change</i> 43 : 601–28 Easterling W E et al. 1998 Spatial scales of climate information for simulating wheat and maize productivity: the case of the US Great Plains. <i>Agricultural and Forest Meteorology</i> 90 : 51–63
Transects	Kromm D E, White S E (eds.) 1992 <i>Groundwater Exploitation In The High Plains</i> . University Press of Kansas, Lawrence, KS Moran E, Brondizio E 1998 Land use change after deforestation in Amazonia. In: Liverman D et al. (eds.) <i>People and Pixels: Linking Remote Sensing and Social Science</i> . National Academy Press, Washington, DC, pp. 94–120
Gradients	Allan W 1965 <i>The African Husbandman</i> . Barnes and Noble, New York Porter P W 1965 Environmental potentials and economic opportunities—a background for cultural adaptation. <i>American Anthropologist</i> 67 : 409–20
Time	Broek J O M 1932 <i>The Santa Clara Valley, California: A Study in Landscape Changes</i> . Oosthoek, Utrecht, The Netherlands Carlstein T 1982 <i>Time Resources, Ecology and Society</i> . George Allen and Unwin, London
Events	Blaikie P M 1985 <i>The Political Economy of Soil Erosion in Developing Countries</i> . Longman, London
Analogs	Burton I, Kates R W, White G F 1993 <i>The Environment as Hazard</i> . Guilford Press, New York Glantz M H (ed.) 1988 <i>Societal Responses to Regional Climatic Change: Forecasting by Analogy</i> . Westview Press, Boulder, CO Wood J S, Knight C G 1974 Cross-cultural perspectives on human adjustment to arid margins. In: Newman J L (ed.) <i>Drought, Famine and Population Movements in Africa</i> . Syracuse University Foreign and Comparative Studies/Eastern Africa Series no. 17, Syracuse NY, pp. 75–126
Theory	Boserup E 1966 <i>The Conditions of Agricultural Growth: The Economics of Agrarian Change Under Population Pressure</i> . Aldine, Chicago Worster D 1985 <i>Rivers of Empire: Water, Aridity, and the Growth of the American West</i> . Pantheon, New York

different social communities to ascertain, for example, generalities surrounding a particular cultural pattern (relations between matrilineal systems and matrilineal residence); social experience (immigrants in the new world city); or community structure (comparative political systems). In human–environment studies,

comparison between communities is also important. Examples of such comparative case studies include investigation of traditional or vernacular architecture; studies of pastoral life-styles and range management strategies; cross-cultural studies of local environmental relations; analysis of ethnic farming populations in

the same locale; and studies of traditional navigation technologies and knowledge systems used by island inhabitants.

2.2 Place

Places which provide examples of a shared geographical setting or similar human–environmental process may be comparatively studied. Examples include studies of human–environment relations in specific kinds of environment, such as mountains, flood plains, and island or coastal areas; research on societal responses to extreme population density; and investigation of places affected by technological transformation, such as impacts of transportation change (super highways, dieselization of railroads), agricultural mechanization leading to rural depopulation, management of water resources, and transitions in household energy use in developing countries.

2.3 Perspective

An alternative to selection of different cases analyzed from a similar perspective is to bring multiple perspectives to a single case. Such an approach explicitly invokes a multioperationism desirable in all case studies—the convergence of different kinds of evidence toward a more complete and rich understanding. One common method for an approach based on perspective is the *emic* vs. *etic* familiar from anthropology—the understanding held by the individuals involved in contrast to the analytical viewpoint of the researcher. This distinction is now an important element of postmodern qualitative inquiry across the social sciences. The use of perspective in comparative study includes landscape analysis, analysis of environmental controversy, and multiperspective community or place studies.

2.4 Structure

Comparative case studies may be developed on the proposition that selected cases reflect the functioning of a common underlying structure, somewhat analogous to comparative linguistics. For cross-cultural translation of color terms, a formal Munsell color chart can provide a common structure on which to map various language terms for color. Less formally, comparative analysis may be based on typologies which provide the framework for comparison. In the human–environment realm, structural frameworks used as the basis for comparative case studies include Linnean plant classification compared to indigenous

ethnobotanical plant classification; fundamental requisites of a society; comparative farming systems; and human response systems in the face of natural hazards.

2.5 System–Subsystem

A comparative case study may be organized to look at various levels of a complex system to establish relationships, mechanisms, and emergent properties that vary by the level under investigation. Examples include studies of the public role in environmental management at local, state, and national levels; studies of agricultural change viewed at the family, village, society, and regional levels; and studies of the driving forces of deforestation, from family- and farm-level decisions, through regional and national forces to the global economy.

2.6 Scale Analysis

Case studies of the occurrence of phenomena at different spatial scales may be undertaken to ascertain variance or invariance of form or process. Scale analysis has become an important element in the study of the natural environment, including hydrological, climatological, and ecological systems. In human–environment investigations, scale analysis is undertaken in studies of the causes and potential impact of climate change to suggest local and larger-scale determinants of greenhouse gas emissions; appropriate levels of generalization of impacts, such as climate change on farming; and initiatives for mitigation.

2.7 Transects

Geographical transects provide one kind of spatial sampling that researchers can use to develop information and insight that can be generalized across a society or region. A transect provides one means of minimizing sample bias in an attempt to develop generalizations and analysis of differences within a society or region that might reasonably be assumed to be homogeneous. Such studies include investigation of market towns along a transect in an agricultural region; research on the dynamics of colonist settlement along new roads in rainforest areas; or exploratory archeological investigation of settlement remains in a region. In transect studies, the area being investigated may be found not to be homogeneous, suggesting comparative study of social differences (such as ethnic enclaves in an agricultural region) or environmental

differences (such as local availability of ground water in the High Plains of the USA). One special kind of transect selected in recognition of difference rather than homogeneity is the gradient.

2.8 Gradients

Cultural ecologists in geography and anthropology often undertake comparative case study analyses using sample cases selected across environmental, spatial, or social gradients rather than in areas representative of a homogeneous area. One principle underlying such studies is that society adapts to environmental differences; thus the purpose of study is to understand the nature of adaptations and how they function in the face of environmental change. Examples include studies of agricultural adaptation along an altitudinal or climatic gradient; investigation of the impact of proximity on market involvement in a developing country; or research on the human–environment interactions in different societies occupying similar environments, either across a social gradient within a similar environmental region, such as farmers in Africa, or living interspersed in the same location, such as pastoralist and agriculturalist land use management.

2.9 Time

A temporal approach to comparative case study examines the same place at different times, different places at the same time which reflect temporal differences, or different places at different times. In the first approach, a single place is examined at different times. Such a study might show the development of successive human–environment relations as population or other variables change through time. When a geographical location was occupied by different societies through time, such studies in geography were offered as evidence of ‘sequent occupance (sic),’ one of the strategies used to refute environmental determinism. Investigation of different places at the same time could be seen as a quasi-temporal process when diffusion of an innovation or process of change is spreading geographically and some areas represent an older core location with earlier adaptation while others represent preadaptation or emerging stages. An example might be the diffusion of a cash-cropping complex in a traditional rural society where various locations along a space–time gradient represent stages of adaptation of a new technology.

2.10 Events

In the evolution of comparative case study in the human–environment realm, one salient mode of analysis focused on human response to natural hazards. Natural hazards research sought common-

alties, first, in human response to floods, followed by attention to other natural hazards, to the hazardousness of places, and eventually to technological hazards. In all of these realms, comparative case studies provided the core strategy for the development of general frameworks for natural hazard analysis. Seen more generally, selection of hazard case studies for comparative research may be organized around extreme geophysical events, such as tornadoes, hurricanes, and floods, or longer term environmental threats such as drought, deforestation, or soil depletion and erosion. Comparison was also based on spatial dimensionality and temporal occurrence of hazards, as well as on hazard geometry. The event-based approach can be extended to other kinds of events: human and ecological impacts of dams; local resistance to imposed political or technological change; or suburban sprawl.

2.11 Analogs

An analog is based on the premise that objects or events having some similarity may also have other similarities. Thinking by analogy in fact underlies the comparative case study approach, in that the similarity among cases makes them in fact analogous in some respect. As a formal organizing principle, analogs inherently involve two or more cases that share sufficient commonality that insight can be gained by comparison. There is no formal procedure for developing analogs; rather, good analogies depend on ‘pragmatic considerations regarding contexts, interests and purposes’ (Jamieson 1988, p. 81) and represent *a priori* insight before the full development of the comparative cases. In comparative case study using analogs, there is no necessary juxtaposition of the cases in time or space; rather, comparison originated in a conceptual linkage that appeared promising to the researcher. Analogous case studies are strengthened when they are selected with explicit similarity criteria.

Comparative case studies developed within an analog framework include studies of the impact of modern technology in developing countries; society’s response to acid rain, ozone depletion, and future greenhouse gas stabilization; responses to past climate change; and adaptation to drought. Often, comparative case studies built upon analogs are future-oriented, asking whether insights gained from the analogs can be applied to future decisions.

2.12 Theory

Comparative case studies can address theory, wherein the cases are selected to test hypotheses from the theory. Examples include comparative case studies of the process of agricultural change with respect to population growth; modeling society’s progress through economic, demographic, disease, or mobility

transitions; testing Wittfogel's concept of the hydraulic civilization; or explicating a framework for understanding the processes of environmental adaptation. In such studies, individual cases are not only compared with other empirical cases, but compared also with the expectations from *a priori* theory.

3. The Comparative Case Study Method

The comparative case study approach draws upon the same kinds of method that are common to the case study and comparative methods. The researcher links a hypothesis to certain cases using one or more of the strategies described previously, gathers information about the cases, develops evidence from the cases, presents the cases, and constructs explanation or generalization from the evidence. Insight, creativity, intuition, and even serendipity cannot be dismissed as important elements that the researcher uses to formulate hypotheses and to develop the strategy of research.

The depth of analysis of the constituent cases in a comparative study selected for *instrumental* purposes may not be equivalent to that of an individual case examined for its *intrinsic* interest (Creswell 1997). Focus on an issue or hypothesis draws attention to specific relevant issues or variables in the cases. Yin (1989) referred to this difference as *holistic* versus *embedded* analysis, the latter connoting the selection of cases within a wider research question.

3.1 Tools for Analysis

Researchers developing new case studies for comparative analysis have a wide variety of available tools and data sources, including interviews, surveys, participant observation, ethnographic research, field observation, officially collected information, archival research, environmental analysis, remote sensing, qualitative analysis, quantitative analysis, systems analysis, structural analysis, and various modes of modeling. In contrast, researchers undertaking meta-analysis of previous studies may be confined to information developed through other researchers' selection of tools and information; the researcher may undertake further study to amplify information on the case. As with individual case studies, the results of comparative analyses are usually narrative, although normative, probabilistic, statistical, simulation, or graphic models are possible modes of conceptualizing and communicating research outcomes.

4. Evaluation of Comparative Case Studies

Although there are no formal rules for evaluating comparative case studies, there exist reasonable criteria against which studies can be assessed. These include establishing the relevance of the cases;

thorough identification and consultation of information sources; the adequacy and accuracy of reflecting information from sources; correct interpretation of sources; objectivity in reporting and interpretation; and internal consistency as evidenced by triangulation or multioperationism. If the cases purport to be a 'natural' experiment, controls provide an important strengthening of the argument. Additional criteria include the possibility of replication; reporting of contrary evidence and/or interpretation; specification of exceptions and special circumstances; recognition of and explanation for contrary cases; rigor versus superficiality; and value-added insight and understanding. Primary case studies must also meet the canons of field or archival work in the relevant discipline; comparative case studies built from meta-analysis of previously published work have potential for review by original researchers.

5. Conclusion

Comparative case studies offer a richness of insight that may be lost in the abstractions of statistical analysis, while maintaining the tradition of the case study as a means to explore generalities in the complexity of the real world. The intriguing variety of strategies for organizing comparison means that this approach to understanding human relationships to the environment will continue to be a vital part of human–environment scholarship.

See also: Comparative History; Comparative Method, in Evolutionary Studies; Comparative Studies: Method and Design; Global Environmental Change: Human Dimensions; Human Ecology; Human–Environment Relationship: Indicators; Human–Environment Relationships

Bibliography

- Creswell J W 1997 *Qualitative Inquiry and Research Design, Choosing among Five Traditions*. Sage, Thousand Oaks, CA
- Eisenstadt S N 1968 Comparative study. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York, Vol. 14, pp. 421–9
- Hamel J 1993 *Case Study Methods*. Sage, Newbury Park, CA
- Jamieson D 1988 Grappling for a glimpse of the future. In: Glantz M H (ed.) *Societal Responses to Regional Climatic Change*. Westview Press, Boulder, CO, pp. 73–93
- Leach E 1968 The comparative method in anthropology. In: Sills D L (ed.) *International Encyclopedia of the Social Sciences*. Macmillan, New York, Vol. 1, pp. 339–44
- Oreskes N, Shrader-Frechette K, Belitz K 1994 Verification, validation, and confirmation of numerical models in the Earth sciences. *Science* **263**: 641–6
- Ragin C C 1987 *The Comparative Method: Moving Beyond Qualitative and Quantitative Analysis*. University of California Press, Berkeley, CA
- Ragin C C, Becker H S 1992 *What is a Case? Exploring the Foundations of Social Inquiry*. Cambridge University Press, Cambridge, UK

Stake R 1995 *The Art of Case Study Research*. Sage, Thousand Oaks, CA

Yin R K 1989 *Case Study Research: Design and Methods*. Sage, Newbury Park, CA

C. G. Knight

Human–Environment Relationship: Indicators

In 1987 the World Commission on Environment and Development introduced a resonant idea: ‘Sustainable development,’ the commission wrote (1987, p. 43) ‘is development that meets the needs of the present without compromising the ability of future generations to meet their own needs.’ Sustainable development names a hope—that human prosperity can be reconciled with the natural world upon which all material economies ultimately rely. It is a notion that has exercised a broad influence in public discussions, notably in the United Nations Conference on Environment and Development in 1992 and the document *Agenda 21* (see Sitarz 1993) adopted at that conference in Rio de Janeiro.

Finding an operational definition of sustainable definition has proved difficult, however. In a dynamic global economy where substitutes for a wide range of goods and services are readily available—at a price—it

is difficult to define ‘needs’ in a way that makes clear what must be conserved or preserved so that future generations can meet their needs (as they define them). In addition, we have yet to develop a generally accepted accounting system that can render authoritative judgments when an ecosystem verges on environmental bankruptcy. But there is no question that biodiversity is being lost at a historic pace, that fossil fuels cannot be replaced on the timescales on which they are now being burned, or that soil erosion diminishes the agricultural capability of some desperately poor people.

It is in this context that indicators play a significant role as people and societies search for ways to meet the requirements of nature and humans. Sustainable development is a long-term, large-scale phenomenon, if it is feasible at all, and it is likely to emerge slowly and unevenly in the world economy. Like other historical phenomena, sustainability may be hard or impossible to see at the level of everyday experience until it is well underway. Indicators rooted in scientific measurement provide an alternative way to perceive and to assess progress toward a sustainable future.

Indicators are quantitative measures of human well-being, economic activity, and natural processes and conditions; they are needed to sense the degree to which human activity may be continued or expanded in the future. The most useful indicators are repeated, precise measurements of natural and social phenomena. Such data series illuminate trends and identify changes; accordingly, indicators can sound alarms, define challenges, and measure progress. The measure-

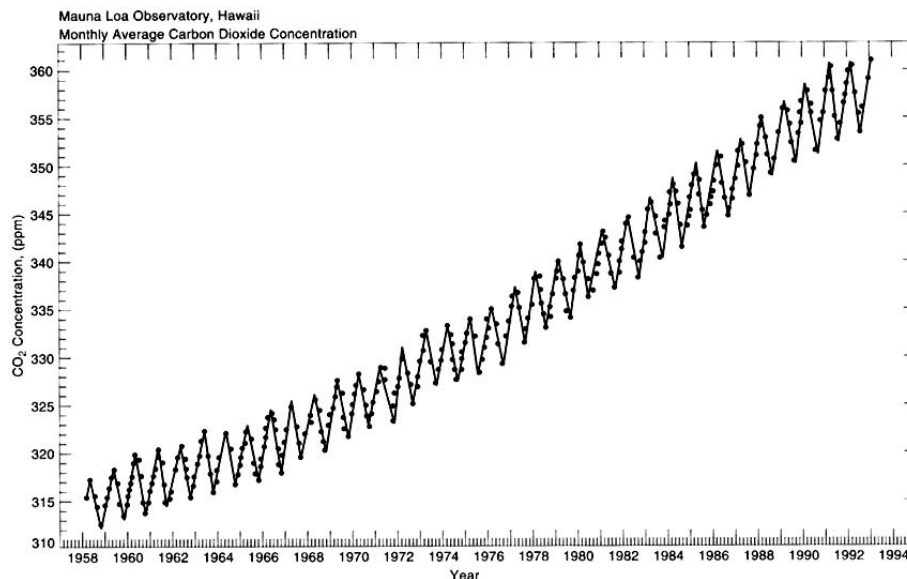


Figure 1
Atmospheric carbon dioxide concentrations at Mauna Loa, Hawaii

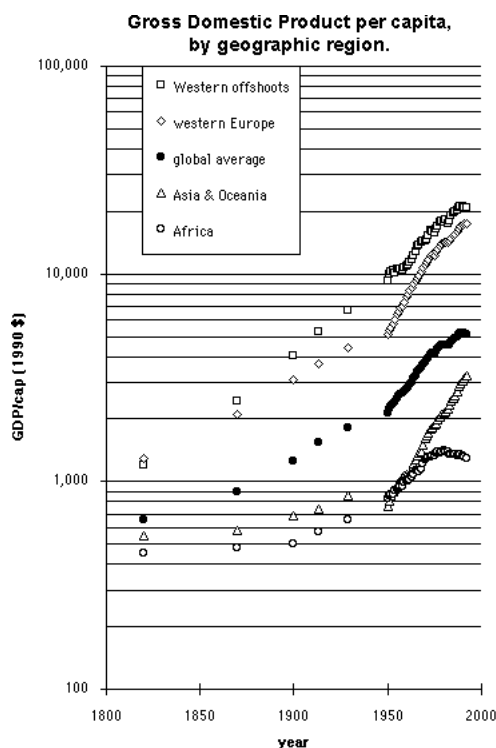


Figure 2

Estimated per capita gross domestic product, by region (source: Maddison 1995)

ment and interpretation of indicators requires sustained efforts and resources from governments, firms, nongovernmental organizations, and the scientific community. Indicators communicate, sometimes evocatively and often across cultural and national boundaries. The indicators described below illustrate how both social and natural realms have been affected by human action over the long term and on large scales.

Measurements of carbon dioxide levels in the atmosphere warn of climate change, but the warning is ambiguous in policy terms (see Fig. 1). Like other gases, carbon dioxide mixes and spreads in the atmosphere, so that its impacts are widespread even when its sources are localized. (Note the seasonal oscillations, however: blending within the northern and southern hemispheres takes place faster than global mixing.) Like other changes in climate, responses are delayed in time and reflect multiple perturbations, many not of human origin. Nonetheless, the steady long-term trend reflected in this data series has triggered worldwide scientific attention and international agreements. The attention commanded by this indicator is a measure of the influence of environmental concerns in formal decision-making, the arenas in which quantitative indicators are indispensable for communicating credible information.

Estimates of economic output demonstrate that economic growth has been substantial in all parts of the world over the past two centuries. Although even the poorest have seen their incomes increase, the gap between rich and poor has also widened (see Fig. 2). Over the next several decades, indicators will be needed to monitor and report on human welfare and life support systems, with measurements on different spatial and temporal scales to inform social choice and debate. Some of these indicators may come to define whether human populations are heading toward sustainability, in the way that infant mortality and life expectancy symbolize progress in public health. Some indicators can also aid navigation, helping individuals and societies to steer a course toward sustainability.

This article sketches current work on indicators and outlines an approach grounded on a widely shared set of normative objectives.

1. Indicators of Sustainable Development

Humans have made repeated, precise measurements of some phenomena since ancient times; astronomical and meteorological signs have guided agriculture for millennia. The use of indicators expanded as industrial societies sought to measure and manage a widening variety of environmental and societal parameters (see Porter 1995).

Today, many people routinely scan prices, news and weather reports, and other sources to inform action and expectation in the short run. Indicators can perform similar functions with respect to long-term, large-scale changes. A striking example of the influence of indicators is provided by the adoption of a common European currency at the turn of the century. In preparation for monetary union, the governments of the European Union agreed to harmonize economic policies by meeting and maintaining budget deficit targets, defined as specified fractions of gross domestic product.

For complicated social objectives such as sustainable development, no single indicator provides a definitive benchmark. Instead sets of indicators have been proposed, identifying multiple parameters that together provide assessments of three basic dimensions—prosperity, environmental quality, and human well-being. Indicators have usually been interpreted using the Pressure-State-Response (PSR) framework, which posits linkages between human action and environmental consequences (Adriaanse 1993). Human activities exert *pressures*, such as burning gasoline in cars, that alter the *state* of environmental variables, such as the quality of air in a city. Impaired states, in turn, elicit *responses*, such as regulations governing pollution-control technology. PSR makes use of data that often are already collected for administrative purposes. The approach captures an important element of sustainable development:

humans can impair the life-support systems of the natural world, calling forth responses aimed at protecting environmental quality. It is straightforward to expand PSR to include the impact of economic development on human welfare.

There is widespread agreement on the utility of PSR but no convergence on a single set of indicators for the measurement of sustainability. There have been numerous efforts to assemble indicators of sustainable development, however, ranging in jurisdictional scope from municipal to international and spanning ecological scales from watersheds to the planet as a whole. (For a compilation, see International Institute for Sustainable Development 2001.) Many indicators in these sets do not have long time series, so the ability to discern trends is limited. Coverage is uneven. Although environmental indicators outnumber the rest, there is greater depth in economic and some social dimensions. Many variables are available for nation-states but few at smaller spatial scales. Information on land cover and land use is weak, especially in developing countries, and institutional indicators are sparse. Some proposed indicators reflect considerable imagination, notably those proposed by the international Balaton Group (Meadows 1998) and in many community-scale projects. For instance, gardening activity was used as a measure of community vigor in Seattle, Washington.

Indicator sets assembled by the UN Commission for Sustainable Development (CSD) and the World Bank complement one another in an illuminating fashion. The CSD indicators (1996) were assembled using the PSR framework. Selected through a consensus process without an agreed operational definition of sustainable development, the CSD indicators are numerous, diverse in the methods used to measure development or sustainability, and include a large number of indicators where reliable measurements do not exist.

The World Bank (1997), in contrast, estimated three national capital accounts. Each attempts to capture the value of an aspect of national economies. 'Produced' capital includes physical capital and financial claims, both familiar in standard accounting practice. A second account measures natural capital, the resources and services provided by the natural world. In principle this includes standing timber and soil fertility. Estimates of natural capital are as yet primitive. A third component of wealth, quantitatively the largest, is human resources, the economic value of labor, knowledge, and social institutions.

The Bank's indicators are highly aggregated and have been estimated using drastic assumptions, but they are conceptually clear. The wealth of nations should be considered in three parts. At least at the margins, wealth can be transferred from one account to another in ways advantageous to people—as when productive farmland is fenced off to create a nature reserve. By contrast, the United Nations CSD indicators do not warn unambiguously of imminent

hazards in any ecosystem or society, nor do they provide guidance on how to pursue sustainable development.

The United Nations indicator set does include variables that are not now being measured, however. This is a notable strength of its process, reflecting a realization that much of what needs to be sensed about sustainable development remains unclear. The UN also has sponsored a wide-ranging research effort under the aegis of the International Council of Scientific Unions (Moldan et al. 1997).

In sum, the existing indicators of sustainable development are limited in their effectiveness by the lack of agreement on what to develop, what to sustain, and for how long—that is, a lack of agreement on the meaning of sustainable development. Proposed indicators of this politically popular idea have drawn upon work done in past decades on the measurement of human welfare and the condition of the environment. These efforts bring together many sources of illumination but have yet to produce a set of measurements of social and natural conditions that could plausibly lead to prosperity for all while conserving the life-support systems on which human economies rest.

2. Trends and Transitions

That we do not have a clear path to a sustainable future is hardly surprising: identifying indicators that point to a sustainable future is a normative task, one that is inescapably difficult in a fractious world of multiple cultures. One approach, used recently by the US National Research Council (1999), suggests directions in which indicators of sustainability might evolve over the next several decades. In a report entitled *Our Common Journey*, the Research Council's board on sustainable development (BSD) focused attention on a sustainability *transition*, a decades-long process in which human economies might move toward a durable relationship between meeting needs and conserving nature. Instead of the elusive endpoint of sustainable development, the BSD sought to highlight the scientific, technological, and societal innovations required to 'feed, house, nurture, educate, and employ many more people than are alive today, while preserving the basic life support systems of the planet and reducing hunger and poverty.'

The underlying metaphor is the demographic transition, the long-term slowing of human population growth which is projected to lead to a stable population size in the latter part of the twenty-first century. The transition in population, already underway for more than two centuries, has brought the human species from a situation of relative stability with high birth rates and high death rates to a situation of stability at low birth rates and low death rates. Because death rates were lowered before birth rates, there has

been a large net increase in population size. A sustainability transition would correspondingly carry human economies from an agro-ecological past, through a rapidly growing industrial phase, into a future in which materials would be used far more efficiently, generating less pollution. As population sizes stabilize, moreover, wide disparities in material standards of living may be moderated by addressing hunger and poverty. Like the demographic transition, a sustainability transition would evolve over decades.

The sustainability transition is an optimistic scenario; signs that it may be in progress are mixed. The indicators above show rapid, perhaps exponential growth rather than stabilization. Yet forests have regrown in advanced industrial nations; famine has been significantly reduced in poor countries, even as the number of hungry people in the world continues to expand; and environmental concerns arose, perhaps surprisingly, in societies that suffered from some of the worst pollution. None of these indications points to a turning away from unbounded expansion, and there is nothing so clear as the peak in the rate of growth in human numbers which occurred in the late 1960s.

Calling for research to make a sustainability transition possible is an unsurprising message from groups of natural and social scientists. What is less apparent at first sight is the intellectual strategy and its implications for indicators. The sustainability transition eschews prediction in favor of a *normative* definition of the task ahead, to meet human needs while preserving the life-support systems of the planet. The BSD found a basis for defining human needs in treaties and the declarations of international conferences devoted to nutrition, housing, social welfare, education, and employment. While such statements articulate aspirations with few resources or implementation plans to back them up, on the decades-long timescale of a sustainability transition, one can imagine quantitative changes large enough to make a qualitative difference with respect to both human needs and the conservation of the natural world.

3. Navigating Transitions

Like the carbon dioxide trend above, a sustainability transition may well require quantitative indicators to be seen at all. The BSD proposed two families of indicators, one to perceive transitions and one to guide social choice in pursuit of deliberate transitions.

Seeing a sustainability transition using the BSD definition implies sensing the degree to which human needs are met and the degree to which life-support systems are put under stress. Human welfare has been monitored since the depression of the 1930s drew attention to the role of governments in meeting basic needs. Appropriate quantitative indicators have been developed, but at the turn of the century they were not yet being reported regularly in a forum, such as the

annual report of the United Nations Development Program, where crises could be highlighted and emergent transitions discerned.

Environmental indicators are numerous and need conceptual organization. The BSD's proposal was geographic, focusing attention on four scales of significance to a long-term transition:

(a) *Global circulatory systems* are affected by human activity, including the atmosphere, climate and ocean circulation, trade and travel, and the spread of exotic species and diseases. Trouble in the circulatory systems is important because the scale of circulation can implicate the entire planet more rapidly or persistently than governments can regulate. Although greenhouse gases and ozone-depleting substances were carefully studied by the end of the twentieth century, biological and information-intensive flows were not yet being assessed as components of a global circulatory system.

(b) *Regions vulnerable to critical environmental damage* are being identified (Kasperson et al. 1995). For instance, the Aral Sea and the Valley of Mexico are experiencing severe, possibly irreversible damage. Because the governing factors—including a vulnerable natural setting, lack of feasible alternatives, and governing institutions unable to take effective action—vary among regions, there is not a single set of indicators that can monitor the combination of social and natural factors that lead to irreversible damage.

(c) *Ecosystem services for urban areas* are only beginning to be monitored systematically (e.g., Baltimore Ecosystem Study 2001), even though urbanization is continuing to accelerate in developing countries. Because many ecosystem services are unpriced (e.g., flood control) or allocated by monopolies (e.g., water supply), market signals are absent or misleading. Yet ecosystem services tend to be irreplaceable when damaged and inflexible because they are supplied by costly infrastructure.

(d) *Parks and protected areas*, managed to enable their biota to persist indefinitely, have been identified on a place by place basis, rather than through a consistent set of appraisals of their long-term sustainability. Attempts to conserve biodiversity on larger scales (Olson and Dinerstein 1998) are just getting started.

In each of these settings, indicators form an indispensable but incomplete part of the intelligence needed to discern a sustainability transition in the decades to come.

A complementary function of indicators is to inform public policy and other decisions, large and small. Some navigational aids are available, although it seems likely that others may also come into wide use.

(a) *National capital accounts*, emerging from the World Bank's work, could provide a simple but crude measure of sustainability: each of the three kinds of capital might be transformed by human activity, but so long as a nation's total capital increases over time,

its trajectory may lie in roughly the direction to contribute to a sustainability transition. The word 'roughly' is important: a utilitarian metric is inadequate to assess sustainability. But the capital accounts draw attention to transformations among forms of wealth—transformations that will continue through, and beyond, any long-term search for sustainability.

(b) *Assessing policies* using indicators assembled under the PSR framework can lead to adaptive management. Adaptive management treats policies as experiments, designing them so that lessons may be learned reliably from the implementation of policies, even those that fail (Lee 1993). In the PSR context, this means using pressure and state indicators to test the effectiveness of responses.

(c) *Monitoring ongoing transitions* provides analysis of trends, particularly those that carry large future commitments, such as population growth. Even if the population continues to stabilize, it will be essential to improve technology and energy efficiency so that humans can accomplish more, economically, with less impact on the natural world.

(d) *Surprise diagnosis* will be needed because a sustainability transition is an improbable development. Reflecting on environmental surprises over the past three decades, Kates and Clark (1996) concluded that we can learn from surprises, so that we can better anticipate, avoid, or mitigate their consequences. But knowing how to improve management is also a temptation to operate closer to the edge. Surprises could therefore become *more* frequent as humans gain better knowledge of the world; this possibility qualifies the conventional notion that science is valuable because it improves our ability to control or at least to predict danger. From this perspective, surprises are valuable indicators in themselves. It is well accepted that surprise should produce humility. Surprise should also produce curiosity. On the timescale of the sustainability transition, curiosity and the learning it prompts are likely to be important, whether or not control can be extended in the short term.

4. Indicators and Social Learning

There is no agreement about which indicators to use to measure the movement of economies toward or away from sustainability. One should accordingly expect spirited debates over the value, biases, and meanings of indicators. In the related sphere of economic policy one can observe over the past half-century sharp disputes in the US over economic growth, the incidence of poverty, unemployment, and inflation. All these have been indicators that American politicians think will influence voter behavior (Tufte 1978). Remarkably, the independence of the data-gathering and analytical organizations has survived, despite their location within government agencies (see Duncan and Shelton 1978).

This is an important lesson: the independence of science is central to the social value of scientific information. Another lesson is that surprise is itself a valuable indicator. Governments and societies should anticipate unexpected things to happen. In a policy context, this idea is a kind of precautionary principle: because surprise is likely, action should be undertaken with thought, humility, and caution.

Indicators used to report on a sustainability transition are likely to be biased, incorrect, inadequate, and indispensable. Getting the indicators right is likely to be impossible in the short term. But not trying to get the indicators right will surely compound the difficulty of discovering, inventing, and achieving sustainable development.

See also: Development: Sustainable Agriculture; Ecology, Political; Environment and Development; Environmental and Resource Management; Environmental Challenges in Organizations; Environmental Change and State Response; Environmental Cognition, Perception, and Attitudes; Environmental Justice; Environmental Planning; Environmental Policy: Protection and Regulation; Human–Environment Relationships; Sustainable Development; Sustainable Transportation

Bibliography

- Adriaanse A 1993 Environmental Policy Performance Indicators. A study on the development of indicators for environmental policy in the Netherlands. April. Netherlands: Sdu Uitgeverij Konintginnegracht
- Baltimore Ecosystem Study 2001 Web site: <http://www.ecostudies.org/bes/> Visited 8/20/01
- Duncan J W, Shelton W C 1978 *Revolution in United States Government Statistics, 1926–1976*. US Government Printing Office, Washington, DC.
- International Institute for Sustainable Development 2001 *Compendium of Sustainable Development Indicator Initiatives and Publications*. Web site: <http://iisd1.org/measure/compindex.asp>. Visited 8/20/01
- Kasperson J X, Kasperson R E, Turner B L II (eds.) 1995 *Regions at Risk: Comparisons of Threatened Environments*. United Nations University Press, Tokyo
- Kates R W, Clark W C 1996 Environmental surprise: Expecting the unexpected. *Environment* 38 (2): 6–11, 28–34
- Keeling C D, Whorf T P 2001 Atmospheric CO₂ records from sites in the SIO air sampling networks. In: *Trends: A Compendium of Data on Global Change*. US Department of Energy, Oak Ridge, TN
- Lee K N 1993 *Compass and Gyroscope. Integrating Science and Politics for the Environment*. Island Press, Washington, DC
- Maddison A 1995 *Monitoring the World Economy, 1820–1992*. Development Centre of the Organisation for Economic Co-operation and Development, Paris
- Meadows D 1998 *Indicators and Information Systems for Sustainable Development*. A report to the Balaton Group. The Sustainability Institute, Hartland Four Corners, VT, pp. 78. Web site: <http://www.sustainer.org/pubs/>. Visited 8/20/01
- Moldan B, Billharz S, Matravars R 1997 *Sustainability In-*

- dicators. *A Report on the Project on Indicators of Sustainable Development*. Wiley, New York
- National Research Council (USA) 1999 *Our Common Journey: A Transition Toward Sustainability. Report of the Board on Sustainable Development*. National Academy Press, Washington, DC. Web site: <http://lab.nap.edu/catalog/9690.html>. Visited 8/20/01
- Olson D M, Dinerstein E 1998 The Global 200: A representation approach to conserving the earth's most biologically valuable ecoregions. *Conservation Biology* **12**: 502–15
- Porter T M 1995 *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press, Princeton, NJ
- Sitarz D (ed.) 1993 *Agenda 21: The Earth Summit Strategy to Save our Planet*. EarthPress, Boulder, CO
- Tufte E R 1978 *Political Control of the Economy*. Princeton University Press, Princeton, NJ
- United Nations Commission for Sustainable Development 1996 *Indicators of Sustainable Development. Framework and Methodologies*. United Nations, New York
- World Bank 1997 *Expanding the Measure of Wealth: Indicators of Environmentally Sustainable Development*. World Bank, Washington, DC
- World Commission on Environment and Development 1987 *Our Common Future*. Oxford University Press, New York

K. N. Lee

Human–Environment Relationship: Prehistoric

Human–environmental relationships comprise a web of culturally imposed constraints in a complex environment, within which people adopt a variety of subsistence pursuits. These may or may not have long-term effects on that environment. During the course of Pleistocene prehistory, humans were few in numbers and constrained by their technology and social organization from unlimited spatial expansion or assuming the trophic role of a super predator. But they may have influenced megafaunal extinctions or vegetation change through manipulation of fire. By the beginning of the Holocene, greater subsistence success allowed exponential population growth, coinciding with agricultural origins in several world regions. Domesticated plants and animals became artifacts, allowing humans to achieve ecological dominance and incrementally transform their environments. But the management strategies followed and the landscapes created were constructed culturally.

1. Inter-relationships

Human–environmental inter-relationships are complex and nuanced. On the one hand, resources vary over time and are unevenly distributed in geographical

space, so that strategies for utilization will differ for the short and long term, and at small or large scales (Butzer 1982). On the other, human communities subscribe to a wide array of subsistence and cultural modes, and attempt to reconcile new information with traditional values. Alternative pathways for possible change are evaluated through economic and cultural choices, which presume a major role for cognition, information, and decision-making. Environmental factors set ‘boundary conditions’ to human behavior at a particular technological and organizational level only in so far as they pose socially unacceptable risk. Short- or long-term environmental crises may be met by routine accommodation, active mitigation, or shifts in subsistence pursuits or social organization. Outcomes therefore are unpredictable.

2. A Pleistocene Prelude to Change

The roughly two million years of Pleistocene prehistory—the Paleolithic—were critical for human evolution and presumably influenced the ‘hard-wiring’ of basic human behavior. Emerging human–environmental interrelationships not only suggest a datum to examine such patterns in more recent times, but may have set a trajectory for later environmental behavior. Given the rudimentary archaeological record and in the absence of direct ‘observation,’ it is next to impossible to grasp the cognitive assessment of environmental problems, or even to infer a response or to understand attenuated processes of systemic change. A very real danger in Pleistocene prehistory is to posit a cause and effect relationship, based on little more than coincidence of several categories of proxy data. There are no simple answers to the broad, interdisciplinary questions about Pleistocene human–environmental relationships.

2.1 Human Vulnerability to Pleistocene Environments

The expansion of early human groups to higher latitudes and new continents during the course of the Pleistocene gives the impression that they learned progressively to cope with cold or arid environments, where resources were different than in tropical Africa. But the Paleolithic record is strikingly discontinuous in space and time, raising doubts as to the ability or choice of Paleolithic peoples to colonize ‘marginal’ environments with widely dispersed but clustered resources (Butzer 1982, Chaps. 12–14, 1991).

In what is now the semiarid interior of southern Africa, settlement was limited to periods of substantially moister climate and closely linked to water sources. Even some valleys with dependable rivers were unutilized for repeated spans of many tens of

millennia, and not all periods of moister climate were accompanied by an automatic territorial expansion of settlement. Effective human occupation across the dry interior of southern Africa was delayed until well into Holocene times, after 10,000 before present (BP).

Paleolithic people probably avoided this region because food resources would have been few and thinly spread during the dry seasons of poor years. Plant foods would have had low predictability and low net productivity, while the available game would have been difficult to hunt without bow and arrow, sophisticated traps, or complex game drives. If bands of hunters did penetrate such dry regions, they would have been so widely dispersed as to encounter neighboring bands only sporadically, leaving them with incomplete information as to the status of resources, e.g., animal movements or the condition of plant foods (Butzer 1991). During periods of severe drought stress, starving animals also lack fat while excessively lean meat poses severe metabolic problems for starving people (Speth 1987).

The true capacity of drought-prone, low productivity, and low predictability environments (see Butzer 1982) to sustain human life is a great deal less than its overall resources, averaged over time and space, would suggest. A similar ‘marginality’ model is offered by Bocquet-Appel and Demars (2000) for western Europe between 40,000 and 11,000 BP. In glacial-age France settlement was concentrated in a ‘refuge zone’ located in northern Aquitaine. During less harsh, subarctic conditions prior to 22,000 BP small temporary sites extended northeastward to the Rhine, but then retracted to southwestern France during the glacial maximum, c. 22,000–16,000 BP. Thereafter, as the climate ameliorated, an increasing number of larger sites are found in areas previously not occupied, as far afield as Britain and Germany. Across these 30 millennia, regional population may have increased from 4,000 to as many as 40,000 people, in part facilitated by improved hunting skills.

There was a vast ‘marginal’ environment in western Europe where not drought, but the low productivity and predictability of a subarctic environment, with dangerous turns of weather and hard, snowy winters, made glacial-age occupation hazardous except for summer-season hunting forays. Risk-prone environments such as semiarid thorn bush or arctic tundra seem to have been avoided, in favor of demographic clustering in dependable, productive environments. The persistent and extensive spatial gaps of the Paleolithic archaeological record in better-investigated areas appear to be real, rather than simply a function of site visibility and preservation.

Until the close of the Pleistocene, human groups did not demonstrably ‘push the envelope’ by venturing far or long into marginal environments. Incremental human expansion into higher latitudes, especially during the late Pleistocene, probably targeted environmental mosaics capable of supporting sizeable

population clusters on a long-term basis. That implies a cognitive ability to explore new regions, so as to assess viability and risk, in the context of available technology and subsistence skills.

Changing hunting proficiencies are demonstrated by quantitative differences in the animals slaughtered in South African or western European sites of different periods. Before about 40,000 BP hunting focused on relatively docile animals, whereas in Late Palaeolithic times more dangerous animals or those from difficult montane environments were hunted regularly, and aquatic resources were systematically exploited on defined, seasonal rounds (Klein 1999). In part this reflected the invention of the bow and arrow, more versatile throwing javelins, and a variety of harpooning and fishing gear. But better tactical and collaborative organization of hunting and fishing activities will have been just as important.

The exponential increase in hunting efficiency evident across Eurasia between 25,000 and 10,000 BP implies a revolutionary growth in food-energy exploited, making possible rapid demographic expansion during the closing millennia of the Pleistocene. In both Cantabrian Spain and interior South Africa site sizes or artifact concentrations increase exponentially during the Late Pleistocene, and there were some 30 to 70 times as many sites (per millennium) about 10,000 years ago than there were about 100,000 BP. The ‘take-off’ in site numbers and implicit population growth begins at different times in different regions, namely 22,000 BP in Cantabrian Spain, 16,000 BP in France, 14,000 BP in northwestern Siberia, 10,500 BP in North America, and 8000 BP in temperate Australia (Butzer 1991, Bocquet-Appel and Demars 2000).

This carries implications for the relative proficiency of prehistoric peoples to utilize environmental resources. Paleolithic populations remained very small and vulnerable across nearly two million years, precisely because they lacked the technology and cognitive ability to move higher up the trophic pyramid. They only achieved the role of a major predator during the closing millennia of the Pleistocene.

2.2 Late Pleistocene Megafaunal Extinctions

The disappearance of mammoth, mastodon, woolly rhino, great sloths, or sabre-tooth ‘tigers’ from the Holarctic fauna at the end of the Pleistocene continues to intrigue. Giant kangaroos and cow-sized diprotodons from Australia also belong on this list of ‘megafauna’ (Martin and Klein 1984). But unlike the end-Cretaceous extinction of the dinosaurs, those of the terminal Pleistocene did not eliminate the dominant orders of larger animals, but fell within the scope of evolutionary ‘turnovers’ incompletely documented for earlier phases of the Pleistocene.

Species extinctions are a grim reality of recent centuries, and can be attributed to chemical impacts,

habitat destruction and harassment, predation by overhunting, or deliberate extermination. Some such extinctions or near-extinctions were rapid, others followed many centuries of sustained attrition. But for Late Pleistocene megafaunal extinctions there is neither direct testimony nor a promising analytical method to identify cause and effect, such as for the demise of the carrier pigeon, or the near-extinction of bison, bald eagle, or whooping crane.

The scope of this problem is illustrated by the catastrophic extinctions on New Zealand after its first colonization by the Polynesian ancestors of the Maori, less than a millennium ago (Anderson and McGlone 1992). Two families of large wingless birds (moas) became extinct, as well as an array of songbirds and marine taxa, all prior to European intrusion during the early 1800s. Although moa extinction within 4 centuries is well-enough documented, together with butchery, woodland fragmentation, and introduced dogs and rats, we will never know the human dimensions of this event, on the very threshold of oral history. Yet catastrophic extinctions on isolated islands of the Pacific Ocean during the last 2 millennia involved fewer variables than the response of more complex, continental ecosystems to a familiar human predator before 10,000 BP. These recent examples do not therefore provide a model applicable to the Late Pleistocene.

The case of extinctions in Australia makes the point. Between perhaps 30,000 and 14,000 BP a total of 42 larger vertebrates, mainly marsupials, and several taxa were subject to size reduction. Two of the vertebrate families that disappeared were huge, clumsy and slow-moving browsers, reminiscent of large mammals of the early Tertiary. The kangaroo family (Macropodidae) lost six genera (or 30 species), but it could be argued that the surviving kangaroos and wallaby (10 species) are 'modernized,' efficient, and smaller forms of the same lineages (phyletic replacement). In the big picture, this suggests the consequence of intergeneric and interspecific competition.

There also were several major environmental discontinuities during the same time span, including filling and drying of lakes, expansion and 'downsizing' of river systems, and extensive activation of eolian sands (Ross et al. 1992). Pollen profiles suggest re-expansion of browsing habitats about 10,000 BP (Dodson et al. 1992). In combination with the sudden, drastic climatic perturbations for which Australia is noted today, that implies repeated episodes of fragmentation and reduction of habitats during the Late Pleistocene, rather than a single and unidirectional change at its close. Environmental flux of sufficient amplitude would selectively disadvantage the demographic viability of 'archaic' forms, accelerating competitive exclusion among genera with overlapping niches.

People settled parts of Australia before 40,000 BP, but they were few in number, their distribution limited

to more productive environments, and they introduced no animals other than the dingo, at a much later date (Kohen 1995). Artifacts have been found at a handful of faunal sites that lack patent evidence of butchering activity. How does one link the coexistence of human hunters or scavengers with extinctions over a time range of 25 millennia? Alternatively, how does one distinguish an evolutionary turn over from human predation? Temporal correlations do not provide process-links.

The case of North American extinctions is equally puzzling. Seven common genera disappeared 12,000–10,000 BP, at a time of demonstrated human predation, but of the 31 other extinct genera or species (all poorly dated), at least several phased out while even a human presence on the continent is contested (Grayson 1991). Twenty-seven of the extinct genera and six extinct species were found only in North America. In effect, this great faunal turnover primarily impacted New World forms, many of which were 'archaic,' to select for a smaller group of expansive ruminants (bison, deer, pronghorn), and carnivores (bear, wolf, coyote). But not all niches were refilled, and Spanish equids successfully recolonized grassy shrublands of the American West and Argentina.

Late Pleistocene environments in North America have no modern analogs: both grasslands and woodlands were 'patchy,' with fine-grained mosaics, greater local plant diversity, and unfamiliar seasonal productivity patterns (Graham et al. 1996). Both the extinct grazers and browsers were unspecialized, filling their energy requirements with a broad range of plants, as is typical for monogastric ungulates. At the close of the Pleistocene, such 'patchy' biomes were replaced by monotonous grasslands or closed-canopy forests, more suitable for specialized ruminants (bison, deer, caribou), whose stomachs can degrade plant toxins and more rapidly assimilate energy. All the large, monogastric grazers (mammoth, horse) and mixed browsers (mastodon, ground sloths, camelids), which had long gestation periods, became extinct. Also symptomatic of environmental stress, but inadequately understood, is the progressive reduction of size ('dwarfing') of the disappearing taxa during the Late Pleistocene, as is the 'dwarfing' of phyletic replacements among surviving genera worldwide.

Unlike Australia or North America, the contiguous continents of the Old World were little affected by extinctions (Martin and Klein 1984): a few phyletic replacements here and there, and the more dramatic disappearance of elephant and rhinoceros in Palearctic Eurasia for the first time since the Tertiary. Unimpeded contact and faunal interchange in the Old World implies a steady intergeneric competition for overlapping niches, which may have driven Old World faunas to modernize earlier. As on the other continents, that does not exclude human predation as a triggering mechanism for extinction of some vulnerable taxa. But neither does the present evidence single

out humans as super-predators ten millennia ago, whatever the environmental inputs and biological feedbacks of end-Pleistocene extinctions.

2.3 Fire Ecology and Vegetation Change

The human use of fire is a potent tool of habitat alteration or destruction. That is supported by both ethnohistorical evidence and quantification of charcoal particles from pollen profiles. Fire depends on the availability of fuel, climate, and ignition. ‘Natural’ fires, ignited by lightning strikes, have facilitated expansion of grassland and savanna since mid-Tertiary times (Kershaw et al. 1997). Less clear is the degree to which prehistoric peoples manipulated fire so as to alter vegetation ecology, because it is difficult to separate the evidence for artificial vs. ‘natural’ fires.

In Australia, indigenous people continue to use fire for many purposes: to drive game, to clear undergrowth for visibility and stimulate fresh growth, and to serve a range of defensive, domestic, and social purposes. The popular truism that indigenous fires were ‘controlled’ habitually is not supported by the facts, nor is it consonant with the experience of wildfires on other continents. Fire was an instrument of environmental management in that it once kept most Australian eucalyptus woodlands grassy and free of shrubby undergrowth, although such woodlands had already evolved across five million years in an interactive relationship with fire.

Pollen profiles from south-eastern Australia that cover most of Holocene time show charcoal particles increasing exponentially during the last 2 millennia or so, to culminate shortly before or after AD 1800, even though there is no consistent climatic trend during that period. It is possible that the elaborate burning culture recorded after 1788 only evolved during the late Holocene, after changes in cultural practices. The density of archaeological sites in south-eastern Australia doubled during the last two millennia, compared with the previous three (Dodson et al. 1992), suggesting that increasing use of fire primarily reflected demographic expansion or greater mobility. Given the much smaller and disjunct Australian populations of the Late Pleistocene, it seems improbable that human manipulation of fire could have produced a striking charcoal peak 38,000 BP in northern Queensland (see Kershaw et al. 1997), since the ‘event’ saw rainforest replaced by eucalyptus woodland, which instead implies climatic change. To argue that early use of fire was a co-agency in megafaunal extinctions will require a comprehensive, ecological case and an archaeological context.

Unlike Australia, which has a desertic core, sub-Saharan Africa is dominated by grasslands with scattered trees (savannas), mosaics of savanna and open woodland, or open woodland with patches of grass. Although extensive tracts of such open vegetation reflect either sandy substrates or extensive

seasonal flooding, others have been modified and maintained by recurrent fire. It was once believed that Africa’s savannas were created and expanded progressively by deliberate burning through much of human prehistory, but that is no longer tenable. The great diversity of cursorial ungulates for which Africa is noted began to evolve during the Miocene, while Pliocene deposits reveal abundant pollen of grasses as well as evidence of charred wood. During glacial-age intervals of drying, such as 30,000–10,000 years ago, much of sub-Saharan Africa was an almost treeless grassland, with woodland or forest reconstituted at the beginning of the Holocene.

Pollen profiles from the highlands of Malawi show that a mosaic of grassland and forest remained unchanged from the Pleistocene to the present (Meadows 1984). In interior South Africa, the distribution of grass and semidesert shrub shifted according to climatic trends and was little affected by the first appearance of indigenous herders. Until agricultural activities began to create patches of disturbed vegetation in tropical Africa, mainly after 3000 BP, increases in grass pollen and charcoal currently are attributed to episodes of drier climate, thus the opening of the Cameroon ‘Grass Fields,’ formerly believed to be anthropogenic (Zogning et al. 1997). In Africa and elsewhere it is important to distinguish slash-and-burn agriculture, as a focused form of micromanagement within forest environments, from the modification of savannas or open woodlands to attract game or improve pasture for livestock. Prehistoric fire ecology primarily concerns the latter, and assumes that deliberate burning reinforces natural processes.

3. Changing Holocene Relationships with the Environment

Regional population growth at the end of the Pleistocene sooner or later set in train a variety of adaptive responses to new or changing environmental opportunities. In some areas, increasingly skilful hunter–foragers turned to a more specialized exploitation of coastal resources, while others remained more conservative generalists, despite technological innovation. Elsewhere hunter–foragers chose to diversify their ecological niche by investing more time in the harvesting of large numbers of small mammals, at least on a seasonal basis. Tapping into a more predictable food supply, such strategies repaid the effort by allowing a more sedentary residential pattern, even if small hunting parties fanned out in pursuit of bigger game at opportune times. Longer-term residence at central encampments, in caves or out in the open, in turn allowed a more persistent exploitation and manipulation of diversified plant foods.

In several world regions this strategic shift led to the domestication of plants or animals, or both. This was

not a socioeconomic revolution, but a long trajectory of co-evolution, during which people experimented with and eventually ‘managed’ a selection of plant and animal resources. They favored and tended more desirable strains until new domesticated species emerged, to be maintained through deliberate planting and breeding. This inaugurated a novel set of relationships between people and the environment in that both plants and animals became artifacts with which people fundamentally changed their habitats, involuntarily or deliberately replacing native biota with cultivated plants and domesticated livestock.

3.1 The Impacts of Early Agriculture

Agricultural origins have been amply documented by multidisciplinary archaeological research in the Near East and in southern Mexico; similar processes are apparent, but imperfectly understood in Andean Peru, sub-Saharan Africa, and south-eastern Asia. Morphological (‘hard’) evidence for domesticated grasses and legumes is available for sites along the Near Eastern steppe-woodland ecotone by 10,500 BP, implying that the process of experimentation and manipulation began several millennia earlier. Agricultural origins in the New World may have been even more attenuated, but that database is spotty.

As long as communities practicing complementary agriculture remained few, small, and confined to the natural habitats of their cultivated plants and livestock, environmental impact would be minor. But these new pursuits were contagious across the Near East and much of Europe, introducing new plants or animals to environments where indigenous biota were displaced. Agricultural villages appear on the coasts of Greece, Crete, and Cyprus by 8,000 BP, and a distinctive new agrosystem had emerged around the coastal plains of the western Mediterranean about 7,000 BP. In some areas, resident hunter–foragers domesticated local wild cattle prior to agricultural penetration (Yugoslavia), while elsewhere expanding agriculturalists allowed outbreeding of livestock or cultivars (Hungary), to produce new domesticated variants. Dispersal across western Europe before 5,000 BP proceeded in stages, each with a probable sequence of ecological readjustment in new settings, indigenous innovations, interchange between resident hunter–foragers and intrusive farmers, and eventual indigenization of agricultural traits.

This early, Neolithic agriculture was localized and often ephemeral, and most pollen evidence shows that woodlands soon regenerated. But strong counter-examples verify repetitive fire clearance, cultivation, and stockgrazing that converted extensive areas into grass, weeds, or secondary vegetation 6,000–5,000 BP (Poland) (van Zeist et al. 1991). In parts of the Mediterranean Basin, early Neolithic pastoralists probably used fire clearance to open up pasturage in native woodlands as early as 7,000 BP (Greece,

Dalmatia, eastern Spain), with reforestation delayed up to five centuries. Such impacts were local rather than general, but in some areas they may have initiated or reinforced a trend to more xeric forest types. Yet settlement was spatially discontinuous, even in ‘preferred’ environments, until about 4,000 BP, judging by repeated episodes of woodland recovery and presumed local abandonment (Butzer 1996).

3.2 Plow Agriculture, Pastoralism, and Ecological Change

Towards 3,500 BC (5,400 BP) the invention of the plow (Mesopotamia) marked a significant transformation in the complexity and productivity of agriculture. The plow required a draft animal and opened larger fields to cultivation, while also exposing soil to potential erosion. Animals no longer served only as meat on the hoof. Cows, sheep, and goats were milked, and the milk converted to butter or cheese. Wool was skeined and woven into cloth as new breeds of woolly sheep were selected. The fertilizing properties of animal manure were probably recognized, as animals and their products became interlinked systemically with the range of subsistence activities.

More work was invested in a plot of land, including preparation, planting, weeding, and harvesting, to increase productivity. Plowing along hillsides deliberately or inadvertently created stepped slope profiles that may have suggested the principle of terracing. Swamps were drained and heathlands ameliorated with transported soils. In arid environments, opportunistic local manipulation of runoff developed into more complex methods of water control, including canal irrigation. Orchard crops, olive oil, and wine emerged as commercial products. Exchange networks and growing urban markets further reduced risk and, spurred by metals, metal products, and pottery, began to favor economic integration over wider areas.

Particularly in the Mediterranean Basin, much land was unarable, but animals were mobile and could be moved to rough upland pastures when the snows melted or lowlands became subject to summer drought. Short and long distance mobility patterns (transhumance) emerged that brought stony uplands into productivity. Controlled burning began to be used to manage such pastures on a regular basis, keeping woodlands open (but not deforested) and burning back woody or thorny shrubs. By the end of the Bronze Age (1,200 BC), much of the Mediterranean world was used and modified by farmers and herders, even though vast tracts of montane forest remained. A familiar cultural landscape was in place (Levant, Greece), expanding to the far west by Roman times, and bringing a simplified version to higher latitudes during the Middle Ages (van Zeist et al. 1991).

Intensified agriculture represented a quantum jump in production, demography, and interregional

networking, in managed environments that were increasingly artificial. Native biota were replaced progressively, with cultural controls over natural selection, succession, and diversity. To maintain such managed ecosystems required constant or increasing investment of labor, in order to maintain artificial equilibrium conditions and avoid ecological degradation (simplification). This was the case in the Bronze Age Mediterranean Basin and Near East, as well as China. A millennium later, similar transformations are apparent in parts of Mesoamerica, the Andean world, and several nuclei in southeastern Asia.

Changing human–environmental relations were not limited to agricultural regions. In the central USA sedentary utilization of rich, riparian resources incorporated controlled use of a variety of minor plant foods well before the appearance of standard domestication. In arid lands and the sub-Arctic, or in the high Andes, mobile herding patterns emerged that were based on new domesticated genera such as camel, reindeer, and llama. Even in the Arctic, groups of hunter–foragers devised new collaborative methods to effectively exploit coastal resources, by mobile strategies such as those of the evolving ‘Eskimo’ cultures (Alaska to Greenland). A mastery of the environment is broadly evident that was qualitatively and quantitatively different from that of the Late Pleistocene.

With increasing environmental exploitation by expanding populations, the potential for ecological impairment or degradation was rapidly enhanced. But the fact that discernible damage did not keep pace suggests that farmers built on the cumulative experience of trial and error in familiar environments, to articulate more conservationist strategies (Butzer 1996). Ideally, land use seeks to minimize both long-term environmental damage and short-term subsistence risk. Pollen and geomorphologic records show that for the most part, Mediterranean farmers and herders were successful in doing so during the last four millennia, despite sporadic ‘accidents’ of mismanagement in some areas. Only during Late Medieval times, when population was substantially greater than in Classical times, was supplementary food imported from outside the Mediterranean region. Until the early 1900s, productivity was maintained or enhanced, which implies sustainability across 7000 years of agricultural land use (Butzer 1996).

An anomalous episode of disturbance is dated to the transition of the Bronze and Iron Ages, c. 1,000 BC. It saw intense and often prolonged vegetation disruption (Spain, northern Greece, Northwest Turkey) by pastoralists moving into the Mediterranean environment from temperate Europe or the Balkans. Presumably they misapplied management methods better suited to humid environments. Much the same happened during the Middle Ages, after desert nomads moved into the Levant and North Africa, or Slav herders occupied northern Greece. In other cases initial agricultural occupation led to a flush of soil erosion (Greece), and

soil loss was locally problematic after pastoral expansion or agricultural decline, thus during the later first-millennium BC (Palestine, western Turkey, Greece, Italy, Spain) and again in Byzantine or Medieval times (Palestine, Greece, Italy, North Africa, western and central Europe). Although the cultural landscape and its humanized woodlands were repeatedly rehabilitated, soil thinning had to be compensated for with manure, but cumulatively this does not compare with watershed destruction by mechanized agriculture or industrialization since the 1950s.

In the New World, where the focus of recent discussion has been on the impacts of European-style land use, more pragmatic work shows extensive forest disturbance or removal, and even soil erosion, during indigenous demographic expansion many centuries before 1492 (Central America, Mexico, eastern United States) (Pohl et al. 1996, Butzer and Butzer 1997, Peacock 1998).

The human imprint on the landscape at the threshold of history was profound. In combination with the devastation of island ecosystems, such as New Zealand or Madagascar, it includes episodes of biotic impairment and even extinction, with enduring imprints. Yet most environments today, even ‘wilderness’ areas, are humanized to some extent and many if not most agricultural landscapes provide esthetic comfort and psychological anchoring in their particular cultural contexts. The human ascent to unprecedented demographic success and ecological dominance has impoverished floras and faunas the world over, often replacing them with vast monocultures or a handful of livestock species—even before the side-effects of industrialization could pollute the waters or turn machinery loose on the soil mantle. But until the dawn of a global consciousness, the transformations of the last 10 millennia were guided by community decisions, made through culture-specific lenses, with the goal of subsistence success across the generations, guided by imperfect information, despite conservative and conservationist strategies. That is the *reality* of human–environmental relationships, however they are judged today in the face of global ecological crisis.

See also: Ecology, Deep; Environmental and Resource Management; Environmental Cognition, Perception, and Attitudes; Human–Environment Relationships; Hunter–Gatherer Societies, Archaeology of; Hunting and Gathering Societies in Anthropology; Land Degradation; Sustainable Development

Bibliography

- Anderson A, McGlone M 1992 Living on the edge: Prehistoric land and people in New Zealand. In: Dodson J (ed.) *The Naive Lands: Prehistory and Environmental Change in Australia and the South-west Pacific*. Longman Cheshire, Melbourne

- Bocquet-Appel J P, Demars P Y 2000 Population kinetics in the upper palaeolithic in Western Europe. *Journal of Archaeological Science* **27**: 551–70
- Butzer K W 1982 *Archaeology as Human Ecology: Method and Theory for a Contextual Approach*. Cambridge University Press, Cambridge, UK
- Butzer K W 1991 An old world perspective on potential mid-Wisconsinan settlement of the Americas. In: Dillehay J D, Meltzer D J (eds.) *The First Americans: Search and Research*. CRC Press, Boca Raton, FL
- Butzer K W 1996 Ecology in the long view: Settlement histories, agrosystemic strategies, and ecological performance. *Journal of Field Archaeology* **23**: 141–50
- Butzer K W, Butzer E K 1997 The ‘natural’ vegetation of the Mexican Bajío: Archival documentation of a 16th-century savanna environment. *Quaternary International* **43**(44): 161–72
- Dodson J, Fullagar R, Head L 1992 Dynamics of environment and people in the rested crescents of temperate Australia. In: Dodson J (ed.) *The Naive Lands: Prehistory and Environmental Change in Australia and the South-west Pacific*. Longman Cheshire, Melbourne, Victoria
- Graham R W, Lundelius E L, FAUNMAP 1996 Spatial response of mammals to late quaternary environmental fluctuations. *Science* **272**: 1601–6
- Grayson D K 1991 Late pleistocene mammalian extinction in North America: Taxonomy, chronology, and explanations. *Journal of World Prehistory* **5**(3): 193–231
- Kershaw A P, Bush M B, Hope G S, Weiss K F 1997 The contributions of humans to past biomass burning in the tropics. In: Clark J S, Cachier H, Goldammer J G, Stocks B (eds.) *Sediment Records of Biomass Burning and Global Change*. Springer, Berlin
- Klein R G 1999 *The Human Career: Human Biological and Cultural Origins*, 2nd edn. University of Chicago Press, Chicago
- Kohen J 1995 *Aboriginal Environmental Impacts*. University of New South Wales Press, Sydney
- Martin P S, Klein R G (eds.) 1984 *Quaternary Extinctions: A Prehistoric Revolution*. University of Arizona Press, Tucson, AZ
- Meadows M E 1984 Past and present environments of the Nyika Plateau, Malawi. *Palaeoecology of Africa* **16**: 353–90
- Peacock E 1998 Historical and applied perspectives on prehistoric land use in eastern North America. *Environment and History* **4**: 1–29
- Pohl M D, Pope K O, Jones J G, Jacob J S, Piperno D R, de France S D, Lentz D L, Gifford J A, Danforth M E, Josseland J K 1996 Early agriculture in the Maya lowlands. *Latin American Antiquity* **7**: 355–72
- Ross A, Donnelly T, Wasson R 1992 The peopling of the arid zone: Human–environment interactions. In: Dodson J (ed.) *The Naive Lands: Prehistory and Environmental Change in Australia and the South-west Pacific*. Longman Cheshire, Melbourne, Victoria
- Speth J D 1987 Early hominid subsistence strategies in seasonal habitats. *Journal of Archaeological Science* **14**: 13–29
- van Zeist W, Wasylikowa K, Behre K E (eds.) 1991 *Progress in Old World Paleobotany*. Balkema, Rotterdam
- Zogning A, Giresse P, Maley J, Gadel F 1997 The late holocene palaeoenvironment in the Lake Njupi area, West Cameroon: Implications regarding the history of Lake Nyos. *Journal of African Earth Sciences* **24**: 285–300

K. W. Butzer

Human–Environment Relationships

The relations between human activities and their biophysical milieu have never been a subject of deep and sustained interest to the Western social and behavioral sciences. Such attention as they have been given has been fitful and partial, and it has arguably always been much less than their importance merits. Environment—in the sense of the biophysical world—has not been a topic that the orthodox human sciences have felt comfortable in addressing. This article traces the general scope of human–environment relations and of the past and present study of them in the human sciences.

1. The Term ‘Environment’

‘Environment’ is a troublesome word in that it can mean quite different things in different contexts. In some fields, it has long been used to refer to the social surroundings of human existence and behavior. In that sense, it encompasses much of the standard subject matter of the social and behavioral sciences. Elsewhere, it denotes the biophysical surroundings of human activity. In that sense, it encompasses much of what the social and behavioral sciences have deliberately chosen not to study. An emphasis on ‘environment’ is associated confusingly in the one case with the rejection of ‘nature’ as an important factor (as in the ‘nature–nurture debate’) and in the other with its affirmation (as in ‘nature–society relations’). An ‘environmentalist,’ depending on the context, can be one who believes that social influences outweigh hereditary ones in human development (a frequent meaning of the term today in psychology and sociology); that human beings are molded by climate, terrain, and other elements of their biophysical surroundings (the meaning given the term in the 1968 *Encyclopedia of the Social Sciences*); or that vigorous action should be taken to slow or reverse human-induced change in earth systems (the usual meaning of the term in ordinary parlance today). Only the second and third of these meanings correspond to that of ‘environment’ taken, as here, in the biophysical rather than the social sense. Both of them, moreover, exemplify a second problem: the frequent usurpation of the term by a narrow and particular sense of it that obscures its full scope.

Even if ‘environment’ denotes not the totality of the biophysical universe, but only those portions of it that are relevant in some way to human activity, its boundaries are not easy to fix. They can and often do change as human activities change or as understanding of their relations with their surroundings does. At or near the boundaries, the environment can include the mineral deposits of the deep ocean floor (now that the technology for their exploitation exists), the strato-

spheric ozone layer (now that its role in protecting life on the earth's surface is recognized), the stars when they were used for navigation, and a comet or an asteroid threatening to hit the earth. If the outer limits are always uncertain and always in flux, however, the core elements are easily defined: climate, soil, terrain, bedrock, mineral and fossil fuel deposits, atmosphere, fresh and salt water, and biota. They are the basic elements that compose what the Russian geochemist Vernadsky (1929) called the biosphere: the zone in which life as we know it can exist and which life, in turn, can affect and transform.

To speak of nature or the biophysical world as the environment of human existence is at least to some degree to separate the two realms. Such a separation is often challenged and resisted. The question has long been debated whether humans are fundamentally a part of, or apart from, the natural world. The grounds on which they are asserted to, or not to, belong to nature are of the most varied sorts. Social arrangements can be described as given by nature in order to justify and 'naturalize' them and rebut criticism of their injustices. Nature, as in the concept of natural rights, can at the same time be used as a standard by which existing social arrangements are measured and found wanting. During the nineteenth century, Darwinian and Lamarckian theories of human origins tended to dissolve certain boundaries between humanity and nature by emphasizing the importance of the organism's evolution from its surroundings and its continuous shaping by the demands that these impose. They were and have continued to be adopted and employed for widely varying ends by both conservatives and radicals.

At the same time, the American geographer Marsh (1864), a contemporary of Darwin's, tended rather to emphasize the distinctions between the human and the natural. He insisted forcefully, and with a wealth of evidence, that human actions modifying the earth were distinct in kind rather than merely degree from natural processes of change. They were, Marsh pointed out, far more rapid and abrupt in occurrence and far more radical and destabilizing in their character. But the very scale and scope of human impact, which have both grown enormously even since Marsh's time, tend in another way to blur the boundaries between society and environment: by making the latter less and less of a distinct entity. Undoubtedly human impact has altered and even transformed many aspects of the earth. Human actions, both deliberately and incidentally, have reshaped them into what is sometimes called a second nature that can no longer be regarded as something quite separate from society. Many people now spend their lives in climates that, at the scale of the built environment, are controlled closely and deliberately; that locally and regionally have in many cases been altered by human actions though rarely by conscious human intent; and that globally may or may not

already bear the marks of greenhouse gas emissions from fossil-fuel burning and land-cover change. Little of the environment can be considered wholly 'natural' if that term means pristine and unaltered.

Much of the environment has not been transformed, however, and all of it, however much altered, remains distinct from the human realm in another sense. Human beings possess certain characteristics of reasoning and reflexivity different in kind or at least in very significant degree from the rest of the known world. No other species or biophysical agent has equivalent attributes allowing it to carry out equally elaborate projects affecting the environment, nor have the actions of any changed the globe nearly so drastically in so short a time as have those of the one species possessing them. That humanity and environment are in some important ways separate is implied by the very existence of the social and behavioral sciences (and the humanities) as core fields of study distinct from the natural sciences.

Yet their separation very often is challenged from the margins by two broad programs for reducing one to a branch of the other and asserting their unity in terms of both substance and appropriate methods of study. Extreme forms of biological determinism propose to account for human and social action as the product of heredity and evolutionary imperatives, environmental determinism as the product of the shaping forces of climate and other elements of the human habitat. Both seek to annex the social realm to the natural. Annexation of another kind is envisioned by the more radical approaches found in the social studies of science, which propose to treat the knowledge generated by the natural sciences not as a means of explaining social action and processes but as itself one of their products. Science is seen not as a privileged source of fact but as a social activity like any other and governed by the same rules and to be analyzed in the same terms. These perspectives offer valuable correctives to an unthinking separation between the human and the natural. Weaker forms of them—evolutionary psychology, for example, or the study of the social correlates of environmental attitudes and activism—have been fertile sources of challenging new findings and explanations. Thus far, at any rate, some basic distinction between humanity and the biophysical world has continued to be accepted as unavoidable in practice.

2. Human–Environment Relations

But even if environment and society are distinct realms, not so simply and unproblematically related as reductionists on either side would argue, they are closely interdependent and interacting ones. If the autonomy of the social from the natural sciences (or of the natural from the social) is justifiable, what is less so is the thoroughness with which the social sciences, by

and large, have cleansed themselves of contact with the biophysical world. They have, in general, ignored the fact that all human activities, though not reducible to the biophysical plane alone, nonetheless have a biophysical side to them. All human actions require energy and matter from the environment and release both back into it. All have certain prerequisites in the form of the basic necessities of human existence—air, food, water, a tolerable temperature range, and so on—that the environment must provide; and these only begin the list of significant relations between society and its biophysical milieu.

Broadly speaking, the environment can represent a source of services (a term now coming into use to replace the narrower one of ‘resources’). Such services include background benefits from the environment that are enjoyed even though they are not consciously and deliberately exploited as natural resources. Among these are the key life-support systems of the biosphere. Other environmental services range from the absorption of wastes to the provision of spiritual and esthetic satisfaction and natural amenities such as scenery and agreeable weather. The environment can equally serve as a source of disservices, ranging from the hazards of violent natural events such as storms, earthquakes, and epidemic and endemic diseases to disamenities and constraints on human activity. To the extent that the environment acts as a determinant of human qualities, its effects can again be classified into services (the formation of valuable qualities) and disservices.

But there is no such thing as a service or a disservice without a recipient thereof. If some aspects of the environment are essential to the existence of any kind of human life as we know it—though not necessarily as it might exist in the future—whether and how any other plays either role depends crucially on the variables of what it is that people do and wish to do. Environmental features that help them do what they wish to do represent services, those that interfere represent disservices. Actors differentiated by location, occupation, class, status, gender, and any number of other variables will benefit or suffer differentially. The significance of the environment, then, is never determined solely by the characteristics of the environment; it is determined also by the ways of life that people lead within it.

As a result, a change on either side may be sufficient to change the relationship. It is by now clear that the environment on its own cannot be presumed to be stable or equilibrium seeking. Natural processes of change and fluctuation on many different scales of time and space are ubiquitous. Over time, they have increasingly been supplemented and, in some cases, overshadowed by human-induced changes, both those made consciously and those resulting unintentionally from actions taken. These may enhance or lessen the biophysical services and disservices received by different groups and individuals. Possible environmental

instability must be considered in the study of past environment–society interactions and in planning for future ones.

But environmental change, while it can be sufficient to alter environment–society relations, is not necessary for an alteration to take place. Equally sufficient can be shifts occurring on the social side. Changes in technology may mean that some aspects of the environment will cease to provide services or disservices, while others will begin to do so. Falling water was a vital source of industrial energy through direct-drive waterpower in the era before widespread fossil fuel use, and it became one again with the development of the means of hydroelectric power production. Uranium became an important resource only through the development of nuclear power generation. Some technologies are developed to alleviate existing problems in environment–society relations: those of weather monitoring and forecasting, for example. Others incidentally create new problems: the new role of clouds, fog, and downdrafts as hazards created by the development of aviation. Technological changes themselves are entangled closely with changes in social organization, which themselves may also alter society–environment relations. Such variables as the structure of ownership and exchange, household and gender relations, and political system, for example, all play an important mediating role between natural events and their possible human consequences. Famine, for example, may follow drought that drastically reduces the crop harvest, but a severe drought is neither a necessary nor a sufficient condition for famine. One of the few well-established regularities in the social sciences, documented by the economist Amartya Sen, is that no modern famine has occurred in a country, however poor, with well-functioning democratic institutions. They act, it would seem, to prompt effective government interventions preventing disaster. A certain environment, environmental feature, or environmental event, even if unchanged, will thus bear different relations to society at different times. Changes in the environment itself, of course, whether themselves natural or human-induced, can also alter the role of the environment in the lives of its inhabitants.

But an account of human–environment relations in these largely utilitarian terms remains only a partial one. It excludes, most notably, questions about environmental ethics or the proper relations between humanity and environment. The social sciences have found it difficult enough to address issues of values where only human actions are involved. Their particular focus has almost disabled them from assimilating or even recognizing arguments about the rights of other species and other features of the environment: the American biologist Leopold’s (1949) ‘land ethic,’ to take one example. Their usual tendency is to treat such beliefs—and understanding of the environment more generally—as expressions of social position

rather than to grapple with their possible validity. But the rise of Green parties and ecocentric groups and movements and of expressed support for notions of the rights of nature pose external challenges to the use of purely human-centered criteria, ones that have grown in urgency as human power to alter the environment, with likely effects including the widespread extinction of other species, has also grown.

3. *Human–Environment Relations in the Social Sciences*

These observations suggest the wide range that human–environment interactions cover. The actual range of interactions recognized and studied by the mainstream social sciences at any one time has always been much narrower. The most important early form in which environment impinged on the fledgling social sciences was that of environmental determinism. This doctrine offered one promising and plausible way of solving the formative puzzle of the social sciences: explaining why human behavior and institutions varied so widely from one place and people to another. The French political and social theorist Charles-Louis de Secondat, Baron de Montesquieu, in the eighteenth century treated climate and soil in *L'esprit des lois* as important factors in social life. He asserted, for example, that the people of cold or barren or mountainous areas tended to be vigorous and freedom-loving and those of warm, fertile plains to be more indolent and passive. Political and social institutions suited to one environment, he maintained, would be inappropriate to another. Many thinkers of the French and the Scottish Enlightenment echoed Montesquieu in seeking to explain differences between societies as at least in part growing out of differences in climate or other aspects of the physical environment.

But the concept of environmental influences soon changed from a useful part of the social theorist's conceptual baggage into an irksome burden. The shift proceeded in step with the organization of the social sciences into separate disciplines of sociology, anthropology, history, geography, economics, law, and so on. A few nineteenth-century historians—Henry T. Buckle in England, John W. Draper in the US, and V. O. Klyuchevsky in Russia among them—continued to assert the fundamental importance of environment in human affairs, but the trend of studies ran strongly in the opposite direction. Far more typical was the stance taken by Emile Durkheim in sociology. In his chief empirical work, he rejected meteorological and climatic explanations of patterns of suicide in favor of ones involving the character of social organization. At a more general level, he proposed that social facts be explained solely in terms of social causes. The approach excluded not only psychological or individual-level explanations in sociology, but also biological and physical-environmental ones. Not all of sociology was

Durkheimian, but the rejection of environment as a significant factor in social life was not one of its claims to be contested widely within the discipline. It was paralleled by the work of Franz Boas and others in anthropology who likewise insisted on the liberation of social inquiry from environmentalist as well as biological theories of cultural differentiation. Environmental determinism lasted longest in geography, but even there it faded under strong criticism from within and without the discipline.

Classical political economy from the outset recognized environmental conditions as determinants of scarcity and cost. It attributed the key phenomenon of land rent to differences in the inherent quality of the soil. Malthusian population theory presented the disparity between limited natural resources (the ability of the land to produce food) and potentially unlimited human wants as a basic condition of social life. The classical labor theory of value, however, tended to promote human activity over nature as the key source of wealth. Environment was demoted to a secondary role in another and more decisive way in neoclassical economics. Natural services were absorbed into the wider stock of capital and deprived of any distinctive qualities vis-à-vis human-produced goods. Natural capital and services were presumed to be perfectly replaceable by non-natural and the possibility of absolute scarcity of vital elements of nature neglected. Market mechanisms would ensure that nonrenewable resources would be depleted at the socially optimal rate and would also ensure that substitutes for them would be devised as necessary. These assumptions were questioned only by fringe schools of thought that had the standing of heresies vis-à-vis the orthodox consensus: theorists who proposed energy rather than monetary measures of wealth and efficiency, conservationists who warned of the exhaustion by waste of key resources not readily replaced by others.

In orthodox Marxist thought, not only the labor theory of value, but also the violent rejection of Malthusian notions of absolute scarcity as ideological fictions and the emphasis on class conflict and on a universal sequence of stages of evolution left environment little room to play a serious role. The chief exception was the much debated concept of distinctive 'hydraulic societies' expounded in the mid-twentieth century by Wittfogel (1957). Developing the earlier notion of the 'Asiatic mode of production,' he proposed that the formation of stable centralized despotisms was favored by the demands of large irrigation and water-control systems in dry climates—the characteristics of the environment thus playing an important role in social outcomes. But the idea only represented such a heresy within Marxism as energy economics did within the neoclassical discipline.

Elsewhere in the social sciences, determinism was for a time replaced by the opposed doctrine of possibilism, prefigured in the work of late nineteenth and early twentieth century French regional geo-

graphers led by Paul Vidal de la Blache and influentially expounded in the 1920s by the historian Febvre (1922). Possibilism pointed to the fact that different societies maintained quite different relations with similar environments as proof that the environment itself never determined any particular way of life. It offered, rather, a wide range of possibilities for habitation and use, some of which any particular society would be inclined to exploit and many of which it would not. In the postwar years, a possibilist approach was one of the many novelties informing the work of a school of historical research centered on the journal *Annales: Economies, Sociétés, Civilisations*. Close study of nature-society relations remained an important endeavor not only in *Annales* history, as it came to be known, but also in the subfield of anthropology and geography known as cultural or human ecology. It too was possibilistic in showing how society-environment relations, especially those of subsistence in nonindustrial settings, could be understood only by study of the characteristics of culture and habitat both. The very strength of environmental research in these precincts, however, only tended to reinforce the notion that environmental differences mattered and continue to matter chiefly for less advanced societies employing primitive technology and engaged in primary production.

That was not the only belief that tended to discourage wider attention to the environment. Another was the fear that taking the environment seriously led easily to environmental determinism and in turn to political implications of a necessarily reactionary sort. The fear, though excessive, gained plausibility both from some of environmentalism's earlier associations and from an easy analogy with biological determinism. According to the natural world any significant role in human affairs seemed to some, given the unevenness of natural patterns across the earth's surface, a step toward legitimating inequalities of power and position. Environmental determinism was most reliably expelled by expelling environment altogether.

It is understandable and by no means unfortunate that there should have been a strong reaction against a doctrine that had shown itself so intellectually barren and empirically untenable as environmental determinism. It is more regrettable that environmental relations of nondeterministic sorts should have come under the same ban. But their neglect was also a predictable result of the way in which academic research and training increasingly came to be organized. Environmental studies of any sort suffered from the deepening institutional and intellectual division of territory between the social and the natural sciences. Part of the price at which the hard-won autonomy of the social sciences came was the disenfranchisement or marginalization of subjects located in their new boundary zones. Part of the price of the increasing specialization of work within them was the increasing difficulty of acquiring an adequate training in both the

natural and the social elements of any research problem. The way in which the social sciences tended to be defined in opposition to the natural split even geography into human and physical and anthropology into cultural/social and physical branches. It was thereby made most difficult for any topic relating the two realms to be studied with any degree of sophistication.

The 1968 *Encyclopedia of the Social Sciences* appeared at a particular low point in social science attention to the biophysical environment, and the entries reflected it. The entry on 'Environmentalism' was a short account, largely historical, of environmental determinism, and the entry on 'Pollution' discussed not chemical contamination of air and water but ritual defilement in the purely cultural sense. Also included were a short piece on 'environment,' chiefly an exercise in definition (albeit a very useful one), and essays on 'conservation,' cultural ecology, and land and water resources. Biographical sketches of a few important figures such as Marsh and Malthus rounded out its treatment of the theme. Taken together, the entries conveyed no sense of the biophysical environment as an important topic for the social sciences.

But events were already taking place outside of the academic world that would again force environmental questions, at least in some form, on their attention. The environmental movement beginning in the 1960s and 1970s expressed a widespread public sense of concern over the impact of human activities on the earth. The list of new fears included ones as varied as forest damage from acid rain, desertification in arid and semi-arid lands, water supply shortages and water pollution, air pollution at various scales, the loss of biodiversity, and climate change at local, regional, and global scales. Activists called for intervention to protect the environment and combat the social forces—be they overpopulation, overconsumption, capitalist greed, socialist mismanagement, or any other—that they saw as responsible for the degradation and fraying of the biosphere.

Closely following the rise of public concern, attention to human-environment relations has re-emerged in the social sciences, but only in the narrow form of study of the human dimensions of human-induced environmental change. For some social scientists, it has been a characteristic and congenial task to examine environmental concern itself as a social movement having social causes, without necessary reference to any truth that the claims that it makes might possess. Such can be a fruitful path of analysis, one that asks how environmental problems are selected by certain groups and actors for attention among many others that are ignored. It is not difficult to show that public concern about many environmental issues substantially outruns understanding of the scientific basis for concern. Thus, the roots of concern may lie less in the objective situation (even if the scientific account of it is accepted as objective) than in other

areas that many researchers, particularly in Western Europe, have sought to identify. A second line of work has been much more visible and on the whole more generously supported by governments and other funding agencies. This form of attention to human–environment relations has largely accepted the idea that a global environmental crisis genuinely exists or threatens and has sought to shed light on its human dimensions: its causes, its consequences, and the ways by which it might be managed or mitigated. This approach, perhaps the dominant one in the US, evidently stems less from the traditions of social science research than from the pressures of external concerns.

Unfortunately, such research has still more often than not been conducted with the tools and within the assumptions and frameworks of the existing social science disciplines and subfields whether appropriate to environment or not. Orthodox social theory developed with little or no reference to environment is, thus, brought to environmental questions and used to recommend policy as if it required no modification to meet them. Yet its inadequacy to deal with many questions without being modified is plain. The long-standing neglect of biophysical issues of scale, scarcity, and irreversibility in economics has been challenged by the emergence of ecological economics, but to date with little apparent impact on the larger field. Sociological, anthropological, and political paradigms developed in other contexts tend to be applied indiscriminately to environmental questions without careful consideration of how applicable they are. The anthropocentrism of the social sciences as a whole makes it difficult if not impossible for them to deal with notions of the rights of nature and inherent value seriously. Most recently, many adherents of postmodernism in the social sciences have had to struggle with the labor of trying to reconcile their reluctance to acknowledge anything as objectively true with their belief that a genuine global environmental crisis exists and must be addressed.

The long separation between the study of society and the study of the earth has hindered the growth of theory indigenous to and appropriate to nature–society issues that does not oversimplify either side of the relation. The natural sciences have been no less reluctant than the social sciences to modify their favored theories and approaches in order to address the full range of links between society and environment. Where they have not ignored the human role, many have tended to fall back on a narrow and reductionist set of social theories to make sense of it. A characteristic example is the currency in natural-science discussions of environmental problems of the concept of ‘carrying capacity,’ the supposed maximum population that an area can support. It is a concept largely rejected as too mechanistic by social scientists who emphasize the culture-bound nature of the assumptions that must be made to define capacity in

particular cases. The view that human overpopulation and overconsumption are the key driving forces of environmental degradation is more widely held in the natural than in the social sciences, whose practitioners, for their part, are sometimes prone to suppose that the capacity of the environment to support life and activities is infinitely flexible or that the only serious limits are social. Many natural scientists also show an affinity for technical solutions to environmental problems—such as carbon sequestration, albedo modification, or a shift to nuclear power as a response to the threats of global climatic change—without adequate consideration of the crucial problems tied up with their social feasibility or public acceptance.

Human–environment relations may now be far more urgently important a research topic than ever, given the possible dimensions of rapid global environmental change and the measures that coping with it might demand. Yet a long period of neglect means that the knowledge base in this area is weak. Little is known about what environmental change might mean for society because little is known about what environment means for society. The present atmosphere of crisis does offer some hope that at last a detailed understanding of human–environment relations will begin to be constructed. Yet an adequate assessment of many environmental issues cannot be achieved from a background only in the natural or social sciences, and it remains to be seen whether the organization of education and research will prove capable of providing the necessary conditions to nurture such work.

See also: Cultural Geography; Environment and Development; Environmental Determinism; Environmental Policy; Environmentalism: Philosophical Aspects; Nature, Concepts of; Environmental and Ecological; Environmentalism: Preservation and Conservation; Ecology, Cultural; Ecology, Deep; Ecology, Political; Global Environmental Change: Human Dimensions; Cultural Landscape in Environmental Studies

Bibliography

- Arnold D 1996 *The Problem of Nature: Environment, Culture and European Expansion*. Blackwell, Oxford, UK
- Bates M 1968 Environment. In: Sills D L (ed.) *The International Encyclopedia of the Social Sciences*.
- Bramwell A 1989 *Ecology in the Twentieth Century: A History*. Yale University Press, New Haven, CT
- Clark S (ed.) 1999 *The Annales School* (4 Vols.). Routledge, New York
- Coates P 1998 *Nature in Western Attitudes since Ancient Times*. University of California Press, Berkeley and Los Angeles, CA
- Costanza R (ed.) 1991 *Ecological Economics: The Science and Management of Sustainability*. Columbia University Press, New York

- Cotgrove S 1982 *Catastrophe or Cornucopia*. Wiley, Chichester, UK
- Daily G C (ed.) 1997 *Nature's Services*. Island Press, Washington, DC
- Febvre L 1922 *La terre et l'évolution humaine*. Alcan, Paris
- Forde C D 1934 *Habitat, Economy and Society*. Methuen, London
- Glacken C J 1967 *Traces on the Rhodian Shore: Nature and Culture in Western Thought to the End of the Eighteenth Century*. University of California Press, Berkeley and Los Angeles, CA
- Guha R, Martinez-Alier J (eds.) 1997 *Varieties of Environmentalism: Essays North and South*. Earthscan, London
- Leopold A 1949 *A Sand County Almanac*. Oxford University Press, New York
- Marsh G P (1864) 1965 *Man and Nature; or, Physical Geography as Modified by Human Action*. Harvard University Press, Cambridge, MA
- Murphy R 1997 *Sociology and Nature: Social Action in Context*. Westview Press, Boulder, CO
- Nash R 1989 *The Rights of Nature: A History of Environmental Ethics*. University of Wisconsin Press, Madison, WI
- Rayner S, Malone E (eds.) 1998 *Human Choice and Climate Change* (4 Vols.). Battelle Institute Press, Columbus, OH
- Redclift M, Benton T (eds.) 1994 *Social Theory and the Global Environment*. Routledge, London
- Thomas F 1925 *The Environmental Basis of Society: A Study in the History of Sociological Thought*. Century, New York
- Thomas W L Jr (ed.) 1956 *Man's Role in Changing the Face of the Earth*. University of Chicago Press, Chicago
- Turner II B L et al. (eds.) 1990 *The Earth as Transformed by Human Action*. Cambridge University Press, New York
- Vernadsky V I 1929 *La biosphère*. Alcan, Paris (English trans. *The Biosphere* 1998)
- Wittfogel K 1957 *Oriental Despotism*. Yale University Press, New Haven, CT
- Worster D 1994 *Nature's Economy: A History of Ecological Ideas*, 2nd edn. Cambridge University Press, New York

W. B. Meyer

Humanistic Geography

Humanistic geography studies people's relationships to nature, space and time, their behavior, feelings and perspectives on environment. Sharing enquiry with other branches of the discipline, its domains of interest overlap strongly with the humanities and social sciences. Its scholarly value and main contribution to society consists in its ability to raise people's levels of awareness about taken-for-granted ways of life and thought, inviting critical reflection on humanity's relationships to the earth.

'Diversity is at the heart of geography—from the varieties in continents and climates to the interrelationships between natural resources and how people live. Diversity is also central to human rights.' Thus, Mary Robinson, UN High Commissioner for Human Rights, addressed delegates at the 29th International

Geographical Congress in Seoul, August 14–18, 2000, upon receipt of the *Earth and Humanity Medal 2000* from the International Geographical Union. She was honored for not only her championing of civil liberties and economic opportunity, but also for proclaiming access to intact and healthy environments as basic human rights. A few weeks later, scholars from all over the world assembled in Rome to explore the potential role of universities in the quest for a 'new humanism.' Renaissance humanism, one speaker recalled, 'set up a new idea of truth as a dynamic statement, not previously defined or constructed, but something to be discovered and then applied ... the joint venture of new scientific discovery with the humanistic approach to mankind's problems enabled the university to nourish innovation and offer a rigorous critique of institutions and social relations' (Bricall 2000). Such 'airing' of university life, he noted, was sadly lacking today. 'Humanistic geography' may have been a twentieth-century invention, but many of its central themes at the dawn of the twenty-first century resonate with centuries-old melodies.

1. Geography and Humanism

Humanists down the centuries have explored the nature of humanity, its passions and powers, while geographers have studied the earth where humans, among many other life forms, make a terrestrial home. For each facet of humanness—rationality or irrationality, faith, emotion, artistic genius, or political prowess—there is a geography: for each geographical interpretation of the inhabited earth there are implicit assumptions about the nature of humanness. To slogans such as Protagoras' 'man is the measure of all things' and Alexander Pope's 'the proper study of mankind is man,' geographers would note that *Homo sapiens* is a terrestrial species, with projects and plans played out in particular environments, even if diverse cultures have created and been inspired by myths and symbols. *Humanus*, literally, means 'earth dweller.'

The term humanism evokes at least four distinct connotations. First, it implies questions regarding the nature of humanness (*humanitas*) with concerns about individuality and sociality, freedom and responsibility, rationality and hedonism, conservatism and creativity. Second, it refers to humanist modes of knowing which oppose scientific reductionism, aiming to elucidate rather than to explain, emphasizing the subjectivity of consciousness and the need for intersubjective dialog. Third, it is expressed in the humanities—fields of learning in the arts, classical literature, and the cultivation of civic virtue. Fourth, it is evident in an enduring concern about the human condition, e.g. in humanitarianism, which has sought to encourage social responsibility and liberal politics. Historically, from whatsoever ideological stance it has emerged,

humanism has included a call for critical reflection on conventional ways of interpreting reality and a plea for integrity and authenticity in everyday life (von Wright 1978, Buttimer 1990).

Humanistic geography also reveals these four lines of concern. What is subsumed under this rubric varies from one country or language tradition to another, the terms 'social,' 'radical,' 'critical' and 'humanistic' often being seen as virtually interchangeable (Buttimer 1974, 1976, Tuan 1976, Relph 1981, Rowntree 1988, Pellegrini 1992, Claval 1999). For some the main goal is still to restore human subjectivity to a field where scientific objectivism has been so dominant, some emphasizing human attitudes and values, others cultural patrimony; some focusing on the aesthetics of landscape and architecture, others on the emotional significance of place in human identity (Ley and Samuels 1978, Buttimer and Seamon 1980). Several also advocate humanitarian concern and engagement in the resolution of social or environmental problems. Virtually all would agree with the formulation of Yi-fu Tuan that 'the humanist's competence lies in interpreting human experience in its ambiguity, ambivalence, and complexity' (Tuan 1976, p. 275). For over three decades humanistic geographers have explored the aesthetics of landscape, issues of memory and myth, moralities and imaginations, and the multifaceted textures of place (Adams et al. 2001). This 'humanist turn' has thus evoked sharper sensitivity to the inextricable connections between thought and context, experience and expertise, and the adequacy of geography to elucidate contemporary reality.

Some have looked askance on these developments as a retreat from problems or as simply a critique of the status quo (Entrikin 1976, Relph 1981). Anti-humanist strains have also been heard from doctrinaire structuralists, and staunch advocates of positivist spatial analysis. In the 1980s many, disillusioned about contradictions in Western legacies, condemned humanism as the arch culprit, a source of *hybris* and vaunting ambition, a myth overdue for dismissal. Postmodernist scenarios of texts reproducing texts in a labyrinth of self-reflecting mirrors precluded assumptions about human intentionality or meaning in lived reality. By the late 1980s there was a growing awareness of the extent to which the entire intellectual heritage of the West had shaped, and been shaped by, the peculiar social history of Western humanity. By century's end there was a notably increased interest in critical emancipatory elements which had previously remained marginal within the profession. These diverse currents pointed to weaknesses in the inherited lore and have opened up new worlds of exploration. And linked to this, there has also been a re-affirmation of geography as fundamental education for mature citizenship, for cross-cultural understanding and for environmentally sustainable ways of life. The 'new humanism' promises to be far more 'geographical' than its predecessors.

An enduring concern of humanistic geography has been the increased level of specialization among knowledge fields. While the narrowing of analytical focus has yielded sharper insight into particular processes, the overall fragmentation of research horizons has seriously diminished the geographer's capacity to deliver comprehensive perspectives on humanity and environment. In a so-called *Magna Carta* for the 'Humanities,' William Caxton (1422–91) insisted that attention be confined to mankind as an object of study in itself, not as part of 'nature,' which was the concern of biology, nor as an object of divine grace, which was the concern of theology. Hence emerged a clear separation of scientific and humanist modes of knowing. The former was to provide analytical guidelines for the analysis and representation of research results, while the latter was to invite discovery and understanding, processes which might not only vary from one individual to another, but also imply elements of emotion, aesthetics, and moral judgment which, strictly speaking, were to be eliminated from the scientist's agenda. At the dawn of the twenty-first century, there are resounding cries for liberation from such institutional dogmatism and inherited meta-discourses which obfuscate and distract attention from the far more tyrannizing meta-discourses of commercially-driven globalization.

An essential quality of humanist discourses in Western scholarly traditions, it could be argued, is a plea for freedom: freedom from bonds of ignorance, fear, oppression or oblivion, freedom to create and to explore new horizons in thought and life (Buttimer 1990). The emergence of a 'humanist' turn within geography during the latter half of the twentieth century might be described as a liberation cry for aspects of thought and experience which had been ignored or suppressed within the practices of previous generations. At the heart of humanist concern at the dawn of a new millennium is the patent failure of well-intentioned scientifically-based attempts to alleviate poverty, inequalities and injustices worldwide. Ironical indeed is the apparent paradox that, as volumes of environmental research have grown, so too has the magnitude of environmental problems. Questions now arise as to whether our taken-for-granted sectorally specialized kinds of science are adequate for elucidating these problems, or whether, in fact, taken-for-granted application of scientific results is part of the problem (Escobar 1997).

The twentieth century ended with a litany of 'post-' and 'anti-'movements in the scholarly world—emphasis placed on elements from which many sought to be free; horizons toward which a humanist spirit would aspire remained unclear. And yet there emerged a resounding chorus proclaiming plurality and diversity. Far from one orthodox brand of geographical knowledge, there are varieties of geographical knowledges—those based on everyday lived experiences, those of women, children, minorities, immigrants and

emigrants, professionals and clients, rich and poor. In fact, many of the major challenges facing humanity and environment today could be redefined in terms of competing, contested or conflicting geographical knowledges.

2. Humanistic Geography

Geography today is recognized as not only a professional field of expertise, an academic discipline, but also a reality of everyday lived experience. Humanistic geographers have consistently promoted attention to everyday geographical experience, not only for individual human persons, but also for the diverse cultural worlds whose home territories are increasingly invaded by globally organized technological and commercial systems. They advocate coresponsible engagement in matters of territorial and social discretion in humanity's modes of dwelling. Illustrations of contemporary themes for research and reflection in humanistic geography include (a) Dwelling, (b) Nature and Culture, (c) Mindscape and Landscape, and (d) Text, Image and Construction of Regional Knowledges.

2.1 Dwelling: Essential Quality of Humanness

Geography is the stuff of everyday experience for all forms of terrestrial life. Mountain and plain, river and lake, woodland and wildlife may be 'explainable' in the categories of natural science, but in reality each cultural group understands nature, space, and time through its own special filters. To survive on planet earth, all creatures develop a geographical sense, a sense of place, space, time, and movement. How to negotiate our diverse geographies is surely one of the enduring challenges of existence, from the politics of empire to the arrangement of one's kitchen, office, and garden. A geographer tells stories of dwelling, and the interplay of mindscapes and landscapes, in terrestrial environments.

The progressive 'humanization' of the earth over the past half-million years, offers fascinating ground for study and reflection (Glacken 1967). Humanity has assumed a dominant role in shaping landscapes, harvesting resources, tipping the natural checks and balances among other life forms, in culturally and technologically diverse ways of life. Each cultural group has devised its own variety of ways in which the bio-physical environment was understood and how humans should relate to it. Some, as in the Judaeo-Christian and Hellenic traditions, have placed humanity above all other life forms and have regarded intellect as the highest manifestation of human spirit. In Hindu and Buddhist traditions, humanity has been regarded as an integral part of cosmic order, recycling, like other life forms, in a continuous process

of reincarnation. Symbolic transformations are quintessentially human projects: the representation of reality in art and artifact, monument and metaphor, science, technology, and national flags. The human body itself has been a universal catalyst for symbol making in many civilizations, most common of which has been the organic analogy of the human person as a whole (body, soul, personality) as a microcosm of the universe.

Human modes of dwelling, with central themes of place and space, home and horizon, memory and identity, remain a primary focus of attention for humanistic geographers (Buttimer and Seamon 1980, Seamon and Mugerauer 1985). Emphasis is placed on the aesthetics and appropriateness of human settlements and the dramatic differences in built environments throughout the world. Humanistic geographers are skeptical of universal generalizations such as those which glorify or demonize humanity. Latecomers anyhow to earth history, some have argued, humans may now be en route toward self-destruction. While acknowledging some grounds for such gloomy prognoses, humanistic geographers continue to probe diverse landscapes, documenting evidence of human ingenuity in adapting lifeways to environment throughout history, and reiterating the perennial challenge of environment and ecology for global humanity today.

2.2 Nature and Culture

Throughout human history nature has constituted the source of inspiration and foundation for cosmology and human identity. Australian Aboriginals revere rocks as dreaming centers and song-lines evoke a sense of on-going creation. The *Upanishads* celebrate the eternal Asvatta (fig tree) and banyan tree, while North American Sioux build their entire cosmology around the sacred cottonwood tree. For East Asian societies, life was seen to emerge from the sacred lotus, and devout Hindus bathe in the sacred waters of the Ganges. Artefactual and cartographic evidence suggests that most cultures have imaged their own place as the center of the world, with horizons varying in shape and extent depending on their livelihoods and travel. The cargo cults of Polynesian islanders re-enact the stories of colonial occupation, Obijwe and Cree have detailed mental maps of trapping lines across the North American Plains—territory viewed by Euro-settlers as a vast undifferentiated arena for conquest by railway, lumber company, or settler (Brody 1981).

Humans everywhere have looked to nature for insight into the secrets of creation, birth and death, meaning and purpose in everyday life. Human inventions such as the domestication of animals and plants, technological innovations in agriculture, transport, and commerce have been construed and defended in

terms of cosmology and value system. The enduring salience of symbolism in shaping contemporary landscapes constitutes a central platform for 'new cultural geography' (Cosgrove and Martins 2000).

For the 350 million so-called indigenous peoples of the earth today, nature and human identity are inseparable. Multinational companies, often in league with national governments, find themselves confronted with a resounding *cri du coeur* from autochthonous populations eager to protect the integrity of their home environments and maintain discretion over minerals, forest, fish, and wildlife in their ancestral lands. The dramatic clash of human interests which characterizes the use of earth resources has at last found its way to public consciousness. The languages within which such clashes are overtly negotiated are characteristically those of technical expertise in the service of governments, corporations or transnational lobbies; the voices of *dwelling*, by people in place, are often ignored. Humanist geographers seek alternative bases on which human discourse about earth resources might be negotiated. Acutely aware of the extent to which scientific and popular notions of 'nature' have been socially constructed and applied in sometimes contradictory ways, they provide catalysts for critical reflection on taken-for-granted policy and programmes.

2.3 *Mindscape and Landscape*

The visible forms of landscape are the texts that geographers read and seek to interpret. To understand the humanized landscape one needs to understand cultural diversity among societies: not only their economies and traditional skills, but also the sense of sacred and profane in their perceptions of nature, space, and time. Successive groups have migrated over time from one area to another bringing their talents, images, hopes and prejudices with them. Characteristically, they try to recreate the new landscape in the image of the past, usually an idealized version of it. Contemporary landscapes have all been formed through the sedimentation of many layers of human occupation.

Among the strongest advocates of the 'humanistic' turn in geography were scholars who already acknowledged the hermeneutic value of landscape interpretation (Meinig 1979). 'Our human landscape,' Pierce Lewis remarked, 'is our unwitting autobiography, and all our cultural warts and blemishes, our ordinary day-to-day qualities are there for anybody who knows how to look for them' (Lewis 1985, p. 13). Landscape interpretation, adopting some of the models of literary criticism, 'read' landscape texts and signs as products in their own right, while also inviting enquiry into culturally varying modes of symbolic transformation in history. Debates over agency versus structure

in the social construction of landscape features became the *sine qua non* preamble to many research ventures of the 1980s and 1990s, as reflected in new scholarly journals such as *Ecumene* and *Géographie et Cultures*.

To communicate insight on the interactions of mindscape and landscape naturally poses critical issues of language, a challenge which geographers share with colleagues in literature, philosophy and art (Pocock 1981). It has surely been on questions of language and power, semiotics and symbolism, that humanistic geographers found challenging moments of interaction with structuralists. Slogans of 'death to the human subject,' short-lived but militant, scoffed at notions of potential human creativity to transcend structural imperatives. The deconstruction of inherited meanings, as suggested in the work of Derrida, could reveal the multiple and conflicting readings that could be made on particular texts. In the 1980's, post-structuralist discourses emphasized the multiplicity of potential interpretations and the undecidability of scientific propositions. In sharp contrast with the previous generation's deeply held convictions about epistemological certainty and images of the world as a mechanical system, many voices of the late 1980s were casting doubt on all scientific claims to objectivity.

2.4 *Text, Image and Construction of Regional Knowledge*

Humanistic geography is concerned not only with ways in which humans, in their everyday lifeworlds, have interpreted their surroundings. It is also concerned with images of the earth projected in academic (scientific) products (Dematteis 1985, Buttner et al. 1999). As with the earth's landscapes, the mindscapes of professional geographers have reflected the cultures and world-views of their sponsors and audiences. Geography's status as an academic discipline has mirrored the fluctuating fortunes of nations and empires, flowering at moments of societal self-confidence, altruism, or expansionary challenge; relaxing in routine operational tasks of housekeeping and inventory, text-book writing, and functionally specialized research in times of stability or depression. And as active participants in a global community of scholars, geographers have also negotiated their practices within the reigning scientific paradigms of particular periods.

Once national ministries of education established geography in university departments, the discipline's agenda became harmonized with that of its major sponsor, the state. To this day, with some notable exceptions, the practice of geography has borne a distinctively national character. From its beginnings as an academic discipline, geography has been valued as training ground for the exploration and conquest of space and resources; for the imposition of order

deemed rational by the conqueror; for the delivery of information on areas, distances, flora and fauna, peoples and cultures, in language categories and narrative frames which could be understandable and useable 'back home.' It is indeed ironic that while most places are subject today to influences which transcend territorial boundaries, the practice of geography is so tightly shaped by national and institutional contexts.

Humanistic geography is interested primarily in examining the geographical knowledges which have been promulgated in various settings. Starting from the accounts of geographers themselves, it examines the story of ideas in practice, the interplay of creativity and context in the practices of various periods, seeking to evaluate the implications of such practices for everyday life and landscape. Such approaches have not been popular, as Yi-fu Tuan warned in the mid-1970s: 'self-knowledge, the prime reward of the humanistic enterprise, has always been suspect in Western culture' (Tuan 1976, p. 276). While little evidence of its popularity in other cultures is available in geographical literature, it is certainly laudable that Yi-fu Tuan, indubitably the most influential writer of humanistic geography, has now shared his own autobiographical reflections on a fascinating, and possibly unique career journey (Tuan 1999). This account should prod readers to think critically about their own career journeys and taken-for-granted perspectives on life and world.

3. Conclusions

Each discovery in humanity's geographical understanding of the world has heralded changes in its self-images, hopes, and fears. The dawn of a new millennium reveals radically revised perspectives on mind-scape and landscape. Geographers have dared to look inwards, exploring cultural differences in environmental perceptions and experience, while colleagues in the humanities dare to look outwards, freely exploring themes such as landscape, identity and place, space and time. Today's world map of power and politics has taken on radically new dimensions due to transformations in technology and trade. The explosion of information and the transcontinental circulation of people, commodities, and ideas raise exciting new possibilities for thought and life. The spatial juxtaposition of culturally diverse peoples should become the springboard for new creativity in politics and social life, rather than problems to be solved by social engineering. Humanistic geography has much to offer by way of cross-cultural and historical evidence of creativity in discovering environmentally appropriate solutions. Gaia's human envelope, the anthroposphere, needs to be understood more as a complex drama than simply as battleground of ecological versus economic rationality; more as *oecumene*, pot-

ential home for humanity, a species which urgently needs to rediscover the art of dwelling.

Much rhetoric has been aired on the Western world's anthropocentric perspectives, critics usually implying alternatives such as the deification of nature, of the social collective, or the reification of ideologies playing out their purposes with humanity and nature. Perhaps the West, far from overestimating *humanus*, has grossly underestimated it. Recovery of the human subject, and recognition of human agency as an integral part of the lived world, the creativity again apparent on peripheral and previously marginalized regions, all betoken fresh potential for humanistic geography today. As a stance on life, humanism welcomes the challenge of discerning the creative potential of individuals and groups to deal with the surface of the earth in responsible and co-responsible ways (von Wright 1978). Nor is human creativity confined to the intellectual sphere; it involves emotion, aesthetics, memory, faith and will. The humanist turn in geography has delivered inspiration to practitioners of physical, economic, cultural and social geography. Its aims are not well served by investing energy on staking claims for becoming a special branch of the field. Humanism should thus more appropriately be considered as a leaven in the dough rather than as a separate loaf in the *smörgåsbord* of geographic endeavor. The emancipatory *élan* recuperable even from our Western traditions could enable geography itself to perform as leaven in the mass of contemporary science and humanities. The Renaissance of humanism calls for an ecumenical rather than a separatist spirit; it calls for excellence in special fields as well as a concern for the whole picture. From our Western corner, perhaps the greatest challenge is that of bridging worlds through dialog and mutual understanding.

Bibliography

- Adams P, Hoelscher S, Till K (eds.) 2001 *Textures of Place: Exploring Humanist Geographies*. University of Minnesota Press, Minneapolis, MN
- Bricall J 2000 Some recent trends in the universities. *Giubileo delle Università. Meeting with His Holiness John Paul II with the Academic Community* 9 September, 13–18
- Brody H 1981 *Maps and Dreams. Indians and the British Columbia Frontier*. Douglas and McIntyre, Toronto, Canada
- Buttimer A 1974 *Values in Geography*. Commission on College Geography Research Report No. 24. Washington, DC
- Buttimer A 1976 Grasping the dynamics of lifeworld. *Annals of the Association of American Geographers* 66: 277–92
- Buttimer A 1990 Geography, humanism and global change. *Annals of the Association of American Geographers* 80: 1–33
- Buttimer A 1993 *Geography and the Human Spirit*. Johns Hopkins University Press, Baltimore, MD
- Buttimer A, Seamon D (eds.) 1980 *The Human Experience of Space and Place*. Croom Helm, London
- Buttimer A, Brunn S D, Wardenga U (eds.) 1999 *Text and*

- Image. Social Construction of Regional Knowledges*. Institut für Länderkunde, Leipzig, Germany
- Claval P 1999 Qu'apporte l'approche culturelle à la géographie? *Géographie et Cultures* 31: 5–24
- Cosgrove D, Martins L M 2000 Millennial geographics. *Annals of the Association of American Geographers* 90: 97–103
- Dematteis G 1985 *Le metafore della Terra. La geografia umana tra mito e scienza*. Gianfranco Feltrinelli, Genoa, Italy
- Entikin N 1976 Contemporary humanism in geography. *Annals of the Association of American Geographers* 66: 615–32
- Escobar A 1997 Anthropology and development. *International Social Science Journal* 154: 97–115
- Glacken C J 1967 *Traces on the Rhodian Shore*. University of California Press, Berkeley, CA
- Lewis P 1985 Beyond description. *Annals of the Association of American Geographers* 75: 465–77
- Ley D, Samuels M (eds.) 1978 *Humanistic Geography: Prospects and Problems*. Maroufa Press, Chicago
- Meinig D 1979 *The Interpretation of Ordinary Landscapes*. Oxford University Press, New York
- Pellegrini G C (ed.) 1992 *Humanistic and Behavioural Geography in Italy*. Pacini Editore, Pisa, Italy
- Pocock D C D (ed.) 1981 *Humanistic Geography and Literature. Essays on the Experience of Place*. Croom Helm, London
- Ralph E 1981 *Rational Landscapes and Humanistic Geography*. Croom Helm, London
- Rowntree L 1988 Orthodoxy and new directions: Cultural/humanistic geography. *Progress in Human Geography* 12: 575–86
- Seamon D, Mugerauer R (eds.) 1985 *Dwelling, Place and Environment*. Kluwer Academic Publishers, Dordrecht, The Netherlands
- Tuan Y 1976 Humanistic geography. *Annals of the Association of American Geographers* 66: 266–76
- Tuan Y 1999 *Who am I? An Autobiography of Emotion, Mind and Spirit*. University of Wisconsin Press, Madison, WI
- von Wright J H 1978 *Humanismen som livshållning [Humanism as Stance on Life]*. Rabén & Sjögren, Stockholm, Sweden

A. Buttner

Humboldt, Wilhelm von (1767–1835)

1. Politics

When Wilhelm von Humboldt, born June 22, 1767 at Potsdam, resigned from his office as the Prussian Minister of Constitutional Affairs at the end of 1819, this was a major catastrophe for the political future of Germany. For Humboldt himself it was an act of liberation and a step towards what he regarded as his real vocation, the study of languages, to which he dedicated the rest of his life. Chateaubriand, the French poet and politician, considered this strange scientific activity as the spleen of somebody who had to kill his time as a consequence of the loss of political influence. But it was in fact the achievement of a

philosophical and anthropological quest which had started in Humboldt's youth. Studying the languages of the world was not at all something completely extraneous to his political work. The same motivation which would guide his linguistic work had also governed his political activities: it was a quest for the creative center of the human mind and for the achievement of individual felicity and autonomy within the totality of other individualities.

This is exactly what had created his political difficulties. At the end of his political career, Humboldt had elaborated a project for the constitutional transformation of Prussia: a proposal to establish a constitutional monarchy with a gradually developing representational system. The French Revolution, which he had welcomed as a decisive step in the history of humanity, had failed in Humboldt's eyes because of its rationalistic one-sidedness: the transformations of France had been governed too radically by abstract rational principles and had not taken into account historically grown entities. For Humboldt, the creation of the new should always be the result of a 'synthesis,' of a sympathetic union of two entities which, creating a new entity by their union, maintain their identity (the alternative—and lesser—methods of uniting two entities are 'isolation' and 'incorporation,' the former being only the juxtaposition of the two, the latter the brutal destruction of their integrity through their unity). Humboldt proposed a 'synthesis' of the traditional prerogatives of the monarch and the interests of the economically active classes (with a consideration also for the traditional rights of the nobility), a 'marriage' which would have transformed Prussia eventually into something new, a constitutional monarchy—and which would have had a model function for the rest of Germany. Not only for the inner organization of Prussia, but also for the whole of the area called 'Germany,' Humboldt projected a political structure according to the same principle of a 'synthetical' union: a confederation which would have left quite a lot of autonomy to its members but nevertheless have given important political functions to the whole (like warfare, freedom of movement and free trade within the union—very much like the USA).

But, in 1819, the times were not propitious anymore for reforms and for the creation of new political structures as after the defeat of Prussia by Napoleon which was the beginning of so many happy reforms in Prussia, like the reform of the educational system, the renewal of the municipal administration and the emancipation of the Jews. Now Napoleon was defeated and the old classes were striking back and were not willing to make any concessions to the bourgeoisie, to the memory of the 'people's war' or to the experience of freedom and civil rights imported by the Napoleonic armies. Humboldt was the representative of liberal political reforms, and, as such, he had to disappear—in the time of Austria's Metternich, of Hardenberg and a reactionary king in Prussia.

2. University and Bildung

But if Humboldt did not succeed in realizing his constitutional dreams, he seems to have had an overwhelming success with the establishment of the University of Berlin some years earlier. Humboldt is best and generally known for this activity, which in fact occupied him for little more than a year (from February 1809 to June 1810) and which was only a part of the reorganization of the whole Prussian educational system. Humboldt indeed made the decisive administrative step for the establishment of the University of Berlin. But what was established was only partially Humboldt's University. He planned the university as a state-independent, economically autonomous institution, with an independence that would guarantee the freedom of research and instruction. Therefore, Humboldt requested that the university should be given a donation of land as the economic basis of that autonomy. Instead, the Berlin University was from the very beginning a state establishment which—it is true—set unprecedented standards by the quality of its new faculty, by new forms of teaching, by the unity of research and teaching for professors as well as for students. The concept of the research university was successfully realized. As a state-dependent institution it would forge a well-trained loyal elite for the King of Prussia rather than free citizens of the realm of the spirit.

Humboldt's proposals for the establishment of the Berlin University and the other educational institutions became, however, seminal texts for a pedagogical theory whose central concept was that of *Bildung* (educational and cultural development). Following the convictions of German idealistic philosophy and the classical German poets (Schiller and Goethe), Humboldt wanted human individuals to be given every possible means to develop their physical and intellectual capacities. The educational institutions have to grant access to that formation of the human individual. The main aim of *Bildung* is not the acquisition of professional skills but the unfolding of the highest intellectual and artistic aspirations. Ancient Greece as the absolute summit of human development is the cultural reference of that educational program. The concept of *Bildung* certainly gave an ambitious aim to the educational institutions of Prussia and of Germany as a whole. This concept even served as a justification of otherwise unexplainable investments of the state in intellectual enterprises like the humanities, which have no practical ends. *Bildung* was an enormously successful concept as one of the central values of German identity, giving education a very high position in the value system of the nation. But just as Humboldt's University project was only partially realized in the Prussian University, the idea of *Bildung*—a generous and grandiose idea—did not so much create the (also politically) autonomous and free individuals Humboldt was thinking of, but had

more of a compensatory function: consoling the German bourgeoisie for its nonparticipation in political power. Hence, all that wonderful *Bildung* did not prevent the German catastrophe.

3. Anthropology

The establishment of the University of Berlin and *Bildung* secured Humboldt's place in the pantheon of the German nation. But his main activity and—in the long run—also his most successful intellectual achievement, was his linguistic and philosophical work. In this domain, however, Humboldt is not—as some historiographers tend to think—a founding father of the scientific project of historicocomparative linguistics that is linked to the names of Franz Bopp and Jacob Grimm. Bopp and Grimm invented the principles for a historical reconstruction of the unity of Indo-European, Germanic, Romance, Slavic, and other languages and for the comparison of historically related languages. This was to be the successful—romantic—project of the mainstream linguistics of the nineteenth century. Humboldt is, on the contrary, the founder of an alternative linguistic project, which is more the summary of the linguistic ideas of the Enlightenment: the study of the diversity of the languages of the world. *Verschiedenheit*, or diversity, is the key term of Humboldtian linguistics which was to flourish—at least in part—in the twentieth century, when the historicocomparative paradigm was relieved by synchronical descriptive linguistics.

Humboldt's linguistic project was integrated into a larger 'anthropological' project. 'Anthropology' in the eighteenth century was a term which designated the empirical study of human beings, as opposed to philosophy, which investigates the universal nature of the human race. Humboldt sketched a vast research program for the investigation of the cultural manifestations of humankind. Understanding the creativity of the human mind was the philosophical motive that gave Humboldt's anthropological project its special dynamics. Two events which occurred in his lifetime were decisive for this research into creativity: first, as for any young intellectual of his generation, the French Revolution, the creation of a completely new, enlightened, political organization. Hence, one of his first publications was an article on the first French constitution (1791) and for his friends he wrote a booklength political treatise 'On the limits of the activity of the state' (1792), which was printed only after his death and which became a founding document of political liberalism. The other revolutionary event in Humboldt's lifetime was Goethe. Here again, what fascinated the young intellectuals was the 'genius,' the mystery of an unprecedented creativity. Humboldt lived in Jena, as a friend of Schiller's, in close contact with Goethe. He studied Immanuel

Kant, whose philosophy was to become the framework of his own theoretical developments. Humboldt's work is centered around the Kantian concept of imagination, *Einbildungskraft*, the faculty of the human mind that is responsible for the creation of the new.

Since imagination as the mediation between the senses and the intellect had remained rather mysterious in Kant's philosophy, Humboldt set out to find its roots. He discovered them in the difference between the sexes. The opposition of the male and the female in nature is the very basis of every new creation, not only of physical generation (*Erzeugung*) but also of the most 'sublime creature' of the imagination: thought. The loving embrace of the female and the male generates new living beings in the physical as well as in the intellectual world. Kant was shocked by Humboldt's somewhat pre-Freudian interpretation of his philosophy. But despite Kant's criticism, sexual union, 'matrimony' (*Vermählung*), love, would be the structural model for everything Humboldt set out to investigate: art, politics, education, and language. Just as dialectics is the fundamental figure of thought in Hegel, the erotic encounter is the basic structure of Humboldt's thought.

After his reflections on the anthropological roots of creativity, Humboldt wrote a book on Goethe (1798) which contains a very modern theory of aesthetics: art is not only a creation of the artist's imagination but also a (language-like) communicative event between the producer of the work of art and the observers, the readers, the listeners who recreate the work of art in their act of reception and thereby accomplish the creative process.

4. Language

While still in Jena, Humboldt wrote his first reflections on language: 'On thinking and speaking' (1795–6), a short sketch which already contained the main features of his language theory. Humboldt's sojourn in France and in Spain (1797–1801) would be decisive for his making language the very center of his intellectual endeavours. In Paris, he discussed philosophical issues with the *Idéologues*, in whose philosophy language played a crucial role, and here and in his journey to the Basque Country, he encountered the Basque language whose profoundly divergent structural properties intrigued Humboldt. Later, his brother Alexander would bring linguistic material from America, and in Rome Humboldt had access to the linguistic materials of the Jesuits (Lorenzo Hervás). For years, the Amerindian languages were Humboldt's empirical linguistic objects before he turned to the Malayopolynesian languages to which his main—posthumous—work *On the Kavi Language on the Island of Java* (1836–9) is devoted. The introductory volume of that book, *The Diversity of the Human Language-*

Structure, completed before his death on April 8, 1835 in Tegel (near Berlin), is Humboldt's major work. With language, Humboldt arrived at the very heart of his general quest for the creation of the new by that mysterious Kantian faculty of *Einbildungskraft*.

Language is not just a means for the communication of thought, it is the very creation of thought. According to the Kantian system, the senses and the intellect create the concept through their synthesis. But for Humboldt, the concept is not created independently of the word: word and concept form an undissoluble synthetical unity; thought is created as sound. Language is the 'formative organ of thought' (*das bildende Organ des Gedanken*, 1903–36, VII, p. 53), 'labor of the mind' (*Arbeit des Geistes*, 1903–36, VII, p. 46). Like art, language (thought-sound) also needs the dimension of the other: it must be heard and understood. Language as a cognitive activity reckons with the other, language is thought in the dimension of the other: *Mitdenken*, or cocognition. But surpassing art, language is a reciprocal creative activity: the hearer becomes speaker, 'the word formed by me resounds from the mouth of the other' (1903–36, VII, p. 56). This ultimate reciprocity is the completion of the linguistic syntheses. The duality of I and Thou is the 'fundamental type' of language.

Humboldt discovered that language is actually at the center of the Kantian system. Humboldt's language philosophy was a linguistic turn of transcendental philosophy. Since Bacon's discovery of the *idola fori*, European philosophy had become gradually aware of the fact that language is not just a device for the communication of thought but first and foremost a cognitive activity. This insight finds a coherent post-Kantian formulation in Humboldt.

The second important moment of that philosophy of language is the insight that the formation of thought, the linguistic 'transformation of the world into the property of the mind' (1903–36, IV, p. 420), always proceeds according to the special ways of historically particular communities, that is, according to particular languages. There is no such thing as language in general, language comes in languages. Thought generated by language, hence, is not uniform and universal but historically and culturally fragmented in different 'worldviews' (*Weltansichten*). The task of language, the generation of thought, is universal, but that task is realized in thousands of different ways in the different languages of humankind.

Now the new—or Leibnizian—moment of that insight is its positive interpretation: from Bacon onwards, philosophy has lamented over the particularity of worldviews in the different languages and wanted to get rid of those 'prejudices'; the semantics of natural languages is like 'a mist before our eyes' (Locke) which philosophical enlightenment must discard. Only Leibniz gave an optimistic twist to the discovery of the linguistically generated diversity of thought when he envisaged a study of all the languages

of the world in order to know the ‘marvellous variety of the operations of our mind.’ Herder took that idea from Leibniz, and Humboldt not only integrated it into the Kantian framework but also developed a huge program of descriptive linguistics.

5. Comparative Study of Languages

All the languages of the world were to be described according to what Humboldt calls their ‘inner coherence’ (*innerer Zusammenhang*), their particular structure. The languages of the world had mostly been grasped according to the traditional categories of Greek and Latin grammar. But the structural properties of a language can only be captured through its own categories. Humboldt envisaged an encyclopedia of the languages of the world, a new *Mithridates*. He started a whole series of descriptive works on different languages of the world of which only the *Mexican Grammar* (Humboldt 1994) has been published up to now. A second kind of structural studies aimed at a revised form of general grammar (which was also based on the Greco-Latin grammar): comparative investigations of linguistic categories throughout the languages of the world. Humboldt’s article on the dual (1827) is an example of that research. After the completion of these structural descriptions, research into the parental relationships of languages (genealogy) or into structural similarities among languages (typology, classification) can be undertaken.

What is often considered as Humboldt’s main contribution to linguistics, the so-called Humboldtian ‘typology,’ is neither central to his linguistic aims nor did Humboldt propose typology as an alternative to the dominant linguistic school, that is, to the historicocomparative studies of languages. On the contrary, in his last book, his unfinished work on Kavi, Humboldt eventually envisaged a comparative study of the Malayo-Polynesian languages which would follow Grimm’s and Bopp’s work on the Germanic and Indo-European languages. Humboldtian typology is a linguistic myth created by superficial readings of his texts. Humboldt was very sceptical about the classifications of the Schlegels, and eventually he explicitly refuted the legitimization of classifying languages, since languages are individuals and should be described as such.

The summit or keystone of Humboldtian linguistics is therefore the investigation of the ‘character’ of languages. This is just the opposite of classification and typology, of grouping similar entities together. It is a tentative attempt to grasp the very individuality of each individual language. Humboldt was not sure whether ‘characterization’ was still a scientific task. But whatever it is, it has to be tackled since it is the ultimate aim of linguistics to see how the human mind uses the instruments it has created, to investigate the very end of language: speech and discourse. The—necessary—structural description of a language is only

‘comparable to its dead skeleton’ (1903–36, VI, p. 147), the life of a language is discourse (*verbundene Rede*), a life giving ‘character’ to that language. Therefore, only languages that have a multifaceted literature yield the possibility of being described as having characters.

Nearly two centuries after Humboldt’s sketch of a huge linguistic program, it can be said that the structural description of human languages he envisaged was the main task and achievement of linguistics in the twentieth century. It is not by chance that Humboldt is present in the seminal theoreticians of modern linguistics (Saussure (see *Saussure, Ferdinand de (1857–1913)*), Bloomfield (see *Bloomfield, Leonard (1887–1949)*), Hjelmslev). The nineteenth century was not very Humboldtian, with its emphasis on diachronical research as well as in its naturalistic research methods. Humboldt’s ‘keystone’ of linguistics, the ‘characterization’ of languages, was attempted by some literary-minded linguists (e.g., Vossler), but was discredited by nationalistic and ideological interpretations of linguistic structures. Since the structural descriptions of the languages of the world have now virtually been completed, it is not astonishing that in actual linguistics typological attempts are again very important. The so-called Humboldtian typology is often quoted in that context. But it is—even more than the other famous Humboldtian invention, ‘Humboldt’s University’—rather a myth than a creature of the author. Without any foundation in Humboldt’s work is also the widespread conviction that Humboldt discovered fundamental principles of Chomskyan linguistics, like the ‘infinite use of finite means.’ Humboldtian linguistics as a whole is in direct opposition to that kind of linguistic research: it is a hermeneutical research into the cultural diversity of human languages in historical communities, preferentially through (literary) discourse, and not a naturalistic investigation of the structure of the universal human mind considered as a universal grammar.

See also: Creativity and Cognition; Education, History of; Educational Institutions, History of; Educational Philosophy: Historical Perspectives; Instructional Psychology; Language and Literature; Language and Thought: The Modern Whorfian Hypothesis; Language, Continental Philosophy of; Liberalism; Liberalism: Historical Aspects; Linguistic Anthropology; Linguistic Turn and Discourse Analysis in History; Linguistic Typology; Linguistics: Overview.

Bibliography

- Borsche T 1981 *Sprachansichten. Der Begriff der menschlichen Rede in der Sprachphilosophie Wilhelm von Humboldts*. Klett-Cotta, Stuttgart, Germany
Borsche T 1990 *Wilhelm von Humboldt*. Beck, Munich, Germany
Di Cesare D 1998 Einleitung. In: *Humboldt W von. Über die*

- Verschiedenheit des menschlichen Sprachbaues und ihren Einfluss auf die geistige Entwicklung des Menschengeschlechts.* Schöningh, Paderborn, Germany, pp. 9–141
- Haym R 1856 *Wilhelm von Humboldt. Lebensbild und Charakteristik.* Gaertner, Berlin
- Humboldt W von 1836–9 *Über die Kawi-Sprache auf der Insel Java*, 3 Vols. Akademie der Wissenschaften, Berlin
- Humboldt W von 1903–36. In: Leitzmann A et al. (eds.) *Gesammelte Schriften.* Behr, Berlin, 17 Vols
- Humboldt W von 1988 (trans. Heath P) *On Language. The Diversity of Human Language-Structure and its Influence on the Mental Development of Mankind.* Cambridge University Press, Cambridge, UK
- Humboldt W von 1994 *Mexicanische Grammatik*, Ringmacher M (ed.). Schöningh, Paderborn, Germany
- Menze C 1965 *Wilhelm von Humboldts Lehre und Bild vom Menschen.* Henn, Ratingen, Germany
- Meschonnic H (ed.) 1995 *La pensée dans la langue. Humboldt et après.* Presses Universitaires de Vincennes, Saint-Denis, France
- Müller-Vollmer K 1967 *Poesie und Einbildungskraft. Zur Dichtungstheorie Wilhelm von Humboldts.* Metzler, Stuttgart, Germany
- Müller-Vollmer K 1993 *Wilhelm von Humboldts Sprachwissenschaft. Ein kommentiertes Verzeichnis des sprachwissenschaftlichen Nachlasses.* Schöningh, Paderborn, Germany
- Quillien J 1991 *L'anthropologie philosophique de G. de Humboldt.* Presses Universitaires de Lille, Lille, France
- Scharf H W 1994 *Das Verfahren der Sprache. Humboldt gegen Chomsky.* Schöningh, Paderborn, Germany
- Sweet P R 1978/80 *Wilhelm von Humboldt. A Biography*, 2 vols. Ohio University Press, Columbus, OH
- Trabant J 1986 *Apeliotes oder Der Sinn der Sprache. Wilhelm von Humboldts Sprach-Bild.* Fink, Munich, Germany
- Trabant J 1990 *Traditionen Humboldts.* Suhrkamp, Frankfurt am Main, Germany

J. Trabant

Hume, David (1711–76)

David Hume, a leading figure in the Scottish Enlightenment, contributed significantly to the disciplines of psychology, history, political science, and economics—as well as to more theoretical discussions in ethics, aesthetics, epistemology, and philosophy of religion. He was a controversial figure in his own day because of his criticisms of both natural and revealed religion and because of the political opinions he was thought to express in his *History*. His skeptical arguments concerning induction are still discussed by philosophers interested in probability theory, and his account of the foundations of morality enters into current disputes about the source of moral obligation.

1. Life and Writings

Hume was born in Edinburgh on April 26, 1711. The family was of genteel but of modest means: His father, Joseph Home, an advocate and member of the minor

gentry of Scotland, died when he was 2 years old and his mother Katherine, who never remarried, brought up him, his sister, and older brother on their small estate of Ninewells in the Scottish Borders. Hume probably began taking classes at Edinburgh University in 1721 where he studied the Latin classics, moral philosophy, mathematics, natural philosophy, and possibly history. There is no record of his ever having taken a degree (which was not unusual at this period) and, in a short autobiography written just before his death, he notes that, while his family intended that he study law, his own interests were in ‘Philosophy and general Learning.’ He probably began planning his first philosophical work, *A Treatise of Human Nature*, in 1729 when he was about 18 years old.

Suffering from complaints which we would now consider to be largely psychosomatic, Hume decided in 1734 that he needed to pursue a more active life; he left Scotland for Bristol where he worked briefly for a merchant. After a few months he went to France, where he spent the next 3 years writing and studying. He went first to Reims, but soon settled at La Fleche in Anjou, in the vicinity of the Jesuit College where, more than 100 years earlier, René Descartes had been a student. The first two books of his *Treatise of Human Nature* (‘Of the Understanding’ and ‘Of the Passions’) were published in 1739, about a year after he returned to England, and the third (‘Of Morals’) was published in 1740 after he returned to Scotland. The *Treatise*, from which he hoped to gain literary fame, ‘fell dead-born from press,’ though it is taken by modern scholars as an important and seminal work. It was followed in 1741–2 by his very successful *Essays, Moral and Political* which were written to reach a more popular audience.

Hume unsuccessfully applied for university chairs at Edinburgh University in 1744–5 and at the University of Glasgow in 1751–2, where he had strong support from his friends (among them William Cullen) but was opposed by many of the clergy, who accused him of irreligion. While his *Treatise* contained no explicit discussions of religion, this was changed in the rewriting of its first book, which was published in 1748 under the name of *Philosophical Essays Concerning Human Understanding*. It contained his famous argument against rational belief in testimony for miracles, as well as the basis for his later attack on the design argument for the existence of God. Continuing to search for a career, Hume spent some time during the 1740s as secretary to General St. Clair and participated in a rather futile attack in 1746 on the coast of Brittany during the War of Austrian Succession. In 1747 he traveled with the General as his aide-de-camp on a Military embassy to the capitals of Britain’s allies, Austria and Sardinia.

In 1749, Hume returned once more to Scotland and began an immensely productive decade. In 1751 he moved to Edinburgh from Ninewells; he was to live

there much of the rest of his life as a leading figure of the Scottish Enlightenment. Hume was elected secretary of the Philosophical Society of Edinburgh—a group devoted to the study of science and technology. (He was later, in 1762, to present to the Society Benjamin Franklin’s paper describing the use of lightning rods.) In 1751 he also published his *Enquiry Concerning the Principles of Morals*, a reworking of the third book of his *Treatise of Human Nature*. His correspondence at this time also indicates that he had already written substantial portions of his posthumously published *Dialogues Concerning Natural Religion*. In 1752 Hume published his popular *Political Discourses* and began work on what was to become his most successful piece of writing, namely his multivolume *History of England*. He became Keeper of the Advocates’ Library (later incorporated into the National Library of Scotland) in January, which gave him direct access to a large collection of historical sources. The first volume of the *History*, concerned with the reigns of James I and Charles I, was published in 1754 and the remaining volumes between 1756 and 1761. In 1758, Hume published *Four Dissertations*, which included his *Natural History of Religion*, *Of the Passions*, *Of Tragedy*, and *Of the Standard of Taste*.

In 1763, after the end of Seven Years War, Hume was invited to go to France as Secretary to the Ambassador, Lord Hertford. In France, where his writings were admired, he was lionized both by the mighty and intellectuals. He developed close relations with many of the philosophes, especially d’Alembert. Hume was *Charge d’affaires* for about 4 months in 1765 after Hertford left Paris. He returned to England at the beginning of 1766 bringing with him Jean-Jacques Rousseau, on the advice of Parisian friends; Rousseau sought refuge in England from persecution on the Continent. Hume and other British friends procured a pension for Rousseau from the King. However, this whole affair turned into a disaster as a result of Rousseau’s paranoia. Rousseau denounced Hume as his enemy. Hume’s defense of his own conduct was published by d’Alembert and Suard in Paris in October of 1766. Hume’s last public appointment was as an Under Secretary of State from February 1767 to January 1768. He retired to Edinburgh where he continued to publish new editions of his works and died on August 25, 1776 as Britain’s foremost man of letters.

2. General Philosophical and Psychological Principles

Philosophers still disagree about the nature of Hume’s basic philosophical principles. It is agreed that he was a skeptic who argued throughout his writings that our basic commerce with the natural and social worlds is not founded on reason, but on the non-rational

processes of the imagination and passions. What is in dispute is the nature of his scepticism, and the degree to which he sought to radically alter the conceptions of the natural and social worlds as conceived by his contemporaries (Read and Richmond 2000). While early twentieth-century philosophers saw Hume as an early forerunner of positivism, this interpretation has been challenged by recent scholars (Wright 1983). The question concerning Hume’s adoption of conservative or liberal principles has also been a subject of dispute—though he claimed neutrality in discussing the political debates of his own day (Livingston 1998, Stewart 1992).

In his *Treatise of Human Nature* Hume set out to found all the other sciences on the science of human nature, and stressed that this science itself must be based on experiment and observation. Hume’s foundational science of human nature was a science of the human mind. The basic psychological principle that he sought to establish in his *Treatise* was that of the association of ideas and impressions through resemblance, contiguity in space and time, and causality. He followed Locke in rejecting innatism and holding that the elemental building blocks of our ideas must be based on experience—the impressions of external and internal experience; however, he argued against Locke that these sources cannot supply us with our common sense or scientific conceptions of reality. Rather, the latter arise through irrational processes of the imagination and habits, based on the principles of association.

In a famous analysis, Hume argued that the principle of scientific induction by which we infer from past instances to future ones is based on custom and habit, and is not susceptible to rational justification. He denied that we have any understanding of the principles by which the natural world operates; at the same time, he argued that we naturally believe in causation between successive events that we observe to be constantly conjoined.

Hume held that we have as much reason to believe in necessary causes in the human as in the physical world. He argued that ‘the conjunction between motives and voluntary actions is as regular and uniform, as that between the cause and effect in any part of nature’ (1999, p. 154). At the beginning of his *Treatise of Human Nature* he set out to apply the experimental method to man, as it had been applied by Galileo, Newton, and others in the previous century to physics. At the same time, he believed that there is a problem in carrying out controlled experiments in the human sciences, because ‘reflection and premeditation’ would disturb the natural processes of the mind (2000, p. 6). He proposed to base his experiments on an observation of human behavior as it manifests itself ‘in the common course of the world, by men’s behaviour in company, in affairs, and in their pleasures.’ While much of his data throughout his writings is taken from history, in Book 2 of the

Treatise, Hume made a series of observations of contemporary social behavior, testing his hypotheses concerning the causes of the passions of pride, shame, love, and hatred. He asked his reader to do thought experiments in which the antecedent circumstances were varied and the resulting passions were predicted. He regarded the subjects which we today call sociology and psychology as closely intertwined.

3. Historiography, Economics, and Politics

Hume was both a philosopher of history and a philosophical historian. He both articulated and applied principles for the assessment of historical sources. These principles are carefully laid out in his essay *Of Miracles* where he discusses the criteria we ought to use to judge the reliability of witnesses, as well as the need to balance those results against the intrinsic likelihood of the occurrence of an event (1999, p. 169ff). Following discussions of probability in John Locke and other early modern writers (Stewart in Stewart and Wright 1994, p. 171ff), he formulated criteria for the reliability of written or oral testimony: these included considerations such as consistency with other reports of the event, the determination of the character and expertise of the witnesses, the number of witnesses, and the self-interest of the witnesses in convincing others of the truth of their accounts. Hume held that we need to balance these considerations of the reliability of testimony against the absolute likelihood of the occurrence of a miracle. He held that in identifying an event as a miracle (i.e., a violation of a law of nature) we are already committed to the view that it has a 'proof' against its very occurrence. This claim is based on his argument that the only evidence we have for laws of nature is past regularity. He concludes that in the case of miracle reports the witnesses would have to be absolutely unimpeachable—a state of affairs that has never obtained. More generally, Hume argues that we must balance the testimony given by historical witnesses against the probability of the occurrence of any event. Thus, in his essay *Of the Populousness of Ancient Nations*, in *Essays Moral, Political and Literary* (1987, p. 377ff), he denies the claims of his ancient sources concerning enormous populations and armies of the ancient world. He argues, mainly on economic grounds, that the ancient nations, unlike modern ones, could not have supported such high populations as ancient writers claimed (D. Wootten, in Norton 1993, p. 281ff). In his *History of England*, Hume frequently appealed to general principles of human nature—those which he believed to be established by the newly emerging human sciences—to evaluate the credibility of his historical sources.

Hume is an important forerunner of classical economic theory. He argued that economic forces influence political structures, and that the development

of modern liberty and government by law had been due to the development of trade and industry. Indeed, in the *Wealth of Nations*, Adam Smith states that Hume was the first writer to reverse the accepted view that political liberty precedes economic development. Hume saw in the rise of the merchants and tradesmen a check to the despotism that existed in premodern society. He argued that it was in the interest of these classes to live by equal laws and to limit the power of the aristocracy and the monarch. His analysis was based not only on his study of the rise of parliament in the early seventeenth century that led to the English civil war, but also in comparisons of modern European society generally with that of the ancient and medieval worlds. While Hume accepted James Harrington's social determinist view of the development of modern society, his originality lay in his postulation of the commercialism as a key factor in limiting the absolute power of the monarch (D. Wootten, in Norton 1993, p. 281ff).

Hume regarded man's economic behavior as deeply rooted in his psychological makeup. He argued that there are three factors that constitute human happiness—action, pleasure, and indolence (*Of the Refinement of the Arts*, in *Essays Moral, Political and Literary* 1987, p. 268ff). He regarded industry as a key to the leading of a balanced life in which pleasure and indolence follow naturally after a period of engaging activity. But, above all, he claimed that there is no pleasure that feeds on itself so much as the desire of the merchant to see his fortune grow. The habit of industry stimulated by this desire creates a frugality that does not naturally exist in an agrarian society based on inherited wealth. The love of gain which marks the activity of the tradesman and merchant results in large sums of money which can be lent at interest to stimulate further economic growth (*Of Interest*, in *Essays Moral, Political and Literary* 1987, p. 301). Low interest rates will be maintained by continuing increase in wealth.

Hume also argued that 'the encrease and consumption of all the commodities, which serve the ornament and pleasure of life, are advantageous to society' (*Of the Refinement of the Arts*, in *Essays Moral, Political and Literary* 1987, p. 272) and that a consumer society encourages the development of hard-working persons whose labor can be turned to the service of the state in times of national need. Key to his analysis is the idea that habits of hard work motivated by self-interest strengthen the nation-state. Developing the ideas of Bernard Mandeville, Hume argued against those in his day who complained that wealth and self-indulgence of the individual enfeeble both the individual and the society as a whole.

While Hume argued that 'Politics may be reduced to a Science' (1987, p. 14ff), he was skeptical about the ability of theorists in his own day to arrive at more than a few general principles of political behavior that would always apply (*Of Civil Liberty*, in *Essays Moral*,

Political and Literary 1987, p. 87ff). Certainly, he thought that history has taught us some general political principles concerning the different political behaviors of men under the different forms of government such as monarchy, aristocracy, and democracy. For example, from the ancient world we have learned that nonrepresentative democracy results in constant ‘tumult and sedition.’ However, he also argued that, given the relative newness of civilization, and the potential effects of ‘education, customs, or principles’ on human beings, it is impossible to predict how political behavior might evolve under circumstances which have not yet existed. He noted that among a number of other false generalizations, Machiavelli had concluded that if a prince relies on a single advisor that advisor would overthrow his master. Hume acknowledged that the evidence of ancient history, as well as that of ‘the little disorderly principalities of Italy’ of Machiavelli’s own day, encouraged such a generalization; but he pointed out that it had been proved false by the experience of European hereditary monarchies of the seventeenth and eighteenth centuries. In general, while Hume thought that while there were a few generalizations that could be relied on in the science of politics, a cautious scholar needed to put principles forward as mere hypotheses to be modified by future circumstances.

Hume argued that political institutions evolve slowly, and that a necessary condition for civil liberty, namely the balance of power between parliament and the monarch, is the result of complex historical factors; liberty, he believed, is always under the threat of disappearing if one side or the other develops too much power. This caution, and his fear of the mob, made him a profoundly conservative thinker, especially toward the end of his life.

See also: Causation: Physical, Mental, and Social; Causes and Laws: Philosophical Aspects; Smith, Adam (1723–90)

Bibliography

- Hume D 1947 *Dialogues Concerning Natural Religion*, N K Smith (ed.) Bobbs-Merrill, Indianapolis, IN
Hume D 1983 *The History of England from the Invasion of Julius Caesar to the Revolution in 1688*. Liberty Classics, Indianapolis, IN
Hume D 1987 *Essays Moral, Political and Literary*. Liberty Press, Indianapolis, IN
Hume D 1995 *Four Dissertations*. Thoemmes Press, Bristol, UK
Hume D 1998 *An Enquiry Concerning the Principles of Morals*. Oxford University Press, Oxford, UK
Hume D 1999 *An Enquiry Concerning Human Understanding*. Oxford University Press, Oxford, UK
Hume D 2000 *A Treatise of Human Nature*, D F Norton, M J Norton (eds.) Oxford University Press, Oxford, UK
Livingston D W 1998 *Philosophical Melancholy and Delirium: Hume’s Pathology of Philosophy*. Chicago University Press, Chicago

- Mossner E C 1980 *The Life of David Hume*, 2nd edn. Clarendon Press, Oxford, UK
Norton D F (ed.) 1993 *The Cambridge Companion to Hume*. Cambridge University Press, Cambridge, UK
Read R, Richmond K A (eds.) 2000 *The New Hume Debate*. Routledge, London
Rotwein E 1955 Introduction. In: Rotwein E (ed.) *David Hume, Writings on Economics*. Nelson, London and University of Wisconsin Press, Madison, WI
Stewart J B 1992 *Opinion and Reform in Hume’s Political Philosophy*. Princeton University Press, Princeton, NJ
Stewart M A, Wright J P (eds.) 1995 *Hume and Hume’s Connexions*. Edinburgh University Press, Edinburgh, UK and Penn State University Press, University Park, PA
Wright J P 1983 *The Sceptical Realism of David Hume*. University of Minnesota Press, Minneapolis, MI

J. P. Wright

Hunger and Eating, Neural Basis of

Eating is a familiar behavior. In humans, it is strongly influenced by cultural, social, and experiential factors. Consequently, eating has been a subject of considerable interest to social and behavioral scientists. In mammals, including humans, food intake is also a regulatory behavior with the primary function of supporting the continuous energy demands of body tissues. When considered from this perspective, food intake is influenced by hunger, satiety, and the biologic mechanisms that couple eating with internal caloric supplies and a stable body weight. This article discusses the signals that are important in enabling the central nervous system to control food intake, and also describes mechanisms thought to integrate these signals.

1. Historical Perspective

Biologic and behavioral studies on the control of appetite and satiety have shifted their focus markedly during the past three decades. The earlier preoccupation with the ‘dual center’ model, whereby food intake was controlled by a lateral hypothalamic feeding center and a medial hypothalamic satiety center, was based largely on investigations in which the putative centers were destroyed by stereotaxic electrolytic lesions (Stellar 1954). This once-dominant idea has now been discarded by most investigators in light of subsequent observations that have led to reinterpretation of the initial results. Specifically, the prolonged period in which animals do not eat after lateral hypothalamic lesions has been reproduced by destroying the dopaminergic neurons that ascend through the diencephalon; this aphagia and other related behavioral impairments appear to reflect a

decrease in cerebral arousal that precludes all voluntary behaviors, not just feeding (Stricker and Zigmond 1976). Moreover, the syndrome of over-eating and obesity after medial hypothalamic lesions seems to derive primarily from the greatly increased level of insulin in the circulation associated with a decrease in sympathetic neural tone, as a result of which ingested calories are rapidly and preferentially stored via lipogenesis rather than being used gradually for ongoing metabolic demands (Friedman and Stricker 1976, Powley 1977).

More recent studies have emphasized metabolic and humoral factors in peripheral organs, in addition to the brain. An increasing number of investigations have related food intake to the functions of the stomach, liver, and adipose tissue, and have suggested an important role of peptide hormones, especially insulin, cholecystokinin (CCK), and leptin, in the control of food intake. In considering these issues, most contemporary investigators have abandoned the traditional depletion-repletion model, whereby a deficiency of caloric nutrients somehow triggers an appetite for food, the ingestion of which then leads to nutrient repletion and satiation. It is now clear that there are no true shortages of metabolic fuels under normal physiologic conditions. Instead, the key variable in controlling food intake seems to be the alteration in the source of the metabolic fuels used during each of the two phases of caloric homeostasis: during the absorptive phase metabolic fuels are supplied from recently consumed food that still is in the alimentary canal and being assimilated, whereas in the post-absorptive phase metabolic fuels are generated from accumulated triglyceride, glycogen, and protein stores derived from food consumed more remotely in time. Thus, according to this perspective the shift from satiety to hunger is not associated with an absolute metabolic deficit, but instead occurs when animals enter the postabsorptive state and must mobilize calories from tissue stores in order to maintain caloric homeostasis. Moreover, eating does not correct a true deficit or restore some aspect of tissue function to normal; rather, it simply re-establishes the pattern of an ample supply of usable calories from the gastrointestinal tract instead of from stores in adipose tissue, liver, or muscle (Woods and Stricker 1999).

2. Satiety Signals

Because the cells of animals cannot long survive depletion of calories, hunger cannot result from cellular nutrient shortages. Instead, the sensation of hunger apparently does not need to be stimulated but is always present, and animals are constantly stimulated to eat unless they have just eaten, in which case the inhibitory effects of satiety generated from the previous meal are still operative. In support of this

perspective are laboratory findings that (a) experimenters cannot easily precipitate eating in satiated rats but can do many things to stop eating by hungry rats, (b) the size of spontaneous meals eaten by laboratory rats is unpredictable, but however much the animals eat in a given meal, meal size predicts how long they wait until they eat again, and (c) food deprivation does not cause a shortage of caloric nutrients in the circulation. The point here is that food intake appears to be initiated by the gradual disappearance of inhibitory satiety stimuli that were generated by the previous meal, not by the gradual increase in an excitatory hunger stimulus (LeMagnen 1985).

2.1 Gastric Distension

Several signals appear to stimulate satiety after a meal. One results from gastric distension, detected by stretch receptors on the walls of the stomach; the greater the distension, the greater the inhibition of eating (McHugh and Moran 1985). This afferent signal projects to the caudal brainstem via the gastric branch of the vagus nerve; as expected, animals eat unusually large meals when the gastric vagus is cut or its projection sites in the brain are destroyed. Experiments in rats have revealed that an intestinal hormone secreted during the meal, CCK, potentiates the gastric vagal signal and produces a heightened sensation of satiety related to gastric distension. This signal is related to the volume of the gastric contents, not its nutrient contents, because there are no known detectors of calories in the stomach. Drugs that acutely block CCK receptors increase food intake during a meal, as if the animal were less aware of gastric distension signals (Smith and Gibbs 1992).

2.2 Ingestion of Calories

The ingestion of calories, detected in the liver and possibly in the intestines, provides a second satiety signal, which is sent to the brainstem by another branch of the vagus nerve. This stimulus appears to develop only in the presence of insulin, a pancreatic hormone secreted when food is consumed. Thus, satiety normally is not lost as the stomach empties because the gastric signal related to distension is replaced by the postgastric signal related to caloric delivery (except in animals with insulin-deficient diabetes mellitus).

2.3 Thirst and Nausea

Two other meal-related signals also are known to be potent inhibitors of food intake. One is thirst,

resulting, for example, from the presence of relatively large amounts of salt in the food. The other is nausea, which might result from the inadvertent intake of toxins in food. However, these two variables are not relevant to most meals. Receptors that detect dietary salt have been located in the hepatic portal vein, which receives ingested food soon after it has been absorbed from the intestines. Receptors for circulating toxins are located in the area postrema in the brainstem. These two signals, together with gastric distension, provide three noncaloric stimuli that have marked inhibitory effects on food intake (Stricker and Verbalis 1991).

2.4 Body Weight

The influence of body weight on food intake must also be included in any consideration of satiety signals because food intake and body weight have been shown to be closely related. Thus, following a period of food deprivation, food intake is increased until the lost body weight is restored. Similarly, following a period in which food has been consumed in excess, food intake is diminished until the body weight that has been gained is lost.

Changes in body weight largely reflect changes in the storage of calories in adipose tissue. These stores are diminished during food deprivation to provide the calories needed for cellular function, and they are replenished when food intake resumes. Thus, as ingested calories are siphoned off to the adipose tissue, their effects on the liver are diminished, thereby reducing the duration of this postgastric satiety signal. In addition, food empties from the stomach more rapidly, thereby reducing the duration of gastric satiety. The opposite effects occur when calorie stores in adipose tissue have been augmented during a period of excessive food intake; both gastric and postgastric satiety signals are prolonged because the rate at which calories are stored in the swollen adipocytes is decreased (Stricker 1984).

2.5 Leptin

Aside from this indirect influence of body weight on food intake, in recent years a direct inhibitory signal to the brain from adipose tissue has been discovered. This satiety stimulus is a blood-borne protein, leptin, that is secreted from adipose tissue in proportion to adipocyte size; the fatter the animal, the more leptin is secreted and the less inclination the animal has to eat. Leptin receptors have been found in the basal hypothalamus, and when they are stimulated food intake is reduced (Woods et al. 1998).

The discovery of leptin illuminates genetic obesity in two strains of mice. Mice of one strain are born with

leptin receptors in their brain but without the capacity to synthesize leptin; mice of the other strain are born without leptin receptors in their brain but with the capacity to synthesize leptin. As might be expected, an intravenous infusion of leptin reduces food intake in the first strain but not in the second.

3. Central Neural Systems

Food intake is a simple behavior influenced by a very complex array of stimuli and situational variables, the details of which remain to be fully illuminated. However, it is already clear that there are many levels at which the central nervous system exerts control over eating. For example, the spinal cord and brainstem influence all aspects of caloric homeostasis via the autonomic nervous system. The hypothalamus and limbic forebrain receive diverse signals related to the food being eaten and the degree of body adiposity, and integrate them with information about the taste of the food, the memory of that food and previous eating opportunities, competing motivations, and other aspects of the environment. Among the numerous neural controls of food intake, those located in the brainstem and hypothalamus are perhaps the best known, so this section will focus on these two brain areas.

3.1 Brainstem

Considerable evidence has accumulated that some control of food intake can be exerted by the brainstem. For example, it is well established that the nucleus of the solitary tract and the adjacent area postrema in the brainstem receive sensory fibers from gustatory receptors in the mouth and throat as well as afferent information from the stomach, intestines, and liver. Thus, these linked brainstem sites likely are the first locus at which relevant sensory input from the viscera is integrated with taste sensations in the central control of food intake. Other evidence derives from investigations of the decerebrate rat, in which all axonal connections between the caudal brainstem and the forebrain are severed at the midcollicular level in the midbrain (Grill and Kaplan 1990). The animal cannot seek food or initiate spontaneous meals but, like intact animals, it will reflexively swallow liquid food put directly into its mouth. Thus, when investigators deliver sucrose solution through an oral fistula, the decerebrate rat will swallow the sweet fluid, provided its stomach is empty, as after a period of food deprivation. However, it will not swallow water or saline when deprived of food and it will no longer swallow sucrose solution (but instead lets the administered fluid passively drip from its mouth) when its stomach is full, or after having received a systemic injection of CCK. Hence, sufficient control of food

intake for such appropriate responses exists entirely within the caudal brainstem, although decerebrate animals clearly are incapable of receiving input from forebrain sites that could modify their intake based upon learning or other experience.

Recent experimental observations provide additional support for an important role of the brainstem in the control of meal size. When the area postrema is ablated, rats eat much larger meals than normal, although total daily food intake is normal; that is, they compensate for their large meals by eating less frequently, indicating that long-term controls of food intake are not impaired. Similar effects are obtained in rats pretreated with capsaicin, a neurotoxin that destroys (among other things) most gastric afferent vagal axons and thereby eliminates the food intake-reducing effects of CCK, and therefore is assumed to damage the signal of gastric distension.

The food-related sensory information that enters the brainstem is relayed along prominent neural projections to the hypothalamus, amygdala, and other portions of the limbic system, as well as to the thalamus and gustatory cortex. Reciprocal neural connections to the brainstem from these rostral sites allow cognitive functions and emotions to influence the control of eating. Linked elements participate in the central control of the gastrointestinal tract, the liver, and autonomic function generally.

3.2 Hypothalamus

Complementing these controls of meal size are the influence on meal frequency exerted by body weight. As mentioned, changes in body fat stores are reflected in the levels of leptin in the circulation. Receptors for leptin are concentrated in the *arcuate nucleus*, which is located in the ventrobasal portion of the hypothalamus. Receptors actually have been associated with two different types of neurons in the arcuate nucleus. One, located in the medial arcuate nucleus, uses *neuropeptide Y* (NPY) as a neurotransmitter. These cells are of interest because it is well-established that injection of NPY into the brain of rats evokes eating, and that chronic infusion of NPY into the brain leads to obesity. Furthermore, axons of NPY-synthesizing neurons project from the hypothalamic arcuate nucleus to the same focal site in the lateral hypothalamus at which eating is elicited by direct administration of NPY. Food deprivation stimulates activity in these NPY-containing neurons, whereas obesity or systemic administration of leptin depress their activity (Woods et al. 1998).

The other group of cells, located in the lateral arcuate nucleus, contain pro-opiomelanocortin (POMC). Those cells also have leptin receptors but, unlike the NPY-containing cells, the POMC-containing cells increase their activity when stimulated by leptin. The POMC-containing neurons use α -melan-

ocyte stimulating hormone (α -MSH) as a neurotransmitter, and activate melanocortin receptors on lateral hypothalamic cells to decrease their activity and thereby inhibit eating. As might be expected, food deprivation decreases activity in the POMC-containing neurons, whereas their activity is elevated when rats are obese and circulating leptin levels are high. In short, there is a dual control by the arcuate nucleus over lateral hypothalamic activity and its control of food intake, one neural input (NPY) stimulating feeding and the other (α -MSH) inhibiting it; this influence is normally stimulated by alterations in blood-borne leptin reflecting changes in body adiposity (Elmqvist et al. 1999).

The noncaloric signals (i.e., gastric distension, thirst, nausea) all have been shown to stimulate oxytocinergic neurons in the paraventricular nucleus of the hypothalamus (Stricker and Verbalis 1991). Projections from NPY- and POMC-containing cells in the arcuate nucleus also terminate in the paraventricular nucleus, allowing the integration of these various stimuli so as to influence output systems that control autonomic function and, indirectly, food intake. The central neural circuits that integrate the peripheral signals that control meal size (i.e., signals associated with the gastric, intestinal, and hepatic stimuli generated by a meal) with those that control meal frequency (i.e., signals associated with blood-borne leptin and perhaps other signals reflecting adiposity) are also currently under investigation.

See also: Body, History of; Body Image and Gender; Eating Disorders: Anorexia Nervosa, Bulimia Nervosa, and Binge Eating Disorder; Eating Disorders, Determinants of: Genetic Aspects; Health and Mortality, Body Stature as a Measure of; Obesity and Eating Disorders: Psychiatric; Obesity, Behavioral Treatment of; Obesity, Determinants of: Genetic Aspects; Weight Regulation: Psychological Aspects

Bibliography

- Elmqvist J K, Elias C F, Saper C B 1999 From lesions to leptin: Hypothalamic control of food intake and body weight. *Neuron* **22**: 221–32
- Friedman M I, Stricker E M 1976 The physiological psychology of hunger: Physiological perspective. *Psychological Review* **83**: 409–31
- Grill H J, Kaplan J M 1990 Caudal brainstem participates in the distributed neural control of feeding. In: Stricker E M (ed.) *Handbook of Behavioral Neurobiology*. Plenum Press, New York, Vol. 10, Chap. 6, pp. 125–49
- LeMagnen J 1985 *Hunger*. Cambridge University Press, London
- McHugh P R, Moran T H 1985 The stomach: A conception of its dynamic role in satiety. *Progress in Psychobiology and Physiological Psychology* **11**: 197–232
- Powley T L 1977 The ventromedial hypothalamic syndrome, satiety, and a cephalic phase hypothesis. *Psychological Review* **84**: 89–126

- Smith G P, Gibbs J 1992 The development and proof of the CCK hypothesis of satiety. In: Dourish C T, Cooper J S, Iversen S D, Iversen L L (eds.) *Multiple Cholecystokinin Receptors in the CNS*. Oxford University Press, London, pp. 166–82
- Stellar E 1954 The physiology of motivation. *Psychological Review* **61**: 5–22
- Stricker E M 1984 Biological bases of hunger and satiety: Therapeutic implications. *Nutrition Reviews* **242**: 333–40
- Stricker E M, Verbalis J G 1991 Caloric and noncaloric controls of food intake. *Brain Research Bulletin* **27**: 299–303
- Stricker E M, Zigmond M J 1976 Recovery of function following damage to central catecholamine-containing neurons: A neurochemical model for the lateral hypothalamic syndrome. *Progress in Psychobiology and Physiological Psychology* **6**: 121–88
- Woods S C, Seeley R J, Porte D Jr, Schwartz M W 1998 Signals that regulate food intake and energy homeostasis. *Science* **280**: 1378–82
- Woods S C, Stricker E M 1999 Central control of food intake. In: Zigmond M J, Bloom F E, Landis S C, Roberts J L, Squire L R (eds.) *Fundamental Neuroscience*. Academic Press, San Diego, CA, Chap. 41, pp. 1091–109

E. M. Stricker

Hunter–Gatherer Societies, Archaeology of

Hunter–gatherer societies traditionally have been regarded as representing the ‘original’ human society—that is, preagricultural peoples reliant upon ‘wild’ foods and interacting with ‘natural’ landscapes. In this way, they have been differentiated from other, agricultural, societies. From a sociological perspective, hunter–gatherers emerged as the ‘precapitalist’ baseline in Marx’s evolutionary framework—functioning as the original ‘communal’ society, practicing an economy different from those of others (Lee 1988). From a biological perspective, Darwin, among others, envisaged them as the human prototype, a position they still hold in more recent biological–ecological frameworks (Bettinger 1991). Anthropology and archaeology have absorbed these viewpoints, but more recently there has developed a significant change in emphasis—a redefinition of hunter–gatherers incorporating a much wider range in societal variation (Ingold et al. 1988).

Narrower, more traditional definitions had emphasized the following features. Hunter–gatherers had been perceived as small-scale, essentially egalitarian societies, of low population size and density and generally rather mobile. These societies were seen largely as *avoiding* accruing an economic surplus, practicing storage and impacting significantly upon

land and resources. In contrast, recent revisions consider that hunter–gatherers also incorporated more complex societal formations within their ranks. These included extensive social networks and signs of social hierarchy—as well as chiefdoms based on hereditary ranked lineages. Demographics included relatively high population sizes and densities, heightened territorial behaviour, and levels of sedentism. Hunter–gatherer economies can be composed of significantly high levels of intensification achieved through economic surpluses and storage. It is envisaged that they had considerable impact upon land and resources via firing and other practices now recognized by recent ecological impact assessments (Williams and Hunn 1982).

Ethnographic studies of the northwest coast peoples of North America (from northern California to British Columbia) included many of the characteristics described above, some living in permanent wooden-housed villages of up to 1,000 people. These hunter–gatherer peoples placed a strong emphasis upon fishing and practiced no agriculture (Suttles 1968, Ames 1985). The Calusa of Florida, another North American ethnographic example, reflect all these features and more—including a centralized state formation which extracted tribute from subservient villages, a royal family, and a standing army. The Calusa also were without agriculture (Marquardt 1988). Recently the ethnographies of more classic hunter–gatherers have also been considerably revised (Schrire 1984, Lourandos 1997). In all, this evidence points to the possibility of a similarly broad range of pre- and nonagricultural peoples having existed in the past, when most of the habitable environments of the world were occupied by hunter–gatherers. Archaeology can provide answers to these questions, as recent investigations have unearthed.

1. Archaeological Examples

Archaeological examples of the historical evolution of these more varied features of hunter–gatherer society have been studied recently in a number of time periods and regions of the world. We now discuss five of these case studies in turn.

1.1 European Upper Palaeolithic

The Upper Palaeolithic in Europe took place during the final stages of the last ice age, from about 35,000 to 10,000 years ago. Modern people, *Homo sapiens*, had colonized the continent by this time. While the cultures of this lengthy period have much in common, clear trends towards more complex hunter–gatherer patterns are discernible in the last half of this period. For

example, after the coldest, driest phase which occurred around 20,000–18,000 years ago (the glacial maximum), climates were milder, especially in well-located southern latitudes such as southwestern France. At this time, known as the Magdalenian period, spectacular painted rock art, including images of extinct herd animals such as mammoth and bison, appear in limestone caves across the region. Occupation sites at this time are more numerous and much larger than previously. The suggestion is of increased numbers and densities of people living a more sedentary lifestyle. Hunting strategies were to target large numbers of herd animals that migrated seasonally to the highlands and back again, following the bloom of pasture grasses. The cave art which occurs in varied, regionalized styles, suggests the development of more territorial social units, each with their own set of artistic repertoires. Cave art, site numbers, and the economy, therefore, among other things, reflect a more populous, sedentary hunter-gatherer society. Burials, which are more common during this period, indicate that certain individuals were afforded more attention in death, suggesting the appearance of aspects of social differentiation (Conkey 1985, Gamble 1986, Jochim 1983, Mellars 1985).

At this time, out on the frigid and largely treeless southern Russian Plain, hunters resided in permanent hamlets of well-constructed mammoth bone dwellings. These bases were located in the lengthy riverine corridors, such as the Don and Dnepr, through which myriads of large herd animals migrated annually. Meat was stored in frozen subterranean pits. Hamlets were part of an intricate network of long-distance exchange—in exotic stone, shell, and ‘Venus’ figurines—which united peoples resident in a harsh, vast arctic desert. Social inequality is suggested by unequal numbers of storage pits around individual dwellings, and perhaps ‘status’ burials, in a prior phase (Soffer 1985). More intensive economic strategies are evident during this period, as was the case also in southwestern France.

1.2 Eurasian Mesolithic

Traditionally, the Mesolithic of Eurasia was considered a depauperate period when contrasted with the more spectacular Upper Palaeolithic. Recent re-evaluations, however, point to complex trends, culminating in patterns of land and resource management, and ultimately, in some areas, to agriculture. This was a time of major climatic change, as warmer, moister conditions, similar to those of today, followed upon the colder glacial Pleistocene period around 10,000 years ago. The Mesolithic, therefore, is a time of adaptation by hunter-gatherers to changing environments both natural and cultural. The temperate zones of central and northern Eurasia were transformed from arctic deserts to mosaics of deciduous forests, lakes, and marshes. Following upon melting glaciers,

the seas rose, creating, for example, the vast Baltic Sea and its resource-rich coastline in northwestern Europe. Populous, semi-sedentary hunter-gatherer societies appeared on these coastlines, of for example present-day Scandinavia, as sea levels stabilized about 6,000–5,000 years ago. They established large shell middens and practiced a wide range of fishing, hunting, and gathering (including storage of nuts). Comparable societies appeared in estuaries and river valleys throughout Europe (Price 1987, Zvelebil 1986).

At the same time in Japan, beginning around 12,500 years ago, the Jomon hunter-gatherers on the rich environments of the main island of Honshu lived in semi-sedentary subterranean pit houses along the coast and inland. They fished intensively and stored quantities of nuts, acorns, and seeds in underground storage pits. It has been suggested that nut groves may have been an economically managed resource as perhaps also were seeds, including grass seeds. These populous, territorially based and ritually complex hunter-gatherers intensified still further their fishing/storage economies through time. In the final stages of their history they resisted for some time the introduction of agriculture (from the Asian mainland) (Akazawa 1986, Crawford 1992), as did also the Scandinavian hunter-gatherers.

It was in drier regions of the Middle East (southwest Asia), however, that Mesolithic hunter-gatherers developed complex societies and economies that also gave rise to agriculture. The Natufian hunter-gatherers of present-day Israel and Palestine between 12,500 and 10,500 years ago practiced intensive collection and storage of wild grass seeds and nuts. They resided in hamlets of small, round huts, and practiced the long-distance exchange of obsidian and shells. Shrines are found within their hamlets where they also buried their dead. Around 10,000 years ago, the domestication of wild grasses—including wheat and barley—took place, signaling the shift to agriculture. However, all other cultural features were essentially similar to those existing before. Agriculture, therefore, was but an extension of the complex hunting-gathering out of which it had sprung. And later agricultural villages, including the walled town of Jericho, were centered upon previously existing Natufian hamlets (Bar-Yosef 1987, Henry 1985). The dynamics of complex hunting-gathering, therefore, gave rise to agriculture as we know it today.

1.3 Archaic North America

The Archaic period of North America, like the Mesolithic, takes place as climate shifted from glacial to modern after about 8,000 years ago. Long-term trends, taking thousands of years, can be recognized in the rich riverine regions of the American mid-continent, where the densest hunter-gatherer populations resided. In riverine corridors such as those of the Mississippi and Ohio, population grew slowly over

thousands of years, while in the Illinois valley, semi-sedentism in the form of cold-weather bases appeared in time. Intensification of the hunter-gatherer-fisher economy accompanied these developments. It took some 4,000 years for long-term trends towards more complex social formations to develop in these regions during the Middle Woodland period. By about 2,100 years ago complex ceremonialism in the form of extensive mounded burial precincts and exchange and trade in a wide range of exotic material (copper, obsidian, pottery) united most of the regions of mid and eastern North America. The richness of the natural environment cannot be the only factor controlling the development of these complex hunter-gatherer societies, as prime environmental conditions had been established as far back as 6,000–5,000 years ago, long before the above cultural features appeared and flowered. Some of the most complex societies—such as those of the Ohio Hopewell—were located outside the riverine corridor itself, on the more arid plains and were relatively mobile rather than semi-sedentary. Their most common unifying feature lay in the ceremonial realm (Bender 1985, Brown 1985). Although agriculture was known to these peoples, it was not adopted by them until more recent times—held at bay, as we have already observed, for some Mesolithic hunter-gatherer societies.

1.4 Preagricultural Mesoamerica

Less archaeological evidence is available for observing trends towards hunter-gatherer cultural complexity in preagricultural Mesoamerica, the region which lies between the two continents of North and South America. Hunter-gatherers of this region also developed agriculture independently. Very long-term trends, however, spanning some 4,000 years, suggest a gradual shift towards higher levels of 'domesticated' produce since about 7,000 years ago. Even by about 4,000 years ago less than one third of the diet was agricultural. By this time some of the main domesticated plants included the staples of later times, maize and beans. Sedentism, in the form of villages, occurred *after* this time, when agriculture had been more firmly established. Aspects of complex exchange and trade had accompanied these trends, and agricultural villages (and later cities) were constructed around ritual precincts, which strongly suggests that one of the key driving forces behind communal aggregations (that is, sedentism) was inter-group ritual itself (Flannery 1968, 1976, MacNeish 1970).

1.5 Holocene Australia

Recent evidence from the continent of Australia suggests that comparable trends towards more com-

plex hunter-gatherer societies, similar to those of the Mesolithic and Archaic periods discussed above, are evident in Australia, especially from about 4,000–3,000 years ago. Aspects of these trends are also evident long before, for example after the glacial maximum or from about 16,000 years ago. In more recent times, however, long-term trends towards greater numbers of archaeological sites and use of sites occurs in many regions of the continent, including most major environmental zones—arid, tropical, and temperate. Demographic changes, including increasing populations and greater levels of sedentism, or reduced mobility, are suggested here, and are accompanied by other sociocultural changes (Lourandos 1997). In the tropical Cape York peninsula located in the northeast, shifts to more regionalized patterns of painted rock art occur during the last 2,000 years or so, suggesting more territorially based societies (David and Lourandos 1998, 1999). In arid Australia during the same period, a shift occurs towards more intensive seed-based (including grass seed) economies. The ethnographic descriptions of seed use (including grass seeds) by Australian Aborigines indicated they were managed via a wide range of practices including firing, irrigation, and the planting and broadcasting of seeds (Smith 1989, Veth 1989). As in all world regions studied, climatic change—in this case towards drier climates—accompanies many of these trends, but cannot on its own explain them.

2. Discussion

Trends can be observed, therefore, in many areas of the world which incorporate features of broader societal variation among nonagricultural peoples. These have included: increased population sizes and densities; sedentism storage and surplus; intensified and specialized economic practices; the suggestion of land and resource management; and broader social, including more hierarchical, formations. To what extent these prehistoric societies reflect ranking, including chiefdoms and the like, is more difficult to discern clearly. However, aspects of this, and the social context in which these systems are embedded, has been established in a number of case studies. The trends themselves took place over huge spans of time covering, in some cases, thousands of years.

Explanations of the trends in the development of hunter-gatherer society are many but tend to be somewhat polarized. As seen in the above case studies, climate, among other things, strongly influenced many of the general patterns and trends. But the dynamics within hunter-gatherer society itself also need to be considered. For example, incentives to intensify production (economy) may be embedded within the competitive social arena itself. In all, however, such

large-scale processes must have involved a complex set of factors—among them, ecological, demographic, and sociocultural.

Methodological problems also limit our investigations. Difficulties lie in gaining access to finer-grained, middle-range archaeological details. Issues of scale, for example, tend to blur distinctions between macro and micro issues in both time and space. Enormous spans of time and vast regions have been investigated in the general models offered. Access to archaeological data is a slow, meticulous process and answers to many of the questions raised here will take time to resolve.

In all, future directions lie with a much wider range of questions and approaches than were posed in more traditional times. The history of our hunter-gatherer past, quite clearly, has begun to be written anew.

See also: Chiefdoms, Archaeology of; Evolutionary Approaches in Archaeology; Fishing Societies; Hunting and Gathering Societies in Anthropology; Kinship in Anthropology

Bibliography

- Akazawa T 1986 Hunter-gatherer adaptations and the transition to food production in Japan. In: Zvelebil M (ed.) *Hunters in Transition: Mesolithic Societies of Temperate Eurasia and Their Transition to Farming*. Cambridge University Press, Cambridge, UK, pp. 151–66
- Ames K M 1985 Hierarchies, stress, and logistical strategies among hunter-gatherers in northwestern North America. In: Price T D, Brown J A (eds.) *Prehistoric Hunter-gatherers: The Emergence of Cultural Complexity*. Academic Press, Orlando, FL, pp. 155–80
- Bar-Yosef O 1987 Late Pleistocene adaptations in the Levant. In: Soffer O (ed.) *The Pleistocene Old World: Regional Perspectives*. Plenum, New York, pp. 219–36
- Bender B 1985 Prehistoric developments in the American midcontinent and in Brittany, northwest France. In: Price T D, Brown J A (eds.) *Prehistoric Hunter-gatherers: The Emergence of Cultural Complexity*. Academic Press, Orlando, FL, pp. 21–57
- Bettinger R L 1991 *Hunter-gatherers: Archaeological and Evolutionary Theory*. Plenum Press, New York
- Brown J A 1985 Long-term trends to sedentism and the emergence of complexity in the American Midwest. In: Price T D, Brown J A (eds.) *Prehistoric Hunter-gatherers: The Emergence of Cultural Complexity*. Academic Press, Orlando, FL, pp. 201–31
- Byers D S (ed.) 1970 *The Prehistory of the Tehuacan Valley*. University of Texas Press, Austin, TX
- Conkey M W 1985 Ritual communication, social elaboration, and the variable trajectories of Paleolithic material culture. In: Price T D, Brown J A (eds.) *Prehistoric Hunter-gatherers: The Emergence of Cultural Complexity*. Academic Press, Orlando, FL, pp. 299–323
- Crawford G 1992 Prehistoric plant domestication in East Asia. In: Cowan W, Watson P J (eds.) *The Origins of Agriculture*. Smithsonian Institution Press, Washington, DC, pp. 7–38
- David B, Lourandos H 1998 Rock art and socio-demography in northeastern Australian prehistory. *World Archaeology* 30(2): 193–219
- David B, Lourandos H 1999 Landscape as mind: Land use, cultural space and change in north Queensland prehistory. *Quaternary International* 59: 107–23
- Flannery K V 1968 Archaeological systems theory and early Mesoamerica. In: Meggers B J (ed.) *Anthropological Archaeology in the Americas*. Anthropological Society of Washington, Washington, DC, pp. 67–87
- Flannery K V (ed.) 1976 *The Early Mesoamerican Village*. Academic Press, New York
- Gamble C 1986 *The Palaeolithic Settlement of Europe*. Cambridge University Press, Cambridge, UK
- Henry D O 1985 Preagricultural sedentism: The Natufian example. In: Price T D, Brown J A (eds.) *Prehistoric Hunter-gatherers: The Emergence of Cultural Complexity*. Academic Press, Orlando, FL, pp. 365–84
- Ingold T, Riches D, Woodburn J (eds.) 1988 *Hunters and Gatherers 1: History, Evolution and Social Change*. Berg, Oxford, UK
- Jochim M 1983 Palaeolithic cave art in ecological perspective. In: Bailey G N (ed.) *Hunter-gatherer Economy in Prehistory*. Cambridge University Press, Cambridge, pp. 212–9
- Lee R B 1988 Reflections on primitive communism. In: Ingold T, Riches D, Woodburn J (eds.) *Hunters and Gatherers: History, Evolution and Social Change*. Berg, Oxford, UK, pp. 252–68
- Lourandos H 1997 *Continent of Hunter-gatherers: New Perspectives in Australian Prehistory*. Cambridge University Press, Cambridge, UK
- Marquardt W H 1988 Politics and production among the Calusa of South Florida. In: Ingold T, Riches D, Woodburn J (eds.) *Hunters and Gatherers: History, Evolution and Social Change*. Berg, Oxford, UK, pp. 161–88
- Mellars P A 1985 The ecological basis of social complexity in the Upper Palaeolithic of Southwestern France. In: Price T D, Brown J A (eds.) *Prehistoric Hunter-gatherers: The Emergence of Cultural Complexity*. Academic Press, Orlando, FL, pp. 271–97
- Price T D 1987 The Mesolithic of Western Europe. *Journal of World Prehistory* 1(3): 225–305
- Schrire C (ed.) 1984 *Past and Present in Hunter-gatherer Studies*. Academic Press, Orlando, FL
- Smith M A 1989 Seed gathering in inland Australia: Current evidence from seed-grinders on the antiquity of the ethno-historical pattern of exploitation. In: Harris D R, Hillman G C (eds.) *Foraging and Farming: The Evolution of Plant Domestication*. Unwin Hyman, London, pp. 305–17
- Soffer O 1985 Patterns of intensification as seen from the Upper Paleolithic of the Central Russian Plain. In: Price T D, Brown J A (eds.) *Prehistoric Hunter-gatherers: The Emergence of Cultural Complexity*. Academic Press, Orlando, FL, pp. 235–70
- Suttles W 1968 Coping with abundance: Subsistence on the Northwest Coast. In: Lee R B, DeVore I (eds.) *Man the Hunter*. Aldine, Chicago, pp. 56–68
- Veth P 1989 Islands of the interior: A model for the colonization of Australia's arid zone. *Archaeology in Oceania* 24: 81–92
- Williams N W, Hunn E S (eds.) 1982 *Resource Managers: North American and Australian Hunter-gatherers*. Westview Press, Colorado, CO

Zvelebil M 1986 *Hunters in Transition: Mesolithic Societies of Temperate Eurasia and Their Transition to Farming*. Cambridge University Press, Cambridge, UK

H. Lourandos

Hunting and Gathering Societies in Anthropology

'Hunting and gathering' is a conventional term for subsisting on the 'natural produce' of the earth by hunting and trapping wild animals, gathering wild roots, fruits, and berries, fishing, and collecting honey of wild bees, as opposed to producing food by agriculture and herding. By hunting and gathering societies (HGS) ethnographers mean contemporary indigenous people who fully or partly engage in these activities, now or formerly, or who see themselves as hunters. (For brevity, the terms 'hunter-gatherers,' 'hunters,' and 'foragers' are sometimes used synonymously.) The term is also used in general reference to prehistoric populations, living before the agricultural revolution (around 10,000 years BC). In common usage, it also stands for an alien 'Other,' diversely constructed by modernity in order to define itself. In this article, the primary focus is placed on the first meaning, while note is taken of its brush with the other two. The term is first examined historically, with an outline of its origins and changes of usage over time. Next, several clusters of features that are partially shared by these societies are presented in model terms, after which the diverse populations concerned are described by region, with the focus on their historical circumstances.

1. Contested Meanings

An understanding of the expanding meanings of HGS necessitates a return to the origin of the field and perusal of its historical development. In the seventeenth century, writers invoked 'primitive peoples,' who hunt and gather for their subsistence, in discussing the 'natural state' of humankind. By the eighteenth century these were typed 'hunter-gatherers,' in distinction from 'herders,' within a broad 'savage' class, in writings on the origins of property relations. In the nineteenth century evolutionist work continued to draw on these specific peoples, often without referring to the class as a whole, in studies of the origins of a variety of other social institutions, from kinship through politics to religion.

The notion consolidated as a central term in early twentieth-century cultural-ecology theory, most influentially in Julian Steward's work (1936). Concerned with the relation between society and environment, Steward argued that for any given mode of subsistence, successful adaptation determines a certain level of

socio-cultural integration. In the case of hunting-gathering, he argued, this is a 'band,' which is universally patrilineal, exogamous, and territorial. HGS was associated with bands, and hunting peoples were comparatively studied as band-societies (see, for example, Damas 1969), until the 1966 symposium 'Man the Hunter' (Lee and DeVore 1968). This interdisciplinary symposium is commonly considered the beginning of modern hunter-gatherer studies.

The symposium was inspired by the assertion that human beings had subsisted by hunting for over 99 percent of the time of human life on earth. Linking cultural-ecological and (neo)behavioral-evolutionary perspectives, its major objective was to learn about human origins and nature by comparing 'living hunters.' Uniform patterns found among them, it was posited, can be regarded as hunter-gatherers' successful adaptations, present and past.

Serious reservations were raised about its agenda. One objection, for example, was that no contemporary group could be seen as 'living relics,' or as direct descendants, of the Pleistocene. Even if a group had been found to subsist purely on hunting, its people would have had a history as long as any other, and they would not have lived any longer in a world populated by hunters as in the Pleistocene. Another objection was that over the span of 2,000,000 years humans had changed considerably, biologically, and psychologically, enough to render homogenizing them inappropriate.

The ethnographic comparison, at the same time, shook accepted wisdom to the core. Contrary to Steward's persistent thesis, research revealed that most bands are actually socially fluid and nomadic. In many cases, gathering makes no less important a contribution to subsistence than hunting. In contrast to contemporary stereotypes and bourgeois intuition, the comparison showed that these peoples meet their needs with surprisingly little labor. (Marshall Sahlins provocatively dubbed them 'the original affluent society'.)

Outweighing the reservations, this successful outcome entangled HGS with the 'Man the Hunter' project. HGS came to mean a type of people, rather than diverse peoples engaged in a type of action, encompassing populations over a vast evolutionary span of time. The term now embodied the hypothesis that, at their core, these populations have similar socio-cultural forms, determined by adaptation to the environment. From a vague term of reference to 'very simple' peoples, and then to 'band societies,' it connoted a firm, homogeneous, and exclusive social category, a category of humanity defined technoeconomically.

Subsequent ethnography research, which burgeoned in the wake of 'Man the Hunter,' grew diverse in focus and approach, as seen for example in the series of International Conferences on Hunting and Gathering Societies (1978 to the present, e.g., Ingold et al.

1988, Vols. 1–2, Burch and Ellana 1994). Issues arising from late twentieth-century indigenous circumstances and cultures were addressed, including political struggles and self-administration in modern states, and indigenous peoples' own perceptions of their environment, economy, and self. Historical, interpretative, regional, and applied perspectives commanded attention; evolutionary-ecological studies of foraging behavior increasingly took a separate course. Indigenous people started contributing to academic research, as anthropologists and interested parties. This growing ethnography showed a diversity of social forms among hunting-gathering peoples, from egalitarian to hierarchical, and even slave-owning societies. The 1970s meaning of HGS, as a firm, homogeneous, and exclusive social category, was increasingly challenged. In an attempt to save it, the category was diversely divided into subtypes (e.g., see Woodburn 1982); or it was recharacterized, for example, by a collective appropriation of nature (see Ingold 1987) or by a distinct cultural perception of the environment (see Bird-David 1990).

The 1980s and early 1990s saw many controversies (see review by Myers 1988), sparked partially by inflexible readings of the term HGS. Scholars disputed the category as pertaining to a class of societies that are both singularized by hunting-gathering and distinguished from agricultural and pastoral peoples, because the ethnography revealed continuities between hunter-gatherers and other societies, as well as great diversity among the hunter-gatherers themselves. Other scholars questioned the categorization of societies by mode of subsistence, as research showed heterogeneous and changing economies. Hunting-gathering was found to be articulated with other pursuits, going back in some cases centuries and even millennia, and, at any rate, undergoing rapid changes since the mid-twentieth century. Students debated whether any living group sufficiently complies with the subsistence criterion and whether this criterion generates a meaningful category, one whose members share more than the selecting mode of subsistence. Even the field at large was critiqued as spurious, by researchers seeing it in terms of the 1960s objective to learn about prehistory and evolution. Their criticism drew on ethno-historical and archaeological work suggesting arguably that these societies have not escaped world and regional forces. HGS is sometimes used nowadays simply to denote the assemblage of indigenous peoples discussed by practitioners within the field of hunter-gatherer studies (see Kelly 1995).

HGS has acquired multiple layers of meanings. It is a circumlocution for a historically-situated field of research that provides a stimulating comparative context and a scholarly communal framework within which to explore a broad array of distinctive issues. Ethnographers construct 'partial connections' (in Strathern's (1991) sense) between these societies by studying their respective groups while keeping the

others 'in mind.' They call to attention partially shared features, which determine common issues in the field, and contribute distinctive perspectives to anthropology generally. Hunting and gathering activities are a focus of attention, rather than a criterion of demarcation, for scholars as for the indigenous peoples themselves.

2. Issues

This section presents in model terms some of the partially shared features that came to light in the course of ethnographers' comparative discussion of their field experiences and material.

2.1 Hunting and Gathering

In many of these societies, people combine these activities with other economic pursuits (e.g., trade in minor forest produce, horticulture, work for neighbors, wage-labor, social benefits, etc.). However, hunting and gathering themselves carry special significance; they are a total phenomenon encompassing economic, social, and symbolic aspects. Hunting-gathering, moreover, often influences the way in which people carry out such other activities. They often return to hunt and gather at least periodically, when not prevented from doing so. They regard what they collect as 'real food,' and while they do pursue other activities, in some cases hunting-gathering still provides a major part of their diet.

2.2 Band Society

In many of these societies people live in small groups (bands) consisting of several families; the groups are flexible. Making visits, and even shifting residence, from place to place, are common. The total population numbers a few thousand, and may be as low as several hundred. Men and women often know each other personally: they all are (or at least are classified as) family and relatives. Their sense of themselves as family often survives growing integration into nation-states.

2.3 Sharing

The land and its resources in many of these societies are often collectively appropriated (see Ingold 1987). Game is publicly shared; other resources, whether procured from the environment or purchased, are shared in response to requests directed by those who lack something to those whom they perceive have it ('demand-sharing': Peterson 1993). Sharing and the expectation of it is part of broader nonaggrandizing

economic reasoning and conduct. While hard to express in modern urbanite language, sharing extends not only to distributive but also to social experience. People in the group share 'each other' (Ingold 1987): company, dwelling spaces and work, stories, memories, and personal histories. They experience their selves as coalescent, rather than as a group of separate individuals. Relationships between people, rather than being regenerated by coercing external structures, spring from the vivid experience of sharing. Sharing, which is normatively promoted, fosters an egalitarian society in many cases (see Woodburn 1982). While closely connected, people usually enjoy considerable personal autonomy.

2.4 *Giving Environment*

Nature is locally experienced neither as a mechanical system nor as objects for human use and ownership, but as an assembly of sentient beings related to and shared with (Bird-David 1990). These beings (animals, forest, winds, hills, etc.) frequently are approached as relatives, often as grandparents. People care for them, and this helps sustain their collective identity. People engage with these beings through Shamanistic practices and by pursuing hunting-gathering. To hunt and to gather are perceived not just as subsistence pursuits but also as means of communication with the 'natural' relatives. Even when forced to reside away, and to adopt modern life-styles, people commonly practice their rituals, and return to their places to hunt-gather and 'keep in touch' with these relatives. This animistic perception is integrated with deep ecological understanding of their neighborhood. It often propagates responsiveness to and sometimes responsibility for the environment, along with its utilization.

3. *Populations*

The indigenous populations considered as hunter-gatherers are described in this section, by geographical cluster, in keeping with the *Cambridge Encyclopaedia of Hunters and Gatherers* (Lee and Daly 1999). The figures provided should be regarded as suggestive, as politics of identity, especially in First World situations, and dispersal and inaccessibility, especially in Third World countries, severely limit enumeration.

3.1 *Australia*

About 266,000 people identify themselves as of Aboriginal or Torres Strait Islander descent, constituting 1.6 percent of the national population (1991 census). This is the case on a continent inhabited solely by hunter-gatherers until the arrival of Europeans in 1788. Compared with an estimated 200 languages and 600 dialects in that year, they now divide into groups speaking about 50 languages. Among the best known

are Aranda, the nineteenth-century study of whom was pivotal in anthropology, and Walipi- and Pintupi-speakers, desert-dwellers and the last to make contact with Europeans. A disproportionate number of Aboriginal people are now largely dependent on social security payments, and do poorly by any standard social indicator. At the same time, a revival of tradition is on the increase, in part due to the growing national and international markets for Aboriginal art (especially the desert 'dot-painting'). Politically, significant advances have been recently made. First was the establishment in 1990 of the Aboriginal and Torres Strait Islander Commission (an indigenous body formed to make policy and distribute federal funding for Aboriginal people). This was followed by the High Court recognition in 1992 of native land title (the 'Mabo case').

3.2 *North America*

Commonly called American Indian Nations in the US, and First Nations in Canada, there are about 500 indigenous groups in this subcontinent. The stratified Tlingit and Haida of the northwest coast, and the egalitarian Arctic Eskimo-Aleut speakers, are renowned for their role in early anthropology. The sub-Arctic Innu, Cree, and Dene are known for their opposition to contemporary environment-threatening development schemes. Prior to European colonization, two-thirds of the subcontinent had been occupied by hunting and gathering people. Devastated by the colonization, some groups perished, and some survived in fragmented forms in reservations. To date, a disproportionate number experience severe social problems. However, recent decades have seen expansions of indigenous organizations, rapid growth of arts, and revival of traditional spirituality (attracting New Age followers). Hunting and gathering economies are being renewed, especially on the north and the northwest coast, as a result of some judicial breakthroughs that brought a degree of recognition of aboriginal and treaty rights. A major landmark was the 1999 establishment of Nunavut, an Inuit majority territory within Canada's federal structure.

3.3 *South America*

Foraging people in the subcontinent's southern cone were almost extinct by the 1920s, and only a handful of each group remains today, assimilated into other communities. A wide spectrum of forest communities, cultivators and hunters by trade, live scattered in the Amazonian hinterlands. In the 1960s they were viewed predominantly as devolved agriculturists, and only later did their egalitarian and nomadic life-styles earn them the attention of the discipline. The Paraguayan Ache is one of the best known groups. Most of these indigenous people have been converted and restricted to reservations; since the 1960s large-scale economic

development has threatened their environment and life-ways even more. As a result of joint campaigns by indigenous activists and international conservation and human rights movements, a slow but significant movement towards recognition of indigenous territories has been taking place since the 1980s.

3.4 North Eurasia

Latecomers into the hunter-gatherer discourse, there are 26 Arctic and sub-Arctic groups in the Russian north. They are divided into at least six language families, and number approximately 184,500, constituting 0.06 percent of the entire Russian Federation population (in 1990). In the seventeenth and eighteenth centuries the majority moved to reindeer pastoralism as their main subsistence occupation, supplementing it by hunting and fishing. Autonomous districts were established for some of them in the 1930s, although political and economic decisions were still made in Moscow. A substantial majority is now modernized, urbanized, and Russified, but activists (mainly writers) campaign for cultural renewal. Under the mounting economic difficulties in Russia, many are attempting to restore some traditional practices.

3.5 Africa

Bushmen (San) and Pygmy groups caught the 1960s' imagination, not least as their habitat was regarded the 'cradle of humankind.' They number 100,000 and 285,000 respectively; each comprises diverse linguistic groups spread across more than half a dozen countries. A further 55,000 people, largely in East Africa, bring the total number of African foragers, part-foragers, and former-foragers to 440,000. Only a tiny fraction of them subsist intensively on foraging. Most keep close ties with cultivators and pastoralists; the Pygmy groups are known for what is described as symbiotic relationships with their neighbors. With worsening circumstances, some groups sought help from development agencies and church groups and some joined environmental activists in protest against ill-advised large development schemes. The Ju/'hoansi (alias !Kung), the best known hunter-gatherer group owing to Lee's extensive work (e.g., 1979), have initiated self-help schemes intended to develop awareness of nation-politics, and leadership and management styles adapted to it.

3.6 South Asia

Less than two dozen indigenous groups live in the South Asian subcontinent. These total less than 20,000 out of over 1,000,000,000 people (wider estimates, including former foragers, put their number at 1,300,000). They include the exceptional Jarwas and Sentinelese of the Andaman Islands, who by hostile behavior have warded off outside contact up to the

present, and South Indian forest-groups who, going back 2,000 years, have played an important part in the economies of the subcontinent's ancient civilizations. The latter groups have provided the main society with exotic spices, medicinal herbs, honey, and meat, occasionally guiding Hindu pilgrims and officiating in their forest temples. Some groups are being pressed to 'develop' into settled agriculturalists by a national modernist policy that sees them as 'backward people' whose existence shames the country; some combine hunting-gathering with trade in minor forest produce and wage-work. Sri Lanka's Wanniyala-aetto (formerly Veddahs) raised strong opposition to their forced relocation, going as far as the United Nations.

3.7 Southeast Asia

Foraging and swidden farming are closely linked in this region, with historical shifts back and forth between them, frequent combinations, and close trade-links between foragers and farmers. In the Philippines there are about 25 distinct linguistic groups, totaling 15,000; in Malaysia, eight linguistic groups total 2,500 people; in Indonesia the total is 15,000, in Borneo it is 12,000, in China it is 4,500, and in Sumatra and Thailand it is 800 and a few hundred, respectively. Governments in this region encourage commercial exploitation of these people's forest habitat, and they rapidly become poor farmers and laborers.

4. Conclusions

Stimulated by the hope to learn about human evolution and prehistory, the study of 'living hunters' has developed into a fruitful area of study in and for itself. These living peoples still afford a study of behavioral-ecological parameters of hunting-gathering, which are important for them, and which they pursue when they can. Over and above this, however, they open a window to socioenvironmental ethics and praxis, conceptions of society and nature, human relationships and economy, which are markedly different from familiar models. In broadening scholars' imagination, that cannot but be biased by urban life-experiences and intuitions, hunting and gathering peoples can help scholars think about other forms of life: human life in the past as well as possible futures. Indigenous peoples for their part increasingly explore new modernities that integrate their traditional practices with life in twenty-first-century nation-states.

See also: Economic Anthropology; Hunter-Gatherer Societies, Archaeology of; Social Evolution: Overview

Bibliography

Barnard A 1983 Contemporary hunter-gatherers: Current theoretical issues in ecology and social organization. *Annual Review of Anthropology* 12: 193-214

- Bird-David N 1990 The giving environment: Another perspective on the economic system of gatherer-hunters. *Current Anthropology* 31: 189–96
- Burch Jr E S, Ellana L J (eds.) 1994 *Key Issues in Hunter-gatherer Research*. Berg, Oxford, UK
- Damas D (ed.) 1969 *Contributions to Anthropology: Band Societies*. National Museum of Canada, Ottawa, ON
- Ingold T 1987 *The Appropriation of Nature: Essays on Human Ecology and Social Relations*, 1st edn. University of Iowa Press, Iowa City, IA
- Ingold T, Riches D, Woodburn J (eds.) 1988 *Hunters and Gatherers*. Berg, Oxford, UK, 2 vols
- Kelly R L 1995 *The Foraging Spectrum: Diversity in Hunter-gatherer Lifeways*. Smithsonian Institution Press, Washington, DC
- Lee R B 1979 *The Kung San: Men, Women and Work in a Foraging Society*. Cambridge University Press, Cambridge, UK
- Lee R B, Daly R (eds.) 1999 *Cambridge Encyclopaedia of Hunter-Gatherers*. Cambridge University Press, Cambridge, UK
- Lee R B, DeVore I (eds.) 1968 *Man the Hunter*. Aldine, Chicago
- Myers F 1988 Critical trends in the study of hunter-gatherers. *Annual Review of Anthropology* 17: 261–82
- Peterson N 1993 Demand-sharing: Reciprocity and the pressure for generosity among foragers. *American Anthropologist* 95: 860–74
- Steward J H 1936 The economic and social basis of primitive bands. In: Lowie R H (ed.) *Essays in Anthropology in Honor of A. L. Kroeber*. University of California Press, Berkeley, CA
- Strathern M 1991 *Partial Connections*. Rowman & Littlefield, Savage, MD
- Woodburn J 1982 Egalitarian societies. *Man* 17: 431–51

N. Bird-David

Hurst, James Willard (1910–97)

It is hardly controversial that J. Willard Hurst of the University of Wisconsin was the most influential scholar ever within the realm of US legal history. By emphasizing factors beyond legal doctrine and by altering the way people think about time, place, and change itself, Hurst enlivened and transformed an entire academic field. As leading legal historian Lawrence Friedman said: ‘Willard Hurst, as a scholar, did what few scholars have ever succeeded in doing: he created a whole new field of study, and virtually out of nothing.’

1. Biographical Background

James Willard Hurst was born on October 6, 1910 in Rockford, Illinois, where his father ran a movie theater and had other business interests. When Hurst attended Williams College from 1928 to 1932, he majored in economics, became editor in chief of the student newspaper, and president of the International Affairs

Club. He also wrote a prize-winning student essay on international implications of the English Gold Clause crisis of 1931, published in the *Journal of Political Economy* the year he graduated. Hurst then attended Harvard Law School, and reported that he was disgruntled about the emphasis on abstractions and the failure to pay attention to the Depression and to the ideas of the legal realists. A note editor of the *Harvard Law Review*, Hurst graduated in 1934 and spent the 1935 to 1936 academic year as a research fellow for Felix Frankfurter, working primarily on a lecture series which Frankfurter published as *The Commerce Clause under Marshall, Taney, and Waite* (1937).

Hurst next served as law clerk for 80-year-old US Supreme Court Justice Louis D. Brandeis during the 1936 term, one of the court’s most contentious ever. Frankfurter reported that Brandeis said that Hurst ranked as his best clerk ever, ‘professionally speaking.’ Brandeis thought the states to be invaluable laboratories for progressive regulation and lawmaking, and he urged young men who worked for him to return to their home regions. Hurst became one of the very few to heed Brandeis’s advice. He began his lifetime association with the University of Wisconsin in 1937, when he joined the law faculty in Madison, only a short drive north of Hurst’s hometown.

By 1941, Hurst had compiled several books of teaching materials and had coauthored *Law in Society* with his dean, Lloyd Garrison. This pioneering case-book focused on the myriad ways that the plight of a widow—whose husband, a carnival worker, was crushed under a company tractor—intersected with legal institutions beyond the courts. After Pearl Harbor, Hurst and his wife, Frances Wilson, moved to Washington, DC. Hurst first worked for the general counsel of the Board of Economic Warfare and then joined the navy, where his assignments included work on a highly publicized treason case, leading eventually to his *The Law of Treason in the United States* (1971).

Even before returning to the Wisconsin law faculty after World War II, Hurst began to articulate in print the crucial link he perceived between legal history, done in a new way, and ‘opportunities to apply informed, humane reason to influence the course of events.’ The ‘Wisconsin idea’ of enlightened political leadership had lasting appeal for Hurst—and he repeatedly rejected job offers to teach at other prestigious institutions. In the course of rejecting an offer from Yale Law School in 1938, for example, Hurst wrote to his mentor Felix Frankfurter that Wisconsin afforded ‘just about an ideal “laboratory” situation.’ Despite Yale’s offer of higher rank and twice the salary, Hurst advised Frankfurter that he preferred Madison ‘particularly because of this extraordinarily interesting state government machinery here on the doorstep.’ Although Hurst was extremely reluctant to travel from his Wisconsin base throughout his life, he nonetheless was directly involved in launching the

international Law and Society movement, as well as in pioneering a new approach to legal history that has had worldwide impact. He taught full time at Wisconsin until 1980 and continued to publish after retiring from teaching. Hurst died in Madison on June 18, 1997.

2. Hurst's Scholarship

Hurst's scholarly career repeatedly underscored a core idea. In 1971, he summarized the enormous project he had undertaken: 'A realistic history of law in the United States will be a social history of law, taking law as man-made, and as the product of both deliberation and drift, but not of any immanent superior order of reality.' Doing legal history from the outside—from the perspective of sociological jurisprudence—led Hurst to turn his back on the 'great man' tradition and to reject the previously dominant focus on formal judicial opinions. He firmly opposed teleological accounts that tended to celebrate the gradual improvement of law over time. Instead, Hurst focused on economic matters—as shaped by the general culture—and on the ordinary processes of statutes, administrators, low-level cases, and the everyday work of lawyers.

In particular, Hurst concentrated on analyzing what he termed 'our prime inheritance,' which he defined as 'middle-class ways of thinking.' With exceptional tenacity, he set about tracking 'the expansion of middle-class energy.' He sought to describe and to analyze how most people in the USA responded to a central problem. That basic issue, according to Hurst, confronts 'whether men will be able to use the power of the politically organized community to secure the basic conditions of a decent life for the individual, without thereby destroying the very values they are trying to realize.'

Throughout his decades of scholarship and teaching, Hurst combined unrelenting and careful attention to massive low-level details with sweeping descriptions of general patterns. He sought to develop a new integration of the social-science ideas of his era with careful historical consideration of the largely unexamined mass of law as it functioned on the ground, far removed from the doctrines promulgated by selected appellate judges that dominated—and still dominate—US legal education.

Robert Gordon described the great reach of Hurst's innovative academic project as follows:

After expanding the relevant record beyond prior appellate cases and their facts to all prior legal acts and the social influences on them, and the pending proceedings to the overall task of designing workable legal-regulatory arrangements for the present age, Hurst wrote his books as background briefs for present-day lawmakers.

To Hurst, social patterns mattered greatly. But he perceived that generalized thoughtlessness and a

strong tendency to accept inertia and drift often served to counterbalance, or even to disable, other strong, competing US commitments.

Through his prodigious output of scholarship on markets, legislation, monetary policy, and administrative actions large and small, Hurst pointedly emphasized the frequent unthinking clash between core US values. He seemed to delight in clearly describing basic and sobering paradoxes. Thus, for example, Hurst observed: 'Individuals realize their humanity only in society, but they realize their individuality only in self-awareness, which consists in some sense of separateness.' Moreover, he noted that such tensions afford protective coloration for 'shrewd men' who 'warp individualistic symbols of the constitutional ideal to sanction unchecked organized power.'

As Hurst challenged traditional legal distinctions between public and private activity, he pushed himself and those who followed him to develop both extensive local data and expansive theories to capture and explain overarching patterns. In *The Growth of American Law* (1950), Hurst argued that 'law has been more the creature than the creator of events.' This famous book emphasized the impact of physical, technological, and social facts on nineteenth-century law and paid particular attention to the often inglorious role of the legal profession. Some have criticized Hurst for being too much a functionalist, and for inadequately attending to issues of power. Yet Hurst's 946-page *Law and Economic Growth: The Legal History of the Lumber Industry in Wisconsin, 1836–1915* (1964), for example, highlighted the failures of laissez-faire. It also extensively documented Wisconsin's broad and detailed state regulation of a single industry—during a period generally thought to have been at or near the apex of free-market ideology.

3. Hurst's Impact

William Novak asserted in 2000 that Hurst surely ranks as 'a broad-gauged socio-legal thinker of the first order, writing in the tradition of Tocqueville, Weber, Holmes, and Pound.' Hurst was original in his particular emphasis on place and time, particularly through his focus on his native Midwest. He also innovated through his unusual illustration of the difficulty in discerning any clear patterns of progress in US legal developments. Indeed, Hurst concentrated on exploring the intricate network of constraints—some perceived, some never understood, some formally legal, some not legal at all—that repeatedly confined and constrained the freedom of everyone.

In Hurst's view, Americans competed fervently, if often blindly, to manipulate whatever system they found in place in their particular context. He generally found it necessary to bemoan the power of deep habits that often produced irrational development and sub-

stantial depredations over time. Yet Hurst also provided crisp accounts of the constitutional impulse among Pike Creek squatters, for example, as well as broad studies of such issues as the fierce, largely irrational conflict over gold and silver currency in the late nineteenth century.

Hurst sometimes has been inaccurately caricatured as part of the celebratory framework of some forms of ‘consensus history.’ Yet the very broad-gauged quality of his work (an ongoing quest, in Novak’s words, ‘to try to twist the tail of the cosmos’) emphasized the mainstream at the cost of excluding, or at least understating, the importance of minorities and of dissenting viewpoints throughout US legal history. Hurst’s faith in fact-gathering led him to produce not merely a new way of looking at law, but also an extended historical critique of what he called ‘bastard pragmatism’—the propensity of people throughout US history to allow themselves to be dominated by the short term and the bottom line.

Through his entire career, Hurst emphasized the difficult, if not impossible, goal of increasing both public and private accountability. In *Law and the Conditions of Freedom* (1956) he noted a clear historic preference among Americans for ‘the promise of the future over commitments to the past, but also for the ambitions of the present over the claims of the future.’ Somewhat ironically, Hurst nevertheless remained deeply committed to the progressive ‘Wisconsin idea.’ To Hurst facts were crucial. Only through facts could a doubting optimist, such as Hurst himself, imagine enlightened leaders actually enhancing the general pursuit of ‘lives of good conscience and fulfillment.’ If Hurst sometimes seemed unabashed in adopting what Daniel Ernst called ‘a presentist approach to the past,’ he remained first and foremost an innovative, exceptionally influential historian.

The unusual number and breadth of symposia about Hurst’s work only begin to suggest how significantly his emphasis on the impact of markets, legislation, monetary policy, social habits, and administrative actions transcended earlier boundaries. Through his enormous output of scholarship, his legendary generosity in giving scholarly assistance to others, and his decades of noteworthy teaching, Willard Hurst changed the way US legal history is written and taught. By now, as Lawrence Friedman put it, ‘You’re either a Hurstian or a reviser of Hurst.’ Hurst not only redefined an entire academic field, but he did much to open the borders of legal history to other disciplines. Hurst’s career challenges his successors to consider carefully the intricate, crucial, often paradoxical intertwining of legal ideas and social realities within the many realms in which people actually live.

See also: Law: History of its Relation to the Social Sciences; Law, Sociology of; Legal Education; Legal Scholarship

Bibliography

- Diamond S 1979 Law and social order in the United States—Hurst, JW. *Michigan Law Review* 77: 784–94
- Erlanger H (ed.) 1985 Review symposium: The work of J. Willard Hurst. *American Bar Foundation Research Journal*: 113–44
- Gordon R (ed.) 1975 *Law & Society Review* 10: 5–333
- Hartog H 1994 Snakes in Ireland: A conversation with Willard Hurst. *Law and History Review* 12: 370–90
- Scheiber H 1970 At the borderland of law and economic history: The contributions of Willard Hurst. *American Historical Review* 75: 744–56
- Soifer A 1992 In retrospect: Willard Hurst, consensus history, and the growth of American law. *Review of American History* 20: 124–44
- Tomlins C (ed.) 2000 Engaging Willard Hurst: A symposium. *Law and History Review* 18: 7–221
- Tushnet M 1972 Commentary: Lumber and the legal process. *Wisconsin Law Review* 114–32
- 1980 *Wisconsin Law Review*: 1090–442
- 1997 *Wisconsin Law Review*: 1122–210, Hurst bibliography: 1205–10

A. Soifer

Husserl, Edmund (1859–1938)

Edmund Husserl was one of the most influential philosophers of the twentieth century. Inventor of the ‘phenomenological method’ for the analysis of physical and mental phenomena, initiator of the ‘phenomenological movement’ of contemporary human sciences, he became the great critic of naive naturalism in philosophy and psychology (‘psychologism’), and of what he regarded as the loss of orientation, certainty, and thereby of credibility, of modern science. His appeal for ‘going back to the things themselves,’ and for a foundation of all science and philosophy in a clearly understood ‘life-world’ (*Lebenswelt*), inspired the work of a great number of philosophers (e.g., Gadamer, Heidegger, Ingarden, Levinas, Merleau-Ponty, Ricoeur, Sartre) and social scientists (e.g., Buytendijk, Lewin, Schütz).

1. Life and Academic Career

Edmund Gustav Albrecht Husserl was born in Prossnitz, Moravia, on April 8, 1859. He studied mathematics, physics, astronomy and philosophy at the universities of Leipzig (1876–1878), Berlin (1878–1881), and Vienna (1881–1882), where he obtained his Dr.Phil degree in mathematics with Leo Königsberger in 1883. After a short period as assistant

to the mathematician Karl Weierstrass in Berlin, and a year of military service, Husserl continued his philosophical studies with Franz Brentano in Vienna from 1884 to 1886, and with Carl Stumpf in Halle between 1886 and 1887, where he qualified as lecturer in 1887 with a postdoctoral thesis entitled 'Ueber den Begriff der Zahl. Psychologische Analysen' ('On the Concept of Number. Psychological Analyses').

Husserl became a professor in 1894, and moved to Göttingen in 1901. In 1916, he accepted a call to Freiburg to succeed Heinrich Rickert in the chair of philosophy there. Husserl retired in 1929. After the seizure of power in Germany by the National Socialists in 1933, his status of emeritus professor was suspended because of his Jewish descent, and, after a short reinstatement, definitively revoked in 1936 after the Nuremberg race laws. Husserl died in Freiburg on April 27, 1938.

2. Writings, Subjects, Approaches, and Contexts

Husserl's significance for philosophy, which is certainly greater than that for the social and behavioral sciences, rests on his establishment of a new way of dealing with philosophical problems, which he called 'phenomenology,' turning this expression into a 'technical' term by giving it a clearer delimitation and a weightier role than it had enjoyed with earlier philosophers such as Lambert or Hegel. The methods and content of Husserl's phenomenology, however, changed considerably during their development, which may be summarized in terms of three 'turns' of direction and aims: (a) the turn from early foundational work to anti-psychologism, (b) the shift from descriptive to 'constitutive' phenomenology, and (c) coping with the 'crisis' of modern science by a new focus on the life-world.

2.1 From Early Foundational Work to Anti-psychologism

In *Philosophie der Arithmetik* (*Philosophy of Arithmetic*, 1891, Husserl 1950, Vol. XII, pp. 1–283), Husserl made an attempt at a descriptive—psychological foundation of arithmetic, presumably having in mind an extension of Weierstrass's program of basing the vast number of mathematical concepts on a small number of fundamental ones, by providing an even deeper foundation for the latter by grounding them in the psychical acts underlying the mathematician's work. In the course of these investigations, Husserl introduced (independently of C. von Ehrenfels, who is normally credited with this discovery) 'gestalt qualities,' which he treated under the name *figurale Momente*.

Less successful (with hindsight) was Husserl's criticism of G. Frege's definition of number. It invoked an

objective but sharp counter-critique by Frege, which seems to have persuaded Husserl of the impossibility of taking the special science of psychology (which proceeds inductively, falling short of universally valid statements), as a foundation for mathematics or even for all special sciences ('psychologism'). Therefore, in *Logische Untersuchungen* (1901), aiming at a foundation of logic which at that time was usually set up on a psychological basis, Husserl advocated a consequent anti-psychologism which influenced contemporary philosophy profoundly. This concept was extended in his work 'Philosophie als strenge Wissenschaft' ('Philosophy as a Rigorous Science') (Husserl 1950, Vol. XXV, pp. 3–62) to a critique of historicism and (under the label of 'naturalism') scientism (1910–1911).

The positive foundation of logic was to be laid by a new foundational discipline to which Husserl, oddly enough, gave the name 'descriptive psychology,' despite his claim that it differed from empirical psychology as well as from Brentano's descriptive classification of psychical (mental) phenomena; it was to be 'descriptive' in the sense of presenting the immediately given mental facts without presuppositions (*voraussetzungslos*). According to this earliest form of phenomenology, that which is intended in a mental act (perception, judgment, idea, etc.) is transcendent to the act itself and given only in 'adumbrated' (*abgeschattet*) manner, through changing phenomena, and therefore inadequately. By contrast, everything immanent to the act is free of adumbration and can therefore be regarded as given in an absolutely adequate manner. Moreover, intuition of what is immanently given appears as free of structural complications like those attributed to judgment in contradistinction to sensation. The contents of the sphere of immediate experience (*Erlebnisse*) are beyond every doubt; with 'intentionality' at its center, it is the secure ground of a foundational discipline which is capable of justifying statements about essences (*Wesen*) and their relations, by a method which Husserl first called 'eidetics,' but later called 'intuition of essences' (*Wesensschau*).

Particular relevance to a philosophy of the sciences may be claimed for Husserl's remarks on 'pure logic.' Pure logic is an *a priori* theory of science dealing with two kinds of fundamental categories: categories of meaning (concept, proposition, truth, inference, etc.) on the one hand, and categories of objects (object, state of affairs, connection, number, etc.) on the other. Husserl takes the unity of the propositions of a science (or a discipline) to consist in the unity of its explanatory principles, and finds his next task in the uncovering of the conditions of the very possibility of theories. The route is to proceed from preliminary studies, e.g., a 'purely logical grammar' investigating the components and connections of propositions, via the statement of laws obtaining between such categories, to an *a priori* 'theory of possible theory forms.'

These investigations, continued only in *Formale und transzendente Logik* (*Formal and Transcendental Logic*, 1929) and in *Erfahrung und Urteil* (Husserl 1939), share a methodical aim: the ‘fixation of unambiguous, sharply distinct verbal meanings’ (Husserl 1901, Vol. I, p. 244; Husserl 1973 (English translation 1973), Vol. I, p. 238).

This search for ‘verbal meanings’ is also one of the foremost goals of contemporary methodical philosophy and philosophy of science. Husserl, however, did not approach the joint methodical aim by actually constructing a language of science, but recommended an ‘intuitive presentation (*Vergegenwärtigung*) of essence in adequate ideation,’ obviously conceived as independent of and antecedent to any linguistic construction. The ideation to be performed is not an idealization (as in geometry), but pure description, aiming at the elevation of that which is inexact and essentially vague to something typical, a ‘type.’ By setting out in this way, phenomenology increasingly moved away from the other methodical currents of modern philosophy and came closer to the philosophy of life (*Lebensphilosophie*), from which, however, Husserl differs sharply methodically by ascribing a foundational value to experience, after its precise rendering as ‘pure experience’ (*reines Erleben*) in which meaning, anterior to everything real, is believed to constitute itself (philosophical intuitionism).

2.2 The Step from Descriptive to ‘Constitutive’ Phenomenology

Object and method of phenomenology were extended and partly revised in *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie* Husserl (1913), and pure logic was also characterized in a new way. In Husserl 1901, the predicate ‘ontological’ had been restricted (diverging from tradition) to sciences such as geography, astronomy, and anatomy, which deal with concrete objects, and not with abstract objects (like the theoretical or nomological sciences). Now, the old term ‘ontology’ is taken up again in a new sense. Husserl takes being to be subdivided into highest kinds or ‘regions,’ so the sciences which deal with them yield ‘regional ontologies.’ Above these (material) ontologies we find the formal sciences, mathematics and logic, which are aimed at ‘form essences’ (*Formwesen*) and give rise to a *mathesis universalis*. Treating the ‘empty forms of region’ as such, they comprise the forms of all possible material ontologies, and are nothing else but a ‘formal ontology.’

In *Formale und transzendente Logik*, formal logic was subdivided into (a) a pure doctrine of forms of meaning, or logical grammar, (b) the logic of consequence, or logic of non-contradiction, and (c) the logic of truth. The purely formal ontology was characterized

more clearly as the logic of objects by severing it from the ‘apophantic’ logic of judgment, while the nomological sciences are deductive, formal–analytic sciences describing ‘definite manifolds.’ Fundamentally new is Husserl’s assessment of ontology as a naïve–dogmatic discipline that is turned into a science only by its founding on a ‘transcendental logic.’ The term ‘transcendental,’ as explained in 1936, refers to ‘the motif of inquiring back into the ultimate source of all the formations of knowledge, the motif of the knower’s reflecting upon himself and his knowing life’ (Husserl 1954, p. 100). Husserl’s reflective ‘pure consciousness’ of 1913 is now determined further by its function as the ‘transcendental ego,’ having everything as its object, while never becoming an object itself, and granting access only through ‘eidetic intuition.’

The absolute sphere, which is the foundation of ontology and may be distinguished from the ontological sphere as ‘semasiological’ (R. Zocher), is the object of a ‘critical’ transcendental philosophy, which Husserl can only conceive of as a fully developed phenomenology. It makes use of the grasping of essences as a tool for analyses which approach the problem of foundation not as a methodological task, but as a problem of constitution in the classical sense: in this sphere, according to Husserl, valid statements may be obtained only by a sequence of increasingly radical reduction steps, each of which is to consist in a certain ‘abstention’ (*epoché*). The first two steps, i.e., ‘phenomenological reduction,’ leading to the phenomena, and ‘eidetic reduction,’ leading to the essential structure of consciousness, carry us on to the transcendental sphere. Access to it, however, is only given by ‘transcendental reduction,’ which is the ‘absolutely universal reduction,’ opening up the transcendental subjectivity which constitutes the meaning of the world and its contents, and thereby all objecthood.

2.3 The Crisis of European Science and the Life-world

The period between World War I and the seizure of power by National Socialism in Germany abounds with diagnoses of crises, from the foundational crises in most sciences (claimed for psychology, sociology, economics, etc., following the foundational crisis of classical mathematics so dramatically exposed by H. Weyl in the 1920s) to a general crisis of the times, of intellectual and moral orientation, and of European (‘rational’) culture generally.

Husserl gave much thought to this phenomenon and made out a growing loss of confidence in the sciences hitherto regarded as the firm foundation of Europe’s rational culture. He analyzed this situation comprehensively in an article ‘Die Krisis der europäischen Wissenschaften und die transzendente Phänomenologie’ published in 1936 in a new and little-

known Serbian journal and later forming the first two parts of the monograph bearing the same title (Husserl 1954). The sciences, it seemed to him, while ‘functioning’ as well as ever technically, had lost sight of their origin as well as of their aims, and thereby lost their former role and status as a guide and a standard of reliable knowledge and moral responsibility. Husserl hoped for a renewal by recognition of the fact that all scientific knowledge and activity is grounded in and grows out of everyday life. This is a realm usually not reflected upon but capable of thorough analysis as our ‘life-world,’ the domain not only of the physically ‘given,’ but primarily of our social life. Our immediate experience is not that of objects given to our consciousness, but that of situations in which we find ourselves with other people, in social groups and institutions, participating in common meaningful action, handling (linguistic and other) symbols, and approaching with understanding even ‘the great symbolic systems such as language, myth, religion, art, etc.; all of which are essential elements of the *Lebenswelt* and therewith of highest interest to the social sciences’ (Schütz 1959, p. 96). The phenomenological analysis of these largely ‘pre-predicative’ constituents of the life-world was envisaged by Husserl in the *Crisis* treatise and begun in his late and posthumously published work *Erfahrung und Urteil* (1939), where he has recourse to the life-world and to the primacy of praxis over mere contemplation. But in front of them, he places the ‘pre-predicative’ evidence of perception and of judgment based on perception, and thereby once more a prelinguistic foundational domain. The new approach remained to be developed by others.

3. Contexts of Discussion and of Influence

Notwithstanding the fact that ‘Husserl literally “thought through his pen”’ (Embree et al. 1997, p. 327), entrusting an incredibly large part of his ideas to shorthand memos that remained unpublished during his lifetime (and part of them until now), his thinking did not develop monologically, without context. During Husserl’s student years, the careful study of Descartes, Leibniz, the British empiricists and Brentano was recommended to him by T. Masaryk (today less famous for his *Versuch einer konkreten Logik* of 1887 than for his later career as the first president of Czechoslovakia), and Husserl himself testifies to the influence of William James’s *Principles of Psychology* of 1890 on the shaping of his phenomenological thought. Having acquired some fame as a creative philosopher, Husserl corresponded with many experts in science and philosophy, and discussed his ideas at invited lectures at other German universities, and abroad in Vienna, Prague, Paris, Amsterdam, London and elsewhere—events certainly prepared by the translation of the *Logische Unter-*

suchungen and the *Ideen* into several foreign languages.

The beginning of the ‘phenomenological movement’ may be seen in the exchange of ideas with students and collaborators of T. Lipps in Munich, who learned about Husserl’s investigations through their fellow student J. Daubert after he met Husserl in Göttingen in 1902. In spite of having nothing in print, Daubert had considerable influence on the ‘Munich phenomenologists’ (among them M. Geiger, A. Pfänder, and A. Reinach), some of whom (like H. Conrad-Martius, A. Koyré, and E. Stein, later one of Husserl’s private assistants) came to Göttingen to work with Husserl. They became the leading members of the ‘Göttinger philosophische Gesellschaft’ and main authors of the *Jahrbuch für Philosophie und phänomenologische Forschung*, founded in 1913, which published some of the great contributions (phenomenological and other) to contemporary philosophy, such as Husserl’s *Ideen* in 1913, M. Scheler’s *Der Formalismus in der Ethik und die materiale Wertethik* (*Formalism in Ethics and the Material Ethics of Value*) in 1913–16, and M. Heidegger’s *Sein und Zeit* and O. Becker’s *Mathematische Existenz* (*Mathematical Existence*), both in 1927. After its discontinuation in 1930, the publications’ goals and role were taken up by the American journal *Philosophy and Phenomenological Research*, founded in 1940.

In Husserl’s most active period, private assistants helped organize his literary production, defended his positions (e.g. E. Fink against neo-Kantian critics), and prepared lectures and notes for publication (e.g., the *Zur Phänomenologie des inneren Zeitbewußtseins*, prepared by E. Stein and edited by M. Heidegger in 1928, and, posthumously in 1939, *Erfahrung und Urteil* by L. Landgrebe). In 1940, H. L. Van Breda saved most of Husserl’s *Nachlass* (including 40,000 pages of manuscripts, Husserl’s private library and much of his correspondence), which today forms the core of the Husserl Archives at the Catholic University at Louvain, Belgium. Philosophical and related work in the phenomenological spirit is, beyond this, upheld by ‘Phenomenological Societies’ in several countries.

4. Impact on the Social and Behavioral Sciences

Husserl’s own writings did not take up any topics directly relevant to the social and behavioral sciences and therefore had little influence on them during his lifetime. Yet, they greatly stimulated interest in the phenomenological method in other authors, who explored its applicability and fruitfulness within the social sciences, broadening and strengthening the phenomenological movement. Whereas the investigations of A. Reinach into the *a priori* foundations of civil law, and of M. Scheler into the origins of cultural development, are contributions to the philosophy of

the social sciences rather than to social sciences proper, the potentiality of a phenomenological approach was recognized early in the psychology of personality (especially in the psychology of the subconscious and in psychiatry), and in experimental psychology, where it opened up new vistas for social psychology and the psychology of perception (D. Katz, A. Metzger, M. Wertheimer); the development of gestalt psychology (K. Köhler, K. Koffka) and of K. Lewin's field theory are unthinkable without the impact of Husserl's phenomenology (see also MacLeod 1947). In the field of sociology, the *formal sociology* (or sociological formalism) of A. Vierkandt and G. Gurwitsch have drawn explicitly on phenomenological ideas, and more recently, sociological systems theory, social constructionism and ethnomethodology have been clearly influenced by phenomenological thought (P. Berger, H. Garfinkel, E. Goffman, T. Luckmann, Niklas Luhmann, I. Srubar).

According to A. Schütz, 'Husserl's signal contribution to the social sciences consists ... in the wealth of his analyses pertinent to problems of the *Lebenswelt* and designed to be developed into a philosophical anthropology' (Schütz 1959, p. 97). This development was brought about by Schütz himself, who met Husserl personally only in 1932, after publication of his own pioneering work *Der sinnhafte Aufbau der sozialen Welt* (*The Phenomenology of the Social World*, 1972), but established close intellectual contacts for the six remaining years of Husserl's life. Schütz became the towering figure of phenomenological sociology, and it is due to the considerable impact of his work that Husserl's phenomenology, beyond its historical merit of exhibiting the limits of naturalistic and atomistic thinking, has gained a lasting influence on the social and behavioral sciences of today.

See also: Hermeneutics, History of; Phenomenology in Human Science; Phenomenology in Sociology; Phenomenology: Philosophical Aspects; Philosophy, Psychiatry, and Psychology; Psychologism, History of; Reduction, Varieties of; Religion, Phenomenology of; Schütz, Alfred (1899–1959)

Bibliography

- Bell D 1990 *Husserl*. Routledge, London
- Embrete L et al. 1997 *Encyclopedia of Phenomenology*. Kluwer, Dordrecht, The Netherlands
- Herzog M, Graumann C F (eds.) 1991 *Sinn und Erfahrung. Phänomenologische Methoden in den Humanwissenschaften*. Asanger, Heidelberg, Germany
- Husserl E 1901 *Logische Untersuchungen*. Niemeyer, Halle, Germany (trans. into English, 1970 *Logical Investigations*. Routledge & Kegan Paul, London, and Humanities Press, New York)
- Husserl E 1913 (rev. 1976), 1950, 1952 *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie*. Vol. 1 Niemeyer, Halle, Germany; Vol. 2, Biemel W (ed.), Nijhoff,

- The Hague, The Netherlands; Vol. 3, Biemel M (ed.) Nijhoff, The Hague, The Netherlands (trans. into English 1980–89 *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy*. Vol. 1: Nijhoff, Holland, Vols. 2 and 3: Kluwer, Dordrecht, The Netherlands)
- Husserl E 1939 *Erfahrung und Urteil. Untersuchungen zur Genealogie der Logik*. Academia, Prague (trans. into English 1973 *Experience and Judgment. Investigations in a Genealogy of Logic*. Northwestern University Press, Evanston, IL, and Routledge & Kegan Paul, London)
- Husserl E 1954 *Die Krisis der europäischen Wissenschaften und die transzendente Phänomenologie: Eine Einleitung in die phänomenologische Philosophie*. Nijhoff, The Netherlands (trans. into English 1970 *The Crisis of European Sciences and Transcendental Phenomenology: An Introduction to Phenomenological Philosophy*. Northwestern University Press, Evanston, IL.) (Pts. 1 and 2 had already been published in the journal *Philosophia* (Belgrade) 1 (1936), 77–176)
- Husserl E 1950 ff. *Husserliana: Edmund Husserl, Gesammelte Werke*. Nijhoff, The Netherlands, and Kluwer, Dordrecht, The Netherlands
- MacLeod R B 1947 The phenomenological approach to social psychology. *Psychological Review* 54: 193–210
- Schütz A 1959 Husserl's importance for the social sciences. In: Van Breda H L, Taminiaux J (eds.) *Edmund Husserl 1859–1959: Recueil Commémoratif publié à l'Occasion du Centenaire de la Naissance du Philosophe*. La Haye, Nijhoff, The Netherlands, pp. 86–98
- Schütz A 1962–97 *Collected Papers*. Nijhoff, The Netherlands (Vol. I: *The Problem of Social Reality*, ed. Natanson M 1962; Vol. II: *Studies in Social Theory*, ed. Brodersen A 1967; Vol. III: *Studies in Phenomenological Theory*, ed. Schütz I 1966)
- Smith B, Smith W D (eds.) 1995 *The Cambridge Companion to Husserl*. Cambridge University Press, Cambridge, UK
- Wolf K H 1979 Phenomenology and sociology. In: Bottomore T, Nisbet R (eds.) *A History of Sociological Analysis*. Heinemann, London, pp. 499–556

C. Thiel

Hutterite Demography

The unique demography of the Hutterites—a contemporary Mennonite sect of Germanic origin who live in the Great Plains of the USA and Canada—has played an important role in the development of demographic theory and methods. They practice no birth control for religious reasons and their high fertility levels give reason to believe they do not intentionally control their fertility (Eaton and Mayer 1953). Their communally shared life-style, including the sharing of material resources, is also presumed to limit social or cultural variations in Hutterite fertility (Larsen and Vaupel 1993). Because of their uniformly high and apparently uncontrolled fertility, studies of the patterns and determinants of fertility among the Hutterites have helped to shape our theoretical expectations for populations that do not make any conscious efforts to limit childbearing. And, as a result,

the Hutterites have also been widely used in demographic methods as a fertility standard to which other populations are compared.

1. Hutterite Fertility

Many different estimates of fertility and fertility determinants among the Hutterites have been published in different studies (see Larsen and Vaupel 1993). All of these studies agree, however, that Hutterite fertility is exceptionally high. Eaton and Mayer (1953) reported a median completed family size of 10.4 children among married women over the age of 45 in 1950. Recalculations using the data published by Eaton and Mayer report a lower completed family size with a median of 9.4 children, and a mean of 8.97 among the same women (Lang and Göhlen 1985). These estimates are consistent with the modal completed family size of 10 children, and the maximum of 15 children, reported for the Hutterite sample analyzed by Larsen and Vaupel (1993). Even in this very high fertility population, however, fecundability and the probability of conception decline substantially with age. The monthly probability of conception is estimated to decline by about two-thirds between ages 20 and 35 (Larsen and Vaupel 1993), likely reflecting age-related influences such as higher probabilities of fetal loss, lower coital frequency, longer postpartum amenorrhea, etc., among older age women. Nonetheless, the age-specific influences on Hutterite fertility resulting in a decline of fecundability across ages appear to be largely biological in nature. Compared with other populations, the Hutterite population has exceptionally high fertility levels and the age-related decline in fecundability offers no apparent evidence of intentional fertility limitation.

2. The Hutterites as a Natural Fertility Standard

Because of their doctrines and behavior, the Hutterites are presumed to represent a *natural fertility* population. That is, as Henri (1961) argued, their child-bearing patterns reflect no efforts to limit fertility bound to the number of children already born, i.e., parity-specific fertility limitation. In fact, the Hutterites were one of the populations Henri (1961) used to define and illustrate the concept of natural fertility. He also ensured a special standing for the Hutterites as a high fertility standard when he found that they had the highest age-specific fertility levels of 13 historical populations he studied, all of which appeared to be natural fertility populations.

2.1 The Hutterites as a 'Standard Natural Fertility Population'—Coale's 'I' Indices

Hutterite fertility levels and age-specific fertility patterns have been used as standards in some of the most

widely used fertility measures of recent years. The Princeton European Fertility Project (Coale and Watkins 1986) popularized the use of Coale's (1969) 'I' indices which compare fertility in a population with that of the Hutterites through indirect standardization. Coale's index of overall fertility (I_o) compares the number of births to women in a population with the number of births these women would have had if they had experienced the same high age-specific fertility rates as married Hutterite women (i.e., 0.550 births per year among women aged 20–24, 0.502 to those aged 25–39, 0.447 to those aged 30–34, 0.406 to those aged 35–39, 0.222 to those aged 40–44 and 0.061 to those aged 45–59, see Henri 1961). Coale's index of marital fertility (I_g) similarly compares births to married women with the fertility they would have had if they had experienced the Hutterite age-specific marital fertility rates. And, Coale's index of illegitimate fertility (I_h) draws the same comparison for nonmarried women, again using Hutterite age-specific marital fertility rates as a standard. A final fourth index is intended to capture effects of the proportion married upon fertility (I_m) through comparing the fertility married women would have with that which all women would have, if both experienced Hutterite age-specific marital fertility rates.

2.2 Age Patterns of Hutterite Fertility as a Standard—Coale and Trussell's 'M' and 'm' Indices

Coale and Trussell's (1978) more formal mathematical modeling of age-specific marital fertility control also incorporates the fertility pattern of the Hutterites and Henri's (1961) other natural fertility populations as a standard for comparison. Coale and Trussell used 10 of Henri's natural fertility populations, including the Hutterites, to specify a standard age-specific pattern for natural fertility (see *Demographic Models*). The Coale–Trussell model attempts to parameterize the differences in overall levels and age patterns of fertility from these natural fertility expectations in populations with varying levels of fertility control. The level of fertility is then captured by two parameters of the model, Coale and Trussell's 'M' index of fertility levels and their 'm' index indicating departures from a natural fertility age pattern (1978).

2.3 Hutterite Fertility Determinants in Micro- and Macrosimulations

Not surprisingly, demographic simulation models have also made use of Hutterite fertility standards. Ridley and Sheps' (1966) initial (i.e., Cohort A) microsimulation of fertility, for example, used values of fecundability and postpartum nonsusceptibility which were derived from the Hutterite population. Bongaarts' more recent (1976) macrosimulation of fertility also incorporated Henri's (1961) natural fertility estimates as standards for age-specific patterns of

fecundity and for infecundity associated with sterility. In these simulations the Hutterite fertility standards are used to represent biological fertility parameters in the absence of fertility control.

3. Limitations of the Hutterite Fertility Standard

As the theoretical definition of natural fertility has been challenged, so have the measures of fertility control which have utilized the Hutterite fertility standard. Comparisons with a Hutterite standard population, such as Coale's I_0 index, may reflect differences in fertility which do not arise from parity specific control (Santow 1995). And, as in the case of Coale's ' M ' and ' m ' indices, such comparisons may also fail to capture the heterogeneity of fertility behavior in populations being compared with the natural fertility standards (Okun 1994), including the possibility of fertility control which is not closely tied to parity such as birth spacing. Although comparative indices and methods face many of the same limitations as any indirect standardization, incorporating the Hutterites as a constant standard population for empirically-based comparisons has nonetheless been of considerable advantage to demographers. Use of the Hutterite standard has allowed meaningful comparisons of fertility levels and patterns across widely different historical populations. The Hutterite standard has also provided indices with a clearly interpretable and unambiguous reference to an empirically observed high fertility standard. And most criticisms of the measures incorporating the Hutterite standard concern their interpretation in terms of underlying fertility behavior (e.g., diffuse birth spacing or parity-specific stopping) rather than their ability to index the levels and patterns of age-specific fertility across populations.

4. The Hutterites and Bio-demography

Underlying their use as a high-fertility standard in many different methods, the Hutterites in and of themselves offer a unique opportunity for demographers to explore biological aspects of fertility and fecundity. An absence of any evident parity-specific fertility control, along with very high fertility levels, makes the Hutterite population one which may be used to study biological determinants of fertility without the confounding influence of fertility limiting behaviors. However, a troubling concern in the study of historical populations that lack widespread evidence of fertility control is the extent to which there may be many subgroups within the population with different fertility levels and patterns arising from different behavior and sociocultural circumstances. The Hutterites' communal lifestyle and uniformly high fertility give no reason to suspect wide differences

in fecundability and fertility from social or cultural forces. And the uniformity of the population in many demographic behaviors (e.g., universal marriage and early ages at marriage, short durations of breast-feeding and early weaning, short birth intervals, absence of induced abortion, and low levels of premarital conception, etc.) is likely to minimize many sources of variation in fertility compared with other populations. This relative homogeneity, or absence of subgroup variation, is especially important to the recent use (and assumptions) of event history methods to estimate the underlying biological patterns of fecundability among the Hutterites. This allows researchers estimating fecundability with Hutterite data to assume the underlying patterns and variations they discover arise from biological variation (see Larsen and Vaupel 1993 for a review and example of such studies).

As further studies provide more refined estimates of Hutterite fertility and fecundity, these will likely continue to serve as empirically grounded standards for our demographic expectations in the absence of intentional fertility limitation. The theoretical importance of these expectations to demography is evident in the significant role which studies of the Hutterites have had in shaping the concept of natural fertility and debates over the nature of fertility limitation. Their methodological import is clear in the wide variety of methods already grounded in the Hutterite fertility standard.

See also: Demographic Data Regimes; Demographic Models; Demography, History of; Fertility and Culture: Anthropological Insights; Fertility Control: Overview; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations

Bibliography

- Bongaarts J 1976 A dynamic model of the reproduction process. *Population Studies* 30(1): 59-73
- Coale A J 1969 The decline of fertility in Europe from the French Revolution to World War II. In: Behrman S J, et al. (eds.) *Fertility and Family Planning: A World View*. University of Michigan Press, Ann Arbor, MI
- Coale A J, Trussell T J 1978 Technical note: Finding the two parameters that specify a model schedule of marital fertility. *Population Index* 44(2): 203-13
- Coale A J, Watkins S C (eds.) 1986 *The Decline of Fertility in Europe*. Princeton University Press, Princeton, NJ
- Eaton J W, Mayer A J 1953 The social biology of very high fertility among the Hutterites: The demography of a unique population. *Human Biology* 25: 206-64
- Henri L 1961 Some data on natural fertility. *Eugenics Quarterly* 8: 81-91
- Lang H, Göhlen R 1985 Completed fertility of the Hutterites: A revision. *Current Anthropology* 26(3): 395
- Larsen U, Vaupel J W 1993 Hutterite fecundability by age and parity: Strategies for frailty modeling of event histories. *Demography* 30(1): 81-102

- Okun B 1994 Evaluating methods for detecting fertility control: Coale and Trussell's model and cohort parity analysis. *Population Studies* 48(1): 193–222
- Ridley J C, Sheps M C 1966 An analytic simulation model of human reproduction with demographic and biological components. *Population Studies* 19(3): 297–310
- Santow G 1995 Coitus interruptus and the control of natural fertility. *Population Studies* 49(1): 19–43

D. L. Anderton

Hybridity

'Hybridity,' originally a biological term, has been used by authors in the social sciences, literary, artistic, and cultural studies to designate processes in which discrete social practices or structures, that existed in separate ways, combine to generate new structures, objects, and practices in which the preceding elements mix. During the 1980s and 1990s its use has increased, particularly in English and Spanish, whereas in other languages, such as French, the term coexists with adjacent concepts like *mestizaje* and syncretism, favored in French literature. Some authors include these classic terms in 'hybridization,' conveying greater flexibility to the term to include processes of intercultural confluence.

1. The Term Hybridity in Biology and in Social Sciences

The use of hybridity in humanistic and social science studies has been questioned due to the biological origin of the concept. The risk of transferring the sterility commonly associated with the term to society and culture is put forth. Even though this criticism can still be found in recent writings, it is the extension of a nineteenth-century belief, when hybridization was looked upon with suspicion because it was believed to harm social development. However, this opinion has changed since 1870 with Mendel's theory of the enhancement produced by genetic mixing and its twentieth-century examination (Olby 1984). Strictly considered, biology, particularly botany, teems with fertile hybridizations which take advantage of characteristics of cells from different plants to enrich the growth, resistance, quality, and economic and nutritional worth of food stuffs made from these combinations (Callender 1988). Hybridization of human DNA with bacteria to produce protein is also common; this is, in fact, the main means to generate insulin.

Adoption of the term by the social sciences is not legitimized by its biological background. Similarly to

other concepts taken from biology, for example, *reproduction*, for which a new elaboration had to take place in order to talk about social, economic, and cultural reproduction, the theoretical solidity and explanatory power of hybridity in social sciences depends upon the manner in which the term is relocated within networks of specific concepts referring to society and culture.

On the one hand, hybridity is linked with more or less classical notions of social thought that allude to the separation and mix among cultures and strata: race, ethnicity, *mestizaje*, miscegenation, syncretism, identity, multiculturalism, heterogeneity, creolization, transculturation, bricolage. On the other, its treatment and epistemological consistency depends upon the manner in which different authors link the concept with terms and theories, such as tradition, modernity, postmodernity, border, contradiction, globalization, that condition the understanding of hybridity.

The sociocultural construction of hybridity has taken place in several disciplines. As a linguistic process it was elaborated, among others, by Mikhail Bakhtin, who used the term to characterize the co-existence of learned and popular speech in the same group, or even in a single speaker, and by Homi Bhabha, who defines hybridity as a metonymy of presence and places it amidst power relations, noting differences between colonial and resistance hybridization, depending on whether the linguistic and literary mixes serve domination or subversion (Bhabha 1994). Some historians of the Americas speak of hybridity to explain the interweaving of pre-Columbian imagery with the colonizers' Spanish one and with the culture developed later, during modernity (Gruzinski 1999).

2. Mestizaje, Syncretism, Transculturation, Creolization, and Hybridization

Mestizaje and hybridization are as old as exchanges among peoples, migrations, and wars of conquest wars. The Tower of Babel is a story of *mestizaje*. Pliny the Elder spoke of hybridization when referring to the migrants that arrived in Rome in his day. Several historians and anthropologists have illustrated the decisive role of *mestizaje* in certain areas, such as the Mediterranean, since the days of Classical Greece (Laplantine and Nouss 1997). But the more intense—or researched—periods of interethnic and intercultural mixing happened with Europe's expansion into America.

The mix of Spanish and Portuguese colonizers first, then of English and French, with American Indians, to which the African slaves were added, made *mestizaje* the prevailing process within the societies of the so-called New World. Nowadays, less than 10 per cent of the population of Latin America is descended from

indigenous peoples alone (this percentage decreases in the US and Canada). Communities of European descent that have not mixed with native peoples are also a minority. Thus, the makeup of all of the Americas requires the notion of *mestizaje*, both biologically—production of phenotypes from genetic crossbreeding—and culturally: blending of European and native American habits, beliefs, and ways of thinking.

For a very long time, studies on physiognomic and chromatic aspects of *mestizaje* were privileged. Skin color and physical features are still critical in the ideological construction of subordination and discrimination; however, in the social sciences and in more democratic political thought, *mestizaje* presently centers on the cultural dimension of the combination of identities. In anthropology, cultural studies, and cultural policy making, the issue is presented as devising ways to share multicultural life styles (Kymlicka 1995).

Of the diverse crossings that take place among cultures, the religious ones have received the most attention. They are usually referred to by a specialized term: syncretism. European expansion in Africa, Latin America, and Asia gave way to multiple mixed forms where heterogeneous beliefs fused and new divinities emerged (the Virgin of Guadalupe in Mexico; the combination of African ancestors, indigenous figures, and Catholic saints in Brazilian Umbanda). Practically every religion has fundamentalist currents that seek to defend pure forms, but the magnitude of migration and the transcontinental spread of rituals and beliefs during the eighteenth and nineteenth centuries stressed these hybridizations and increased tolerance towards them to such a point that in countries such as Brazil, Cuba, Haiti, and the US, it was frequent to find a double or triple religious affiliation. An example may be a Catholic person participating in an Afro-American ritual or in a New Age ceremony. If we consider syncretism in a wider sense, such as the simultaneous adhesion to several belief systems, not only religious ones, the phenomenon grows remarkably, particularly among the many people who resort to indigenous or Asian medicine for certain ailments, to allopathic medicine for others, and even to Catholic or Pentecostal rituals. The syncretistic use of these health resources usually is combined with musical fusion and multicultural social organization systems, as is the case in Cuban *Santería*, Haitian Voodoo, and Brazilian *Candomblé* (Rowe and Schelling 1991).

Some of these phenomena were what Fernando Ortiz, Cuban historian and anthropologist, thought about when he proposed the term *transculturation* (Ortiz 1995/1947) to designate 'processes in which features are not mechanically aggregated, but give way to a new phenomenon, independent and original.' Bronislaw Malinowski's endorsement of that utterance in an introduction to Ortiz's book, and its future development by literary critics Angel Rama and

Mary Pratt in their study of travel writing, are still appealing; particularly in literature and in Anglo-Saxon and Latin-American cultural studies.

The word creolization has also been used to refer to intercultural combinations, especially in anthropology and linguistics. Rigorously, it designates the language and culture created by variations of the basic tongue with other languages in the context of the slave trade. It includes the mixtures French has undergone in America and the Caribbean (Louisiana, Haiti, Guadeloupe, Martinique) and in the Indian Ocean (Reunion, Mauritius), or Portuguese in Africa (Guinea, Cape Verde), in the Caribbean (Curaçao) and Asia (India, Sri Lanka). But whereas it presents paradigmatic tensions between orality and writing, learned and popular sectors, center and periphery, in a diversity *continuum*, Hannerz suggests its usage could be extended to the transnational field to name 'processes of cultural confluence' characterized 'by power, prestige, and material resources inequality' (Hannerz 1997). Even though he is not the only author to stress the inequality and discontinuity present in hybridizations, the emphasis he places on examining increasing flows between center and periphery, together with the asymmetries among markets, states, and educational levels, helps avoid considering *mestizaje* as simple intercultural homogenization and reconciliation (Hannerz 1996).

3. Hybridization, Modernity, and Globalization

3.1 Hybridization

Whereas the terms mentioned above are nearly used as synonyms by several authors, *mestizaje* takes precedence in studies devoted to the mix of races or ethnic groups, and syncretism to refer to religious fusion or traditional symbolic movements. Besides the fact that *mestizo* has no English equivalent (some authors include the Spanish word in their English writing), research carried out during the 1980s in Spanish, English, and occasionally in French and Portuguese adopted 'hybridity' to include diverse intercultural mixes: ethnic *mestizaje*, syncretism regarding beliefs, and other modern fusion in which arts and crafts integrate with industry, the cultured with the popular and traditional and with mass media messages (Bhabha 1994, García-Canclini 1995, Bernd and de Grandis 1995, Papastergiadis 1997, Webner and Modood 1997).

3.2 Modernity

Modernity, an expression of which are nation states, has coherently organized—sometimes in reality, others in appearance—the diversity of cultures. There

are fewer than 200 nation states which comprise more than 600 languages and some 5,000 ethnic groups. These states have had to organize multiculturalism through diverse policies, at the same time that they facilitated communication among ethnic groups and the different languages included in their territory. Even in the cases in which a single language was imposed over the 20, 50, or 100 groups that formed the nation, that common language adopts diverse modulations in the same English- or Spanish-speaking country when they combine or hybridize with the words, syntactic structures, and semantic references of the speakers of other tongues.

In the meantime, two other characteristics of modern development, transnational expansion of capitalism and of mass communication technologies, advance relationships and exchanges between local traditions and cultures of modernity. The crafts and celebrations of peasant groups travel, together with their producers, to cities and often to other countries, or they enter into contact with foreign tourists. Ethnic music, linked to specific rituals of a certain group, is not only taped in CDs, movies, and videos to be distributed worldwide; it is also mixed with classical and contemporary music, with jazz, rock, salsa, and other local and regional kinds of music. This is how the *chicha*, a mix of Andean and Caribbean rhythms was produced. Other examples include the jazzy reinterpretation of Mozart by the Afro-Cuban group Irakere, the re-creation of English, Indian, and other music by the Beatles, Peter Gabriel, and many other musicians.

The rigid borders established by nation states became porous. Few cultures can be defined today as stable units, with precise boundaries based on the occupation of a confined territory. But the multiple opportunities to hybridize neither imply indeterminism nor unrestricted freedom. Hybridization happens under specific historical and social conditions, in the midst of production and consumption systems that sometimes operate as constraints, as can be witnessed by the lives of many migrants. One of the social entities that support but also condition hybridization are the cities. The multilingual, multicultural megalopolis, for example, Paris in the nineteenth and twentieth centuries, Vienna at the beginning of the twentieth century, London, Berlin, New York, Los Angeles, Buenos Aires, Sao Paulo, Mexico, and Hong Kong are studied as centers where hybridization fosters greater conflict as well as greater cultural creativity (Appadurai 1996, Hannerz 1996).

3.3 Globalization

Globalization stresses these trends of modernity by creating world markets of material goods and money, messages, and migrants (workers, tourists, exiles).

These constant streams and interactions have weakened borders and frontier posts; also the autonomy of local traditions. At the same time, they foster many more forms of hybridization than they ever did in the past: productive, commercial, communicational, and in consumption habits. To the characteristic modalities of hybridization, which came into existence due to migrations and trips, as well as to the educational integration policies by nation states, the mixtures generated by cultural industries have to be added. Industrialization encompasses not only the mass media (radio, movies, television, video), but also industrial production for massive audiences in international exhibitions, folkloric groups' tours, and globalized theme parks such as Disney's (Harvey 1996).

However, the main globalization analysts note that it not only integrates and generates *mestizajes*; it also segregates, produces new inequalities, and stimulates differentialist reactions. Identity alliances emerge (Latins or Arabs in the US or in Europe) as do organizations of regional integration (European Union, NAFTA, Mercosur), which in part aim at strengthening partial groupings to counteract the indiscriminate interaction of globalism (Beck 2000). Some of the actors find resources to resist both globalization and hybridization at once in these alliances.

3.4 The Future

Several anthropologists and cultural studies specialists maintain that the generalized relevance of hybridization requires changes in the ways social sciences are formulated. It is impossible to analyze large units, that is, nations, states, transnational enterprises, without noticing their internal diversities and modulations, as often happens in disciplines devoted to analyzing at the macro level, such as sociology and economics. It is not acceptable either that anthropology, traditionally dedicated to small communities, continues to see solely their homogeneity, and overlooks the heterogeneity and hybridization caused by the transnational goods and messages that intersect these communities. To be able to politically negotiate diversity and hybridizations, we need a '*mestizo* way of thinking' (Laplantine and Nouss 1997, Gruzinski 1999).

Fundamentalisms are not the only opposition to religious syncretism and intercultural *mestizaje*. There is an extended resistance to accepting hybridization because it generates insecurities within cultures and conspires against their ethnocentric self-esteem. It also presents a challenge for modern analytical thought, used to the binary separation: civilized, wild; national, foreign. This schematism leaves aside frequent ways in which culture is shared nowadays, for example, take

someone who is Brazilian by nationality, Portuguese by tongue, Russian or Japanese by origin, and Catholic or related to Afro-American religions.

Hybrid thought will not look forward to working with radically different, homogenous, compact units as much as with intersections, transitions, and transactions; less with identity than with heterogeneity and multiple senses of belonging. To understand the contemporary world it is indispensable to study how cultures mix, to what rhythms and modalities, what they lose and gain in the process, how they agree to cohabit, and how they negotiate to keep the difference alive. To comprehend hybridity and to learn to be cosmopolitan, 'multicultural gourmets', capable of traveling in diverse cultures and savoring the differences, has become necessary to fathom globalization and the ethnic and national societies' crisis. In brief, to understand society.

See also: Creolization: Sociocultural Aspects; Globalization and World Culture; Multiculturalism, Anthropology of; Multiculturalism: Sociological Aspects; Paradigms in the Social Sciences; Religion: Culture Contact; Specialization and Recombination of Specialties in the Social Sciences; Syncretism

Bibliography

- Appadurai A 1996 *Modernity at Large: Cultural Dimensions of Globalization*. University of Minnesota Press, Minneapolis, MI
- Barbero J M 1987 *De los medios a las mediaciones*. Ediciones G., Gili, México
- Beck U 2000 *What is Globalization?* Polity Press, Oxford, UK
- Bernard C 1993 *Altérités et métissages hispano-américains*. In: Descamps C (ed.) *Amériques Latines: une altérité*. Editions du Centre Pompidou, Paris
- Bernd S, de Grandis R 1995 *Imprévisíveis Américas: Questões de Híbrida (ao Cultural nas Américas)*. Sagra-Luzzatto, Porto Alegre, Brazil
- Bhabha H K 1994 *The Location of Culture*. Routledge, London and New York
- Callender L A 1988 Gregor Mendel: An Opponent of Descent with Modification. *History of Science* **XVI**
- Cornejo Polar A 1997 Mestizaje e hibridez: los riesgos de las metáforas. Apuntes. [Miscegenation and Hybridity: The Dangers of Metaphor. Notes] *Revista Iberoamericana* **LXIII** (180): 341–4
- García-Canclini N 1995 *Hybrid Cultures*. University of Minnesota Press, Minneapolis, MI
- Gruzinski S 1999 *La pensée métisse*. Fayard, Paris
- Hannerz U 1996 *Transnational Connections*. Routledge, London
- Hannerz U 1997 Fluxos, fronteiras, híbridos: Palavras-chave da antropologia transnacional. *Maná, Rio de Janeiro* **3**(1): 7–39
- Harvey P 1996 *Hybrids of Modernity*. Routledge, London
- Kymlicka W 1995 *Multicultural Citizenship*. Clarendon Press, Oxford, UK
- Laplantine F, Nouss A 1997 *Le Métissage*. Dominos, Paris
- Olby R C 1984 Historiographical problems in the history of genetics. *Riv. Stor. Sci.* **1**(1): 25–38

- Ortiz F 1995/1947 *Cuban Counterpoint*. Duke University Press, Durham, NC
- Papastergiadis N 1997 Tracing hybridity in theory. In: Webner P, Modood T (eds.) *Debating Cultural Hybridity*. Zed Books, Atlantic Highlands, NJ
- Rowe W, Schelling V 1991 *Memory and Modernity*. Verso, London
- Webner P, Modood T (eds.) 1997 *Debating Cultural Hybridity*. Zed Books, Atlantic Highlands, NJ

N. García-Canclini

Hypertension: Psychosocial Aspects

Hypertension is a complex disorder of physiological regulation involving many interacting biological and environmental factors. Although the prevalence of hypertension has decreased over recent years, the latest US survey placed the age-adjusted prevalence (> 140/90 mmHg) at 20.4 percent. Interest in the role of psychosocial factors in hypertension has a long history beginning with theories about social stress, affective processes, and individual behavioral dispositions in its etiology. Noteworthy are the early psychosomatic writings about the significance of anger expression and suppression for high blood pressure.

1. Direct and Indirect Evidence

Numerous studies provide evidence of a definite role for psychosocial factors in hypertension. Direct evidence is derived from cross-sectional studies of individuals with hypertension compared to control samples and longitudinal studies to determine what factors are associated with the development of hypertension. Indirect evidence comes primarily from studies of psychosocial factors affecting blood pressure in healthy individuals with or without various predisposing characteristics. Psychosocial factors are also relevant to the degree of adherence to medications, engagement in health-related behaviors affecting risk for hypertension, effectiveness of antihypertensive medications, and nondrug methods of managing hypertension based on psychosocial concepts and methods.

2. Blood Pressure Reactivity and the Sympathetic Nervous System

Elevations in blood pressure can be elicited by many different stimuli and circumstances involving psychological and social factors. Examples are exposure to noxious, painful, threatening, or otherwise stressful

stimuli and events, participation in demanding tasks requiring repeated behavioral and/or mental adjustments, continuous striving and active coping, frustration in achieving goals, mental effort especially in solving difficult problems, momentary moods and behaviors involving fear, anger, and other emotional states, inhibition or suppression of emotional states and reactions, conflicts about the expression of angry feelings or other emotions, and interpersonal and group conflict. Socialcultural norms define what are appropriate and desired reactions and behaviors. The extent to which blood pressure is affected is determined by the balance between demands placed and the individual's resources and ability to cope with the demands.

A common theme in the literature is that individual differences in cardiovascular reactivity to various stimuli and events serve as a risk factor for hypertension and cardiovascular disorders in general. Increased sympathetic nervous system activity and associated hormonal changes is considered the neural source of excessive cardiovascular reactivity (Folkow 1989). In established hypertension, sympathetic activity is increased, particularly in individuals in whom the elevated blood pressure appears to have a basis in the nervous system ('neurogenic' hypertension) (Esler et al. 1977). Moreover, most antihypertensive medications appear to regulate blood pressure by interfering with mechanisms of sympathetic activation. Low parasympathetic activity (vagal tone) may also be a factor in hypertension and may contribute by slowing cardiovascular recovery from stress.

The large body of experimental evidence linking excessive cardiovascular reactivity in humans to established hypertension is compelling on empirical as well as theoretical grounds (Light 1987). Moreover, psychosocial factors are also associated with higher casual or ambulatory levels of blood pressure, suggesting that the transformation from excessive reactivity to a chronic elevation may take place with repeated instances of blood pressure increases. Thus, in some samples of subjects, during waking hours and even during sleep, average levels of blood pressure have been found positively associated with such personality traits as hostility and defensiveness (Shapiro et al. 1993). Physiological models have been developed to describe the cardiovascular dynamics underlying the transformation from early heightened reactivity to later established higher levels of blood pressure.

3. Methodological Complications in Human Studies

Despite the vast literature, debates continue about the conclusiveness of the evidence supporting the cardiovascular reactivity hypothesis based on responses to laboratory stressors. The findings are not always consistent or easily integrated. Individuals vary widely

in their blood pressure responses to various circumstances. Variations in the kinds of individuals and methods chosen for study may result in different findings and conclusions. Laboratory settings may impose constraints on the participants so that they may not respond to challenges and threats with their preferred method of coping. Individuals also vary in the degree to which their responses persist over time for any one provocation or generalize across provocations. The association between perceptions of stress and the blood pressure response to such stress is not consistent because some individuals tend to report not feeling stressed, anxious, or angry while at the same time showing increased physiological reactivity to provocations or elevated blood pressure levels in general. These individuals either generally do not feel negative emotions or wish to make a good impression on other people. This characteristic has been labeled 'defensiveness.' Finally, many individuals learn to cope effectively with stress or to keep their emotional responses under control. For example, experienced air traffic controllers or other professionals dealing with emergencies do not necessarily show elevated levels of blood pressure (Sega et al. 1998). By taking these various factors into account, further sense may be made out of some of the apparently inconsistent findings, thus adding to confidence in the soundness of the reactivity model.

Thus, the ability to predict individual differences in reactivity and in the development of sustained high levels of blood pressure and established hypertension has been improved by considering other psychological factors such as various personality traits and coping styles (anger/hostility, defensiveness and emotional suppression, denial and repression, anxiety, lack of control and helplessness, Type A behavior pattern, depression, pessimism) (Johnson et al. 1992, Jorgensen et al. 1996, Suls et al. 1995). In the social domain, contributing factors are low socioeconomic status, social disintegration and family disruption (poverty, high crime rates, high divorce rates), occupational stress (job strain, job dissatisfaction, high job demands, little independence in one's work duties or latitude in the ability to make decisions), low levels of family and social support, social isolation, taxing responsibilities such as providing care to disabled family members, marital conflict, minority social status, certain ethnicities, and certain gender roles (Colhoun et al. 1998, Pickering et al. 1998). For example, anger and hostility do not necessarily have the same implications for men and women.

4. Psychobiological Models

The ability to predict whether repeated elicitations of increases in blood pressure will lead eventually to chronically elevated blood pressure and to cardiovascular disease is probably best improved when

certain biological factors are included in the test models, such as obesity, physical inactivity, poor physical conditioning, salt sensitivity, excitable autonomic nervous system, deficiencies in certain vitamins and minerals, intake of sodium and potassium, presence of other disorders such as diabetes, high glucose, insulin, or lipid levels, certain racial characteristics, and other genetic or constitutional predispositions such as family history of hypertension and cardiovascular disease. For example, in a study of occupational stress in nurses, those nurses for whom both parents had hypertension and who were hostile, defensive, or anxious had elevated ambulatory blood pressure, which was highest during work days (Goldstein and Shapiro 2000). Other psychosocial, life style, and habitual behavior patterns need to be considered as they are relevant to one's diet, activity level, and intake of caffeine, alcohol, and other substances. For example, high levels of social stress or certain personality traits like a high level of trait hostility may be associated with unhealthy health habits such as smoking, low levels of exercise, poor physical conditioning, and poor nutrition.

5. Diagnostic Issues

Psychological factors are relevant to the diagnosis of hypertension. Some patients may show elevations in blood pressure during a doctor's visit and evaluation but not at other times ('white coat effect'). This effect has been related to an alerting response to the examination or anxiety about one's health. In one study, trait hostility was found relevant to this effect. Those individuals who were highly hostile were consistently high in blood pressure during the examination and over the course of the day, whereas low hostile individuals were more likely to show elevated pressure only during the evaluation and not during the day (Jamner et al. 1993). By having the patient record blood pressure at home, useful information is obtained on blood pressure variability at different times and occasions that may guide decisions about treatment and increase the patient's adherence to medications or otherwise facilitate pressure reduction. Reliable devices are now available for this purpose.

6. The Problem of Blood Pressure Variability

Blood pressure can be quite variable both in the short term with a range of 5 mmHg or more over the course of a min and 30–40 mmHg or more over the course of waking hours during the day. These facts provided the impetus for the development of portable automatic 24-hour blood pressure recorders, initially used in the diagnosis of hypertension and the evaluation of drug effects and more recently in the investigation of blood pressure in different settings (Shapiro et al. 1993). Ambulatory blood pressure recording methods but-

tressed by information obtained from diaries of self reports of moods and the situations encountered in everyday life offer many advantages, including a large number of blood pressure measurements under different conditions (work vs. nonwork or off days, sleep vs. waking periods) and in a variety of repeatedly occurring situations, locations, and activities. Valuable data have been obtained from studies of special populations: paramedics, traffic control agents, secretaries, laboratory technicians, emergency dispatchers, nurses, bus drivers, college students. The research is challenging because of the complexities and uncontrolled nature of the observations, but novel analytical tools show great promise.

7. Psychosocial Factors Affecting Treatment

Psychosocial factors play a role in the treatment and management of hypertension. The response to anti-hypertensive drugs may partly depend on psychosocial factors, although research is sparse on this topic and these factors are not usually considered in the choice of medications used to treat patients. In one example, individuals with high levels of hostility were found less responsive to diuretics than nonhostile individuals, suggesting that the former might be more responsive to agents directly affecting the sympathetic nervous system. Ethnicity can affect the response to drug and non-pharmacologic treatments of hypertension (Brownley et al. 1999). Psychological factors influence degree of adherence to medications and may affect the degree to which side effects of medications occur and affect quality of life.

A great deal of research has been devoted to the evaluation of nondrug treatments of hypertension derived from psychological principles (Silverberg 1990). These procedures seem most suitable as alternatives in patients with mild hypertension, especially those who are free of significant cardiovascular disease risk factors (family history, high cholesterol level, smoking) and in other patients as a means of reducing the amount of medication needed for blood pressure control or completely eliminating the need for such medication. Reducing drug requirements may be of special value in dealing with undesired side effects of medications in some patients. The methods include progressive relaxation, simple relaxation, biofeedback-guided relaxation, biofeedback training involving blood pressure, heart rate, and other cardiovascular measures, biofeedback training using measures associated with sympathetic nervous system functioning, cognitive behavioral methods of facilitating coping with stress and dealing with anger ('stress management training'), hypnosis, and autogenic training. As yet there is no consensus about the effectiveness of these procedures, and major trials remain to be done. Other traditional methods such as yoga and meditation may be useful as a means of increasing awareness and control of mental and bodily functioning. Regular physical exercise can

facilitate blood pressure reduction and improve the coronary circulation. Whether or not people exercise, avoid or take medications, or use herbal, homeopathic, or various nondrug methods is determined by many different psychological, social, and cultural factors.

8. Summary and Conclusions

From the standpoint of the social and behavioral sciences, hypertension has attracted a great deal of attention. It forms the backbone of the growing fields of health psychology and behavioral medicine, and benefits greatly from the integration of methods, concepts, and findings of psychology, psychophysiology, social psychology, physiology, and biological and medical sciences. The evidence is supportive of the role and significance of psychosocial factors in the etiology, course, and treatment of hypertension. These factors include personality and coping styles, social and occupational stress, social status, family and social support, ethnicity, and gender roles. This knowledge has the potential for deepening our understanding of the mechanisms involved in hypertension, improving the effectiveness of both drug and non-drug approaches to treatment, and facilitating efforts to prevent the disorder. There is great need for further systematic research and especially for theoretical and empirical integration of psychosocial and biological knowledge of the disorder.

See also: Biopsychology and Health; Cardiovascular Conditioning; Neural Substrates; Coronary Heart Disease (CHD), Coping with; Coronary Heart Disease (CHD); Psychosocial Aspects; Coronary-prone Behavior, Type A; Gender and Cardiovascular Health; Stress and Cardiac Response; Stress and Health Research

Bibliography

- Brownley K A, Hurwitz B E, Schneiderman N 1999 Ethnic variations in the pharmacological and non-pharmacological treatment of hypertension. *Human Biology* **71**: 607–39
- Colhoun H M, Hemingway H, Poulter N R 1998 Socio-economic status and blood pressure: An overview analysis. *Journal of Human Hypertension* **12**: 91–110
- Esler M, Julius S, Zweifler A, Randall O, Harburg E, Gardiner H, De Quattro V 1977 Mild high-renin essential hypertension: Neurogenic-human hypertension. *New England Journal of Medicine* **296**: 405–11
- Folkow B 1989 Sympathetic nervous system control of blood pressure: Role in primary hypertension. *American Journal of Hypertension* **2**: 103S–111S
- Goldstein I B, Shapiro D 2000 Ambulatory blood pressure in women: Family history of hypertension and personality. *Psychology, Health, & Medicine* **5**: 227–40
- Jamner L D, Shapiro D, Hui K, Oakley M E, Lovett M 1993 Hostility and difference between clinic, self-determined and ambulatory blood pressure. *Psychosomatic Medicine* **55**: 203–11
- Johnson E H, Gentry W D, Julius S (eds.) 1992 *Personality, Elevated Blood Pressure, and Essential Hypertension*. Hemisphere, Washington, DC
- Jorgensen R S, Houston B K, Kolodziej M E, Schreer G E 1996 Elevated blood pressure and personality: A meta-analytic review. *Psychological Bulletin* **120**: 293–320
- Light K C 1987 Psychosocial precursors of hypertension: Experimental evidence. *Circulation* **76**: 167–76
- Pasic J, Shapiro D, Jamner L D, Hui K K 1994 Hostility and the response to diuretic in mild-to-moderate hypertension. *American Journal of Hypertension* **7**: 503–8
- Pickering T G, Devereux R B, James G D, Gerin W, Landsbergis P, Schnall P L, Schwartz J E 1997 Environmental influences on blood pressure and the role of job strain. *Journal of Hypertension* **14**(suppl.): S179–S185
- Sega R, Cesana G, Costa G, Fennanio M, Bombelli M, Mancia G 1998 Ambulatory blood pressure in air traffic controllers. *American Journal of Hypertension* **11**: 208–12
- Shapiro D, Jamner L D, Goldstein I B 1993 Ambulatory stress psychophysiology: The study of 'compensatory and defensive counterforces' and conflict in a natural setting. *Psychosomatic Medicine* **55**: 309–23
- Silverberg D S 1990 Non-pharmacological treatment of hypertension. *Journal of Hypertension* **8**(suppl.): S21–S26
- Suls J, Wan C K, Costa P T Jr. 1995 Relationship of trait anger to resting blood pressure: A meta-analysis. *Health Psychology* **14**: 44–56

D. Shapiro

Hypnosis, Neural Basis of

Modern conceptions of hypnosis can be traced directly to the work of Franz Anton Mesmer, a Viennese physician working in Paris just prior to the French Revolution. Within 30 years of Mesmer's work, most of the basic hypnotic phenomena from amnesia to pain control to hallucinations had been reported (see Ellenberger 1970). James Braid, a Scottish surgeon, coined the term hypnosis in the 1840s. Nowadays, hypnosis refers to the acceptance of a social role, usually involving deep relaxation, and engagement in fantasy. During hypnosis a lowering of critical judgment permits a wide range of behavioral, cognitive, and affective responses to suggestion (see, for example, Orne 1970). This article addresses the question of whether response to hypnosis can be related to specific neural mechanisms or events.

During the last two hundred years, a variety of exaggerated and mistaken claims have been made for hypnosis. Claims have ranged from enhanced psychic abilities to past life regression, to revived memories of alien abduction. In the past, such claims have resulted in periods of scientific disrepute for hypnosis. Nevertheless, during the last century fine scientists (e.g., Hull, Hilgard, Orne, Barber, Sarbin and their students) have been drawn to study hypnosis (see Sheehan and Perry 1976). As well as elucidating the domain of hypnosis, their careful investigations have increased understanding of the potential pitfalls that appear in

experimentation with human research participants (see Orne 1959, 1970).

An observer, watching a demonstration of hypnosis with a highly hypnotizable research participant, sees a series of simple verbal suggestions that result in relatively spectacular alterations of behavior, thought, emotion, and perception. For example, the hypnotized burn patient lies quietly and reports no pain, seemingly relaxing while his dressings are changed: a highly painful procedure (Hilgard and Hilgard 1975). The laboratory research participant asked to re-experience the distant past seems to become childlike, entirely captured by the delusion that he or she is many years earlier. Asked to hallucinate the absence of an obstacle between where the participant sits and some other place, the participant claims to see nothing in the intervening space. But when asked to walk to that place (and thus, through the obstacle whose absence is being hallucinated), the participant walks around the obstacle without seeming to notice doing so. Asked to look back and see that there is nothing there, there is easy agreement with no seeming sense that walking around the obstacle has just contradicted this statement. Asked to forget all that has happened until a pencil is tapped twice, the participant will later 'awaken' with amnesia that is (almost always) temporary and reversible.

Thus, in response to brief verbal suggestions, the hypnotized individual seems to see, hear, feel, smell, and taste in apparent contradiction to the stimuli actually present (see Orne and Hammer 1974). Memory, the sense of volition, mood, and even awareness of self may be altered. With appropriate suggestions, such effects may be extended into the posthypnotic period.

Hypnotic phenomena are easy to elicit and robust. A graduate student, given two or three hours training and a standardized script, such as one of the *Stanford Hypnotic Susceptibility Scales*, can elicit all the major hypnotic phenomena in a psychology laboratory (Weitzenhoffer 1959, 1963; see also Shor and Orne 1962). There they can be studied under controlled conditions. Note that for all practical purposes, hypnotic suggestions are likely to be carried out only in settings seen as culturally appropriate. These include therapeutic, classroom, entertainment, research, and forensic settings (Orne 1970).

1. Problems with Seeking Psychophysiological Concomitants of Hypnosis

People commonly speak of being 'hypnotized' or in a hypnotic 'trance.' The underlying assumption is that hypnosis is a distinct, altered state of consciousness. Most early theorizing (and much current clinical thought) about the hypnotic state employed circular reasoning that goes something like this: Q: Why do you say that person is hypnotized? A: Because she is

responding to hypnotic suggestions. Q: Why does she respond to hypnotic suggestions? A: Because she is hypnotized.

As Orne (1959) showed us, the hypnotized participant acts as shared cultural expectations suggest. In fact, the deeply relaxed hypnotic participant lying or sitting quietly while listening attentively to the hypnotist's voice manifests only one of many forms hypnosis has assumed. For example, in the late eighteenth century, Mesmer's Parisian patients sometimes had pseudo-epileptic seizures during hypnosis. In the late nineteenth century, Charcot's Parisian patients demonstrated first catalepsy, then lethargy, and finally somnambulism (but generally did not have seizures) when deeply hypnotized. At the same time in the south of France, Bernheim's hypnotized patients simply restfully reclined (see Ellenberger 1970). Most modern research participants and patients sit quietly and look very relaxed during hypnosis, although they may also be hypnotized with an 'active-alert' induction, involving, for example, pedaling a stationary bicycle vigorously (see Banyai and Hilgard 1976).

Obviously, peripheral neural activity during these varying hypnotic activities differs widely, reflecting the activity, not hypnosis. Nor has there been any evidence of brain activity unique to hypnosis *per se*. Nor is there evidence of specific brain activity that accompanies the entire hypnotic experience. After 50 years of modern research, the lack of evidence for any unique psychophysiological marker of a hypnotic trance state is one of the factors that have led researchers largely to abandon the concept of a hypnotic trance as a causal explanation for response to hypnotic suggestions (see Kirsch and Lynn 1995).

Concepts such as response expectancies, role involvement, self-delusion, and cognitive abilities are currently used to explain hypnotic responses. This does not mean that hypnotizing people does not have specifiable effects or that the experiences of highly hypnotizable subjects are consciously faked or otherwise theoretically unimportant. Rather, hypnotic effects result from making suggestions to willing participants of differing abilities and from the expectations and cognitive models that a hypnotic induction engenders both for the hypnotist and the participant (see Sheehan and Perry 1976).

At present, it is clear that the ability to respond to hypnotic suggestions lies largely in the subject (and not in the skill of the hypnotist as was previously held). Reports of being hypnotized or in a trance are seen as reflecting a subjective experience that is one of the many effects of cognitive and emotional set, social and historical setting, and suggestion (see Kirsch and Lynn 1995). And while all subjective experiences must have some neurophysiological substrate, psychology as a field of study has not reached the point where many can be pinpointed. Nonetheless, if phrased properly, reasonable questions about the psychophysiological concomitants of hypnosis can be asked.

2. Psychophysiological Differences in Response to Hypnotic Induction

A discussion of the neural basis of hypnosis may be phrased in terms of answers to three questions. First, are there measurable differences in brain activity in response to hypnotic induction among those who subsequently respond strongly to hypnosis in comparison to those who do not? Second, when measured outside hypnosis, are there overall psychophysiological differences between participants who usually respond to hypnotic suggestions for major changes in perception, memory, and/or cognition and those who do not? Third, are there patterns of brain activity associated with response (vs. lack of response) to specific hypnotic suggestions? While there are interrelated issues among these three questions, it is probably least confusing to look at these questions sequentially.

First, does hypnotic induction result in measurably different neural patterns among those who subsequently respond to the more spectacular hypnotic suggestions (those calling for hallucinatory or major cognitive changes) and those who do not? The evidence, while mixed, favors a tentative 'Yes, but for reasons that render the difference unimportant.' Brain function studies are divided between those that find highly hypnotizable, well lateralized, right-handed research participants shifting to greater relative right-hemisphere activation after hypnotic induction and studies that find no differences. The lack of consistent findings may be methodological: studies finding differences tend to use ratio measures that are meant to be sensitive to small differences (see, for example, LaBriola et al. 1987). Studies that do not find differences after induction tend to look at absolute values of measures of brain activation, a less sensitive measure (Graffin 1995). So there is some evidence that successful induction of hypnosis often involves a 'shift to the right' a finding in line with task performance measures of the effects of induction.

Why is this unimportant? Activation changes in response to induction are relatively small and disappear once other suggestions are made or tasks required. Relative hemispheric activation, with or without hypnosis, depends on the mental task participants are asked to perform. So the 'shift to the right' seemingly reflects a response specific to the 'task' of induction. And that response might well change were a different type of induction used (see Crawford and Gruzelić 1992).

3. Psychophysiological Differences between Highly Hypnotizable Participants and their Less Hypnotizable Counterparts?

Many studies distinguish between highly hypnotizable subjects and their less hypnotizable peers. Most modern scales of hypnotic responsiveness comprise

standardized, relaxation oriented inductions followed by 5 to 12 hypnotic suggestions that are each scored pass/fail based on an objective criterion. Entire hypnotic scales are often administered by tape recorder and both experimenter-scored individual formats and self-scored group formats are available. Total score, the measure of each participant's hypnotizability, is simply number of suggestions passed (see Kihlstrom 1985).

All social scientists use Pearson's reliability coefficient, a coefficient that varies from +1.00 (relative scores on two measures are identical) to 0.00 (no linear relationship between two sets of scores) to -1.00 (relative scores are identical in size but opposite in sign). Hilgard and his colleagues have found that test-retest reliability of hypnotizability scales usually averages about +.90 over periods of one week and about +.70 over periods from 10 to 25 years (Piccione et al. 1989). These data indicate that hypnotizability is a highly stable difference among individuals. Among personality traits, the reliability of hypnotizability measurement is equaled or exceeded only by multi-task measures of intelligence.

Hypnotizability is not easily modifiable and governs the large majority of the variance in most experimental studies of hypnotic phenomena. Moreover, what has come to be called 'hypnotizability' is observable in response to suggested alterations in perception, memory, and/or cognition with or without formal hypnosis (see Kihlstrom 1985).

There are two fairly well established psychophysiological correlates of hypnotizability. First, highly hypnotizable research participants (highs) show higher levels of resting theta activity than their less hypnotizable counterparts. Spontaneous EEG activity may be viewed as constituted of waveforms of different frequencies (delta -0 to 3 Hz, theta -4 to 7 Hz, alpha -8 to 13 Hz, beta -14 to 30 Hz). Waking EEG records are largely made up of alpha and beta activity, with a small percentage of theta activity. When measured during rest, this small proportion of theta activity is slightly higher among the more highly hypnotizable participants, especially in recordings from the more anterior brain regions. While no convincing explanations for this phenomenon have yet been advanced, the notion that the theta differences between highly hypnotizable participants and their less hypnotizable counterparts are related to attentional abilities is the most popular hypothesis (see Graffin et al. 1995, Sabourin et al. 1990).

Second, if a variety of other factors are controlled (e.g., task difficulty, familial history of right handedness), well lateralized, right-handed research participants will show greater relative left cerebral hemispheric activation during verbal and mathematical tasks and greater right hemisphere activation during mental imagery. This pattern of task specific hemispheric asymmetry is seen more clearly among highly hypnotizable participants than among their less

hypnotizable counterparts. Highs show greater relative left cerebral hemisphere activation during verbal and mathematical tasks than do less hypnotizable participants. Similarly, highs show greater relative right hemisphere activation during imagery than do low or moderately hypnotizable participants. (The inclusion of moderately hypnotizable subjects provides some control for compliance. Moderates show the same level of compliance to behavioral suggestions as the highs, but are largely unresponsive to hallucinatory suggestions.) These findings have been interpreted as showing greater cognitive flexibility for highs and are in line with some findings in studies using task performance measures (see Crawford and Gruzelić 1992).

4. Psychophysiological Correlates of Response to Specific Hypnotic Suggestions

The third question is whether psychophysiological differences can be found between those responding to specific hypnotic suggestions and those who fail to respond. This question can be asked in respect to hypnotic phenomena of central interest to experimental psychologists (e.g., hypnotic hallucinations), clinicians (e.g., healing of skin and immune system related disorders), or both (e.g., hypnotic analgesia).

Until the recent advent of functional magnetic resonance imaging (fMRI), activity in the normal waking human brain has been studied, for the most part, using measures of spontaneous and evoked electrical activity in the brain. Researchers usually utilize one of three electroencephalographic (EEG) technologies in which brain activity is recorded from the surface of the skull using standardized electrode placements. Obtaining EEG signals is similar to making an electrocardiogram except that EEG signals are measured in microvolts rather than millivolts signals studied with an EKG. (Since much greater amplification is required, EEG recording requires great care in avoiding electrical, movement, and other artifacts.)

Once having obtained an artifact free EEG signal, researchers digitize and transform it. Most frequently they use signal processing, mathematical (Fourier) transformations to determine the proportion of time in which alpha, beta, and theta activity occur. Alternatively, one may integrate the total area under the curve of an EEG record and compare total area at various sites in the brain during different tasks. In both methods one examines the signals from many recording periods (usually of 1 second to 10 minute duration) over minutes or hours of total recording time (see Karlin et al. 1980, Perlini and Spanos 1991, Sabourin et al. 1990). Almost all the research underlying the conclusions in the earlier sections of this essay has used one of these two techniques.

A third EEG procedure involves the presentation of brief, repeated auditory, visual, or tactile stimuli to elicit acute changes in electrical brain activity called event-related evoked potentials. Evoked potentials can be seen when the spontaneous EEG, used as data in the first two methods, is mathematically removed and only short-term changes in EEG signals remain. The precise shape of evoked potentials differs in relationship to a variety of factors including the duration, intensity, predictability, meaning, and sensory modality of the stimulus. However, when such factors are controlled, one expects a sequence of alternating positive and negative potentials during the second or so after stimulus presentation.

Several studies seemingly indicate that hypnotic suggestions to hallucinate alter evoked potentials among highly hypnotizable participants in ways consistent with their reported hallucinatory experiences. In these studies, highly hypnotizable participants who claim to experience either a particular hallucination or amnesia are compared to less hypnotizable counterparts who do not report similar experiences. Across visual and auditory stimuli and across experiments, highly hypnotizable participants seem to show augmented or faster positive evoked potentials and/or decreased or delayed negative potentials when asked to hallucinate the presence of stimuli that are not there. Again consistent with suggested hallucinations, when asked to hallucinate the absence of stimuli that are present, highly hypnotizable subjects show decreased or delayed positive evoked potentials, and/or augmented or faster negative evoked potentials compared to their less hypnotizable counterparts. Parallel findings to those found in negative hallucinations have emerged after amnesia instructions. In sum, highly hypnotizable participants show brain activity consistent with their reports of hypnotic hallucinations (see Crawford 2001).

Perhaps because it is the most clinically relevant negative hallucination, the effects of hypnotic analgesia instructions, instructions not to experience pain when exposed to a painful stimulus, have been studied with a variety of measures of brain activity. These have included fMRI, cerebral blood flow using 133-xenon, SPECT, and integrated EEG. Such studies confirm the pattern seen in the evoked potential studies: instructions to experience analgesia leads to measurable changes in brain function indicating central amelioration of the painful stimuli among highly hypnotizable participants, but not in lows or moderates (see Crawford 2001, Karlin et al. 1980). It is also noteworthy that other studies indicate that hypnotic analgesia is not mediated by endogenous opiate systems (Hilgard and Hilgard 1975).

As to immune system and skin disorders, the positive data remain largely anecdotal. There is a long history of treating skin disorders with hypnosis. (For example, hypnosis is still treatment of choice for warts that do not respond to conventional treatment.)

Similarly, asthma may be treated with hypnosis (although there is some evidence that hypnotic effects are limited to relaxation of the upper airways, rather than changes at the level of the lung where oxygen exchange occurs). Finally, there have been a number of studies on the ability of suggestion and imagery to alter immune responding directly. However, findings have been mixed and no clear conclusions can yet be drawn other than that large, robust effects seem unlikely (see Rhue et al. 1993).

5. Conclusions

Hypnotic responses remain intriguing, somewhat spectacular effects produced by simple verbal suggestion. Since the 1950s, careful rigorous work has been done to elucidate underlying mechanisms and any psychophysiologic concomitants of these phenomena. Inducing hypnosis *per se* has no clear physiologic effects. However, when compared to their less hypnotizable counterparts, highly hypnotizable subjects show slightly higher levels of theta activity during rest and greater task consistent asymmetry on strongly lateralized cognitive tasks. This is consistent with the notion that highs are capable of more flexible attention than are their less hypnotizable counterparts. Moreover, when successfully engaged in positive and negative hallucinations, including hypnotic analgesia and amnesia, highly hypnotizable subjects show cortical activity consistent with their reported subjective experiences across a number of measures of brain function.

See also: Electroencephalography: Basic Principles and Applications; Electroencephalography: Clinical Applications; Event-related/Evoked Potentials; Hallucinations, Neural Basis of; Hypnosis, Psychiatry of; Neglect, Neural Basis of; Pain, Management of; Sleep: Neural Systems

Bibliography

- Banyai E, Hilgard E 1976 A comparison of active-alert hypnotic induction with traditional relaxation induction 1976. *Journal of Abnormal Psychology* **85**: 218–24
- Chaves J, Perry C, Frankel F (eds.) 1998 Special Issue: Hypnosis in the relief of pain: Parts 1 & 2. *International Journal of Clinical and Experimental Hypnosis* **45**(4): 354–3355 & **46**(1): 4–5
- Crawford H 2001 Neuropsychophysiology of hypnosis: Toward an understanding of how hypnotic interventions work. In: Burrows G, Stanley R, Bloom P (eds.), *Handbook of Clinical Hypnosis*. Wiley, New York
- Crawford H, Gruzeliér J 1992 A midstream view of the neuropsychophysiology of hypnosis: Recent research and future direction. In: Fromm E, Nash M (eds.) *Contemporary Hypnosis Research*. Guilford, New York, pp. 227–66
- Ellenberger H 1970 *The Discovery of the Unconscious: The History and Evolution of Dynamic Psychiatry*. Basic Books, New York
- Graffin N, Ray W, Lundy R 1995 EEG concomitants of hypnosis and hypnotic susceptibility. *Journal of Abnormal Psychology* **104**: 123–31
- Hilgard E, Hilgard J 1975 *Hypnosis in the Relief of Pain*. Kaufman, Los Altos, CA
- Karlin R, Morgan D, Goldstein L 1980 Hypnotic analgesia: A preliminary investigation of quantitated hemispheric electroencephalographic and attentional correlates. *Journal of Abnormal Psychology* **89**: 591–94
- Kihlstrom J 1985 Hypnosis. *Annual Review of Psychology* **36**: 385–418
- Kirsch I, Lynn S 1995 The altered state of hypnosis: Changes in the theoretical landscape. *American Psychologist* **50**: 846–58
- LaBriola F, Karlin R, Goldstein L 1987 EEG laterality changes from prehypnotic to hypnotic periods. *Archives of Biological Psychiatry* **16**: 1–5
- Orne M 1959 The nature of hypnosis: Artifact and essence. *Journal of Abnormal and Social Psychology* **58**: 277–99
- Orne M 1970 Hypnosis, motivation, and the ecological validity of the psychological experiment. In: Arnold W, Page M (eds.) *Nebraska Symposium on Motivation*. U. of Nebraska Press, Lincoln, NE, 187–265
- Orne M, Hammer A 1974 Hypnosis. *Encyclopedia Britannica*, 15th edn. Helen Hemingway Benton, Chicago, pp. 133–40
- Perlini A, Spanos N 1991 EEG, alpha methodologies and hypnotizability: A critical review. *Psychophysiology* **28**: 511–30
- Piccione C, Hilgard E, Zimbardo P 1989 On the degree of stability of measured hypnotizability over a 25-year period. *Journal of Personality & Social Psychology* **56**: 289–95
- Rhue J, Lynn S, Kirsch I 1993 *Handbook of Clinical Hypnosis*. American Psychological Association, Washington, DC
- Sabourin M, Cutcomb S, Crawford H, Pribram K 1990 EEG correlates of hypnotic susceptibility and hypnotic trance: Spectral analysis and coherence. *International Journal of Psychophysiology* **10**: 125–42
- Sheehan P, Perry C 1976 *Methodologies of Hypnosis*. Erlbaum, Hillsdale, NJ
- Shor R, Orne E 1962 *The Harvard Group Scale of Hypnotic Susceptibility: Form A*. Consulting Psychologists Press, Palo Alto, CA
- Weitzenhoffer A 1959 *The Stanford Hypnotic Susceptibility Scale for use in research investigations in the field of hypnotic phenomena Forms A & B*. Consulting Psychologists Press, Palo Alto, CA
- Weitzenhoffer A, Hilgard E 1963 *The Stanford Profile Scales of Hypnotic Susceptibility Scales: Forms I and II*. Consulting Psychologists Press, Palo Alto, CA

R. A. Karlin and E. Orne

Hypnosis, Psychiatry of

1. Definition

This treatment technique is thought to have received its name from the English surgeon James Braid (1795–1860) whose clinical observations likened the physiological condition of those who were responsive to the intervention to the state of sleep. It is the direct descendent of the seventeenth and eighteenth century

practices then called 'Magnetism,' 'Somnambulism' and 'Mesmerism,' although its roots can be traced to the religious and healing arts of antiquity. Today practitioners of hypnosis would subscribe to the American Psychological Association's (1993) definition of hypnosis as a 'procedure wherein changes in sensations, perceptions, thoughts, feelings or behaviors are suggested.'

2. Relevant History

Chroniclers of hypnosis (see Ellenberger 1970, Gauld 1992) point to the work of Franz Anton Mesmer (1734–1815) as the cornerstone for the establishment of this technique in modern medicine. Mesmer's belief that an imbalance in magnetic fluid led to disease states in humans gave rise to the ritualized application of magnets to individuals with a variety of medical and psychological problems. Not only the metal apparatus, but also certain people were thought to possess the 'animal magnetism' to restore the magnetic equilibrium and health. French government and scientific investigators who tested his work were able to dispute Mesmer's disease theory, but not the results. People continued to recover from a variety of ailments by means of suggestions (i.e., ideas) provided by charismatic individuals, and the technique soon spread throughout Europe and the USA.

Sigmund Freud (1856–1939) made his living as a young psychiatrist by treating patients, almost exclusively, with hypnosis. As a student of the famous neurologist Jean Martin Charcot (1825–1893) and the physician Hippolyte Bernheim (1840–1919), he learned about the efficacy of an idea to facilitate an action or symptom relief. His early technique employed suggestions to deny his patients' complaints energetically and to provide assurance and motivation for the performance of more adaptive behaviors. Freud also observed hypnotized subjects seeming to possess knowledge without realization, carrying out post-hypnotic suggestions without awareness of the source, and blindly following the commands of the hypnotist. From this clinical data and as a translator of Charcot's and Bernheim's writings on suggestion, hysteria and hypnosis, he became convinced that there were powerful, unconscious mental forces extant in humans that influenced behavior.

In Charcot's Salpêtrière, Freud learned that traumatic events create trance states in hysterics, similar to hypnosis, wherein individuals are vulnerable to auto-suggestion, and painful emotions become pathologically trapped in the nervous system. Hysterics, then, suffered painful reminiscences of these traumata that became expressed cognitively, affectively, somatically and/or behaviorally. Since they possessed similar characteristics, hypnotic trance was employed to assist the individual to revisit the scene of the original trauma, re-experience the original affects for cathartic discharge and recover from their symptoms through

catharsis and insight. This technique was the basis for Freud's uncovering treatment in the *Studies on Hysteria* (Breuer and Freud 1893–1895) and hypnosis became one of the underpinnings of psychoanalytic theory and technique.

The power of suggestion, the influence of unconscious mental processes in humans and the use of hypnosis to access memories were not the only contributions of hypnosis to psychoanalysis. Freud was fascinated by the question of 'influence' and what led people to follow the directions of others. He reasoned that group activity and hypnotic treatments encourage people to minimize individual interests, personal control, discernment, personality style, etc., and to develop intense emotional relationships with people for reasons that are difficult to account for objectively. Freud saw the rapport created between doctor and patient in hypnosis to be comparable to that of tribal members for the chief and used it as a paradigm for his understanding of therapeutic influence and transference.

Where Mesmer might be credited with taking hypnosis away from the exorcists and placing it at least on the fringe of traditional medicine, Freud clearly explored hypnosis for what it could offer as a psychotherapeutic technique, and for what it might contribute to his model of the mind. It is clear from a close reading of his work that Freud continued throughout his career to use hypnosis as an archetype with which to compare and study his incipient psychoanalytic theory. However, his report that psychoanalysis began with his rejection of hypnosis relegated this technique to an inferior place among methods for mental health treatment. Very few of his followers would risk alloying the 'copper of suggestion' with the 'pure gold of psychoanalysis.' Freud's efforts to distance himself and psychoanalysis from hypnosis had something of the same dampening effect as Mesmer's failure to validate magnetic theory scientifically.

The birth of psychology and social science research techniques during the twentieth century brought new methods of inquiry into the essence and effects of hypnosis. The rise of interest in mind–body medicine also fueled clinical interest and investment in this area. Both of these developments, along with the creation of several professional journals and international societies for the study and practice of hypnosis, contributed to the continued development of this field.

3. Therapeutic Applications

3.1 Medicine

A technique that can alter sensations, perceptions, thoughts, feelings and/or behavior can be especially helpful in today's medical setting (Covino and Frankel 1999). Patients and practitioners have long experi-

enced anecdotal reports of symptom relief from something that a physician said, did or provided. Researchers in the laboratory and clinical setting have been laboring to establish empirical support for the use of hypnosis with a variety of symptomatic complaints and to manage common emotions in the medical setting to facilitate treatment.

Most clinicians have witnessed the influence of a patient's perception on the subjective experience of pain. Where an objective physiological event almost always serves as the cause of an injury, the meaning that the person makes of the event determines the amount of suffering. In experimental laboratories, hypnotized subjects have demonstrated an ability to tolerate electric shock, ischemic and cold pressor pain better than their nonhypnotized peers or when they employed other pain management strategies. Hypnotic suggestions for analgesia, distraction, and dissociation have been found to be superior to placebo, relaxation, and cognitive interventions in the management of a variety of painful conditions. Clinical research finds success for hypnosis in the management of pain associated with surgery, cancer, reflex sympathetic dystrophy, burns, interventional radiology, angioplasty, bone marrow transplant, and colonoscopy.

A number of experimenters have been successful at producing raucous breathing and decreased pulmonary function among a subset of asthma patients by means of suggestion alone. Whether these patients were invited to believe erroneously that they were encountering an allergen or breathing a bronchoconstricting drug, the idea alone was enough to affect their breathing adversely. Several well-designed studies have found that hypnosis training has a better impact on asthma symptoms, 'as needed' (p.r.n.) medication use, and pulmonary function tests than education, relaxation, guided imagery, and placebo inhalers. Furthermore, asthma patients who were trained to use hypnosis to manage their breathing have demonstrated a superior ability to tolerate methacholine challenge testing.

Allergists, dermatologists, and others have reported successful treatment of warts, eczema, psoriasis and other skin conditions with hypnosis. In clinical studies, suggestions to imagine bathing in salty water, increasing warm and tingling feelings on the skin, reducing blood flow to infected areas and general healing have assisted patients to relieve itchiness, reduce inflammation and decrease the number and size of warts. Likewise, patients with severe burns have been able to refrain from scratching, tolerate dressing changes and reduce their use of pain medications with the help of hypnosis.

Many patients with irritable bowel syndrome experience emotional distress in the form of anxiety and anger that adversely impacts gastric secretion, motility, and pain. A considerable number of these patients find relief from placebo interventions and

several studies report alterations in gut physiology with hypnosis. Gastric acid secretion, gut motility, bowel habits, abdominal distention and abdominal pain have been reported to improve with very brief hypnotic interventions aimed at increasing relaxation, controlling smooth muscle and gut activity and bowel function.

Hypnotic training has been found to diminish the severity and frequency of bleeding episodes among patients with hemophilia. Training a small sample of patients to use hypnosis for relaxation and anxiety reduction has controlled factor usage, spontaneous bleeding and the need for transfusion. In addition, preoperative training in self-hypnosis for the control of anxiety and bleeding has been found to reduce the number of blood transfusions required by surgical patients.

One of the most common causes of noncompliance with chemotherapy among cancer patients is the significant distress that they experience from the associated nausea and vomiting. Anti-emetic drugs often have their own side effects and are typically ineffective with the anticipatory emesis that is often a conditioned response. General anesthesia, also, commonly causes postoperative nausea and vomiting among surgical patients. Well-designed clinical interventions have found support for hypnosis training over relaxation techniques, distraction or individual counseling in the management of nausea and the frequency and intensity of emesis. This has been the case for pediatric as well as adult cancer patients.

Although a number of factors including maternal age, number of previous deliveries and medication use can influence the duration of labor, women trained in hypnosis have experienced speedier and more satisfactory deliveries. Suggestions for relaxation, analgesia and comfort have been successful in influencing analgesic medication use and pain ratings as well as the length of time spent in the first and second stages of labor. Hypnosis trained primigravid women performed more like parous mothers than first time mothers without hypnosis on the above variables.

Numerous methodological difficulties limit confidence and widespread application of hypnotic techniques in medicine. As with most clinical research, the heterogeneity of the patients in terms of illness onset, type, course, previous treatments, co-morbidity etc., limits what can be said with confidence regarding the use of hypnosis in medicine. Also, many experimenters fail to include an adequate description of hypnosis, or might include hypnotic interventions as a component of a more complex psychosocial interventions, or fail to ascertain whether the subject was indeed hypnotized during the test. Nonetheless, qualified empirical evidence exists to support the efficacy of hypnotic techniques in the treatment of pain, asthma, dermatological disorders, irritable bowel syndrome, hemophilia, hyperemesis and obstetrics (Pinnell and Covino 2000).

3.2 Psychotherapy

The addition of hypnosis to cognitive-behavioral treatments for anxiety and habit disorders (smoking, weight control) improves the effect of traditional psychotherapeutic interventions. The ability to alter perception, adjust expectations, increase motivation and imagine more adaptive behaviors facilitates change, especially among those suffering from anxiety disorders.

Where it was once thought that autonomic nervous system hyperarousal was the result of so called stressful life events, recent research highlights the importance of cognitive processes such as perception. One person's strain might well be another's relief, depending upon the meaning that is made of the situation. An invitation from the hypnotist to perceive an event in a more benign fashion creates the opportunity for emotional relief and more adaptive behavior. Imaginative individuals can, also, use their capacity to picture frightening situations vividly to expose themselves to feared stimuli successfully. These patients are able to achieve some of the same mastery as '*in vivo* exposure' without the time and travel. Behaviors can be mentally rehearsed, problems anticipated and resolved with the help of suggestion, focused concentration and imagery. Likewise, unique motivational ideas can be the object of attention without distraction or criticism, and used to provide significant force for behavioral change.

Most efforts to use hypnosis for smoking cessation and weight control have seen modest results. Patients often report achieving relaxation and satisfaction with the technique, but pounds lost have been few and smoking abstinence is rarely achieved by more than 20 percent of program participants. This might well be due to the manner in which suggestions are employed to motivate patients and reduce appetite and desire. Anxiety symptoms related to surgeries and phobias have been especially responsive to hypnotic training, especially when combined with other psychological techniques. Much more is known these days about the complex manner in which patients cycle through processes of habit change as well as the benefits of medication support. Further work in this area would do well to integrate differing hypnotic approaches into the diverse phases and processes that people use in making habit change happen.

A number of clinicians have advocated the use of hypnosis for patients in psychotherapy. The hope in this area is to assist people to recover memories of difficult experiences as well as to teach a coping mechanism for managing extreme anxiety and dissociative symptoms. These therapists propose that the absorption, dissociation and intensity of emotions that often accompanies hypnotic trance are similar to the experiences of many victims of trauma. In a manner similar to state-dependent learning, and in much the same way that Charcot and Freud employed

this technique, patients are invited to experience hypnosis, revisit the initial trauma and have an opportunity to express, integrate and control their associated emotions. In addition, those who spontaneously experience dissociative states are taught to enter into a hypnotic trance to reduce anxiety and to achieve some ability to terminate or control hyperarousal.

In a fashion that mirrors Freud's later work, the use of hypnosis as a psychotherapeutic technique is the focus of considerable controversy at the moment. In general, as was the case with Freud, the main idea of memory recovery sits in the context of credibility and veridicality. For some patients and practitioners it is painful and angering to raise the issue of truth in memory, since so many of those who were physically and sexually abused in their youth were frightened into secrecy. Yet, Freud seemed to see the inevitable intertwining of suggestion with dissociation in hypnosis as limiting the utility of this technique in psychotherapy.

Laboratory research finds that some modest gains in memory can be achieved from efforts to recall, especially when these are made proximate to the experience. Cues, reviews, relaxation, and other motivational techniques also seem to aid the recovery of memory. Nonetheless, hypnotic techniques have generally failed to improve recall of word lists and eyewitness reports. On the other hand, considerable evidence is mounting to suggest that people in clinical and experimental situations can be lead to the erroneous belief that an event or experience did indeed occur. The fact that these circumstances are more likely to take place in situations of heightened emotionality, uncertainty and unfamiliarity makes many clinicians cautious about using hypnosis as an 'uncovering' technique.

4. Essential Elements and Putative Mechanisms

At the end of such a review, it would be nice to be able to say clearly how and why hypnosis is and has been so helpful to patients in a variety of circumstances. Despite its age as a psychological intervention, this is not yet possible. A good deal of research has, thus far, failed to identify a specific neurological marker or neurophysiological substrate for hypnosis. There are, however, several ways that the beneficial effects of hypnosis have been understood.

Suggestions are ideas presented to a subject for uncritical acceptance. In hypnosis, the subject is asked to concentrate on these ideas in a manner that diminishes distraction, lessens logical analysis and maximizes acceptance and acquiesce. Suggestions are not commands, but invitations that can be rejected or accepted. However, like the response of patients to placebo medications, there is often a sense of

involuntariness to hypnotic responses and an attribution of causation to the hypnotist. The physical context, psychological expectations or the pressure to fit in well with a group can also provide direction for a person's thoughts, feelings or behaviors. Some theorists maintain that these psychosocial factors alone or in combination result in successful hypnotic response.

The ability to become absorbed in an idea as if it is real and to use fantasy and imagery to alter cognition and/or behavior is also seen as a determinant of successful hypnotic response. Such special concentration is thought to invite a state of dissociation where consciousness is divided. The ability to separate awareness of sensation, memory, time, behavior etc., is not only part of daily experience (e.g., shifting the gears of a car without awareness, daydreaming), but it can be created by hypnotists and imaginative patients. Such an ability is thought to facilitate the use of hypnosis in management of painful conditions and anxiety. Since the higher cognitive faculties are blocked from elaborating pain or anxiety signals, emotional or behavioral responses are muted or stopped.

There are a number of clinical, and perhaps non-specific, factors that facilitate or hinder the response of patients to hypnosis. Patients must be motivated to follow the suggestions of the hypnotist. A fearful, overly defensive, unwilling person will rarely succeed with hypnosis and one who feels safe, valued, and cared for by an 'expert' most likely will. Research indicates that hypnotizability appears to be a rather normally distributed, stable, trait. Although there are some that claim that anyone can achieve some benefit from hypnotic interventions, many hypnosis experts find that those with heightened ability are the most helped. Certain conditions favor suggestion such as ambiguity, uncertainty, novelty and hyperemotionality. Good hypnotic technique endeavors to recruit or develop some of these elements.

See also: Conscious and Unconscious Processes in Cognition; Freud, Sigmund (1856–1939); Hypnosis, Neural Basis of; Hysteria; Psychoanalysis, History of; Psychological Treatment, Effectiveness of

Bibliography

- Breuer J, Freud S 1893–1895 Studies on hysteria. In: Strachey J (ed.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Hogarth Press, London, Vol. 2, pp. 1–312
- Covino N A, Frankel F H 1999 Clinical hypnosis. In: Nicoli A M Jr. (ed.) *The Harvard Guide to Psychiatry*. The Harvard University Press, Cambridge, MA, pp. 552–6
- Ellenberger H F 1970 *The Discovery of the Unconscious: The History and Evolution of Dynamic Psychiatry*. Basic Books, New York
- Gauld A 1992 *A History of Hypnotism*. Cambridge University Press, New York

- Pinnell C M, Covino N A 2000 Empirical findings on the use of hypnosis in medicine: A critical review. *The International Journal of Clinical and Experimental Hypnosis* **48**: 170–94

N. A. Covino

Hypothalamic–Pituitary–Adrenal Axis, Psychobiology of

This article will examine the role of psychological factors on the regulation of the hypothalamic–pituitary–adrenal (HPA) axis. The HPA axis regulates the primary endocrine responses to stress, which in turn regulate major metabolic and neural processes. The major principle that will be elaborated is that both activation and inhibition of the HPA axis are in part modulated by psychological factors. We will examine first specific interventions that occur during development that permanently alters the responses to stress. Second, we will discuss some of the mechanisms whereby responses to stress can be modulated behaviorally.

1. The HPA Axis

The secretion of glucocorticoids (GC) by the adrenal cortex under conditions of stress is an important mammalian mechanism of adaptation. The response to stress involves complex interactions between the neurosecretory hypothalamic neurons, the corticotrophs of the anterior pituitary, and the adrenal glands. Various neural afferents mediating the stress response converge at the parvocellular neurons of the hypothalamic parvocellular nucleus (PVN). The cell bodies of the PVN neurons are one of the primary sites of the synthesis of corticotropin-releasing factor (CRF) and project to the median eminence. Activation of the PVN neurons by a variety of different stimuli induces release of CRF into the portal capillaries through which CRF is transported to the anterior pituitary.

The binding of CRF to receptors in the anterior pituitary in turn induces the release of adrenocorticotropin (ACTH). Subsequently, ACTH in the general circulation stimulates the secretion of GC, primarily corticosterone in the rodent and cortisol in primates and humans. GC secreted by the adrenal serve a variety of regulatory and permissive functions which recruit the organism's endocrine, metabolic, immune, and behavioral responses to stress (Munck et al. 1984). Further circulating GCs participate in their own regulation by exerting a negative feedback on stress-induced activity. There are numerous negative feedback sites, which include the PVN, anterior pituitary, and possibly the hippocampus.

The HPA axis is highly responsive to changes in both the interior and exterior environment. One of the more interesting developments in this field is that there appear to be different neural pathways that respond to different stimuli that activate the PVN neurons. Currently, five different pathways have been described (Sawchenko 1991). Certain pathways are activated when visceral and interoceptive signals are involved in the activation of the HPA axis. One of the major pathways involves the structures of the principle components of the limbic region of the telencephalon, which includes the hippocampal formation, the septal, and the amygdaloid nuclear complexes.

Not only have the limbic structure been implicated as possible sites for GC feedback but they are also implicated in the processing of emotional responses (LeDoux et al. 1988). The amygdala plays a pivotal role in the 'fear circuit' described by Davis (1992). The hippocampus also appears to play an important role in the processing of affective information. Given the importance of affective stimuli in the activation of the HPA axis it is not surprising that there should exist neural pathways responsible for processing this class of stimuli.

2. Early Experience

After more than four decades of research using a variety of animal species it has been well established that the effects of early experience are diverse and pervasive. One early experience manipulation, early handling, modifies the HPA axis so as to permanently alter the endocrine (ACTH and GC) response to stress and to modify the neural components of the HPA axis that are involved in regulating the stress response. The early handling paradigm involves an apparently innocuous manipulation in which the experimenter removes the rat pup from its mother once daily, placing it either alone or as a group in a novel container for a few minutes and then reuniting it with the dam in the home cage. Early handling is most effective if initiated between birth and 14 days postnatally.

The early studies demonstrated that early handling profoundly altered the GC response to several different challenges. In general, as adults, the magnitude of the GC response was reduced and further, the GC response in the early handled rats returned to basal levels more rapidly indicating a more efficient negative feedback mechanism (Batnagar et al. 1996). The initial studies were limited by the technologies available at that time. With technological advances it became possible to study other components of the HPA axis. Thus, early handled animals showed a reduction in ACTH secretion following stress when compared to non-handled rats. Early handled rats also have significantly lower basal levels of CRH gene expression (Plotsky and Meaney 1993).

Another well-documented effect of early handling is to permanently modify the GC receptors in the hippocampus. In the central nervous system (CNS), two types of GC receptors are found and each functions to regulate different aspects of HPA activity (DeKloet 1991). The mineralocorticoids (MR) are involved in the tonic regulation of the HPA axis, that is, circadian rhythms. In contrast, the glucocorticoid receptor (GR) appears to be an important extrahypothalamic site of an inhibitory feedback effect on HPA activity although it is not clear that the hippocampus is a negative feedback site for GC. Occupancy of available GRs seems to be required to restrain stress induced ACTH secretion.

In an extensive series of studies Meaney and co-workers (see Batnagar et al. 1996) have established that early handled animals have an increase in hippocampal GR sites as well as increased GR mRNA. Early handling has little effect on the MRs. It has been postulated that the increase in GR density increases the sensitivity of the hippocampus to circulating GCs, enhancing the efficiency of the negative feedback inhibition of HPA activity, thus permitting early handled animals to reduce the poststress levels of ACTH and GC.

All the current information on the effects of early handling on the HPA axis indicates that manipulation permanently modifies several of the important components of neural circuitry that regulate the HPA axis. Thus, the effects of early handling can be attributed largely to alterations in the CNS mechanisms that regulate ACTH and GC secretion.

One hypothesis proposed to account for the effects of early handling is that handling affects the pup indirectly by producing changes in the behavior of the dam (the maternal mediation hypothesis). Maternal behavior parallels and compliments the need of the developing organism and the parent–offspring relationship is most intense during the preweaning period when the dam's role in regulating vital neonatal functions is fundamental. Disturbances of mother–infant relationships by brief periods of maternal separation alters the pup's behavior in a manner that would elicit increased pup-directed behavior by the dam, such as licking, grooming, retrieving, etc. This increase in pup-directed behavior could bring about pervasive changes in regulation of the HPA axis. Using a more naturalistic approach (Caldji et al. 1998) demonstrated that pups who received high levels of licking and arched-back nursing from their dams showed the characteristic pattern of HPA activity that closely resembled the early handled pups.

Plotsky and Meaney (1993) have used relatively long postnatal maternal separations as a model of adverse early experience. Thus, pups were separated from the dams for 3 hours daily beginning on day 2 and continuing through day 14. When compared to pups that had experienced short periods on maternal separation GC levels were significantly higher fol-

lowing restraint and further, basal levels of CRH gene expression in the PVN were elevated. Other investigators have used more extended periods of separation, that is, 8 hours every other day, and have also found that HPA activity is altered in the direction of a more exaggerated response.

Thus, it appears as though the HPA axis can be modified bidirectionally depending upon the length of the separation. It is possible that the signals emitted by the pup are different following brief vs. a more prolonged separation and that the effects on maternal behavior are manifested in very different ways resulting in seemingly paradoxical outcomes. The data do clearly indicate that early experiences do result in a permanent and apparently irreversible influence on the adult HPA axis regardless of the direction of these changes.

3. Coping

Based primarily on the initial observations of Selye (1956), it became clear that the endocrine system and more specifically increased secretion of the so-called stress hormones ACTH and GC was a critical component of the response to stress. Selye also described these responses as nonspecific. Thus, regardless of the stimulus that provoked these responses they were general and occurred nonspecifically to any or all challenges.

Mason (1968) presented a point of view which differed radically from the widely-held beliefs of the time. Mason emphasized a crucial role for 'the psychological approaches involved in the emotional or arousal reactions to threatening or unpleasant factors in the life situation as a whole.' He argued that the nonspecific response to stress is primarily behavioral or psychological in nature. However, although Mason recognized the importance of psychological factors associated with the endocrine activation following stress he did not elaborate on what some of these dimensions could be.

In an attempt to further specify the psychological aspects of the endocrine stress response Levine and Ursin (1990) invoked a cognitive approach in which the lack of information (uncertainty) resulted in the activation of the HPA axis. Uncertainty could be a function of the absence of information and/or the failure to fulfill an expected outcome or to avoid a negative one. The absence or loss of control and the lack of predictability also creates uncertainty which when processed by the brain activates the neurobiological events that result in increased HPA activation.

3.1 Control

One of the most critical psychological dimensions that contribute to the regulation of the HPA axis is control. In a remarkable series of experiments Weiss (1972)

convincingly demonstrated the importance of control in reducing the physiological consequences of exposure to stress. Using gastric ulceration as the measure of stress it was shown that rats that could make a response that terminated an electric shock showed a marked reduction in the number of gastric ulcers when compared to a 'yoked' group that received identical shocks but were not permitted to respond. This initial experiment clearly demonstrated the importance of control. In further studies two other psychological operations also ameliorated the extent of gastric ulceration. These were predictability and feedback. These three elements, control, predictability, and feedback, all serve to reduce arousal and thus HPA activity and are postulated to contribute to the process of coping.

There are many examples of the importance of control in modifying the response of the HPA axis. Rats permitted to escape from shock showed a reduction in GC levels although they received shock on every trial (Davis 1977). In contrast repeated exposure to inescapable shock resulted in increased secretion of GC (sensitization). Hanson et al. (1976) exposed Rhesus monkeys to a loud noxious noise. One group of monkeys was allowed to control the duration of the noxious stimulus by making a lever press response to terminate the noise. A comparable group of monkeys were exposed to the identical duration of the noise, but were not permitted to regulate the duration. The monkeys that exercised control showed levels of GC that were similar to nonnoise exposed monkeys. In contrast, the 'yoked' control animals exhibited extremely high levels of GC. In this experiment the monkeys that could control the duration of the noise were re-exposed to the noise on a subsequent day but on this occasion they were not permitted to exercise control. Their GC levels were now elevated and exceeded those of the noise alone group. This finding represents an example of yet another psychological dimension involved in the regulation of HPA activity, the failure of expectancies.

3.2 Expectancies

Physiological activation in response to uncertainty is a central component of the model presented by Sokolov (1963). Although this model was an attempt to account for the neural mechanisms involved in the process of habituation, a broader interpretation of this model could also account for why uncertainty would result in the activation of the HPA axis. When a subject is presented with a stimulus that not been previously experienced or is unexpected an orienting response occurs. Both behavior and physiological responses are elicited by the stimulus. Amongst the number of physiological components of the orienting response is HPA activation (Hennessy and Levine 1979).

If the stimulus is repeated frequently the response tends to diminish and eventually disappears. Sokolov's

model is based on a matching system in which new stimuli or situations are compared with a representation in the CNS of prior events. This matching process results in the development of expectancies. Uncertainty can, therefore, be viewed as a violation of expectancies. From this perspective it is the evaluation of the event or events by the individual that is crucial in determining whether that event will elicit a stress response resulting in HPA activation.

(Bolles 1972) identified two distinct types of expectancies, stimulus and response outcome expectancies. Stimulus expectancies refers to the storage of information that one stimulus precedes another (classical conditioning). Response outcome expectancies results from the learning that a particular response results in a positive outcome, that is, obtaining reinforcement or avoiding a negative outcome. The development of positive outcome responses is dependent on some of the coping mechanisms previously described. Thus, if an organism can exercise control by making an active response that leads to a predictable outcome this information will be stored in the CNS as a positive outcome expectancy.

Thus, when rats can obtain reinforcement on a highly predictable schedule removal of the reinforcement (extinction) produces a marked elevation in GC. Further when rats are shifted from a highly predictable schedule to an unpredictable schedule of reinforcement GC are again elevated. In contrast when the shift in reinforcement schedules is from low predictability to high predictability GC levels significantly decrease. More recently a similar outcome has been reported in primates. Squirrel monkeys were required to obtain food in a laboratory model of foraging. When the foraging demand was high and food was available unpredictably cortisol levels were persistently elevated. When these monkeys were shifted to a low demand foraging schedule cortisol levels decreased rapidly. The effects of violations of expectancies are also observed in aversive situations. Rats that successfully learned to avoid shock show a progressive decrease in GC as their performance reached asymptote. However, if the animals are prevented from making the avoidance response, although shock is not administered, GC levels are once again elevated.

3.3 Social Factors

There is a large literature that has documented the importance of social behavior and social relationships on the HPA axis. At first glance these may appear to be a separate class of events. They can be, however, subsumed within the more general theories of stress and coping. An extensive series of studies using nonhuman primates (Levine and Coe 1988) demonstrated that at almost every point in development social variables can modulate the HPA axis, both in terms of increased activity and the inhibition of the response to environmental perturbations. Disruptions

of social relationships invariably result in increased HPA activity. Conversely, the presence of familiar social partners can reduce, and at times eliminate, the normal increase in HPA activity in squirrel monkeys. One of the most robust increases in GC levels is observed when infant squirrel monkey are separated from their mothers and other members of their natal social groups. These responses occur even when the infants are separated repeatedly and show no signs of habituation. If the infant is separated from the mother and permitted to remain with their familiar social partners the response is markedly reduced. This occurs only when mother infant dyads are group housed. When the dyad is housed individually the magnitude of the GC response is similar regardless of whether the infant or the mother was removed from the home cage. This phenomenon is often called social buffering, and has been demonstrated in the adult monkeys. Thus, when adult squirrel monkeys were exposed to a natural predator (a live Boa Constrictor) placed in a wire mesh enclosure above their cage, individually housed animals show a striking elevation of plasma cortisol, whereas if the monkeys are exposed to the snake when group housed cortisol levels do not differ from basal levels. Other experiments using different paradigms have demonstrated that the presence of social partners ameliorates the HPA response to stress.

These findings are consistent with the proposed effects of control and feedback on the HPA axis. For the infant, the mother is the primary source of emotional security and serves to reduce the infant's arousal levels. Thus, the infant's capacity to achieve contact and maintain proximity to the mother is perhaps the infant's earliest experience with the use of control. For the adult, the availability of stable and familiar social relationships provides a set of predictable outcomes due, in part, to a prior history of social interactions. Therefore, the predictability of these social interactions would serve to reduce uncertainty.

4. Summary

The contemporary view of stress largely is characterized as an imposition model. The voluminous number of studies in which severely traumatic stimuli were used to elicit HPA activation supports this view. However, there are numerous examples of HPA activation being elicited by the omission of a predictable outcome (frustration) or the absence of a salient feature in the environment such as maternal loss. In both classes of events the common element appears to be the lack of information (uncertainty) related to obtaining positive outcomes or the avoidance of negative outcomes.

The view presented in this article focuses on the capacity of the brain to process information pertaining to obtaining of positive outcomes and/or the avoidance

of negative outcomes. Stimuli that induce HPA activation do so as a consequence of the lack of information (uncertainty) concerning the obtaining of positive outcomes or the avoidance of negative outcomes.

See also: Anxiety and Fear, Neural Basis of; Motivation, Neural Basis of; Neuroendocrinology; Post-traumatic Stress Disorder; Psychobiology of Stress and Early Child Development; Sexual Behavior and Maternal Functions, Neurobiology of; Stress and Cardiac Response; Stress and Coping Theories; Stress: Measurement by Self-report and Interview; Stress, Neural Basis of; Stress: Psychological Perspectives

Bibliography

- Batnagar S, Shanks N, Plotsky P M, Meaney M J 1996 Hypothalamic-pituitary-adrenal response in neonatally handled and nonhandled rats: Differences in facilitatory and inhibitory neural pathways. In: McCarty R, Aguilera G, Sabba E, Kventnansky R (eds.) *Stress: Molecular Genetic and Neurobiological Advances*. Gordon and Breach, New York
- Bolles R C 1972 Reinforcement, expectancy and learning. *Psychological Review* **79**: 394–409
- Caldji C, Tannenbaum B, Sharma S, Francis D, Plotsky P M, Meaney M J 1998 Maternal care during infancy regulates the development of neural systems mediating the expression of fearfulness in the rat. *Proceedings National Academy of Science* **95**: 5335–40
- Davis M 1992 The role of the amygdala in fear and anxiety. *Annual Review of Neuroscience* **15**: 353–75
- DeKloet E R 1991 Brain corticosteroid receptor balance and homeostatic control. *Frontiers in Neuroendocrinology* **13**(2): 95–164
- Hennessy J W, Levine S 1979 Stress, arousal and the pituitary-adrenal system: A psychoendocrine model. In: Sprague J M, Epstein A N (eds.) *Progress in Psychobiology and Physiological Psychology*. Academic, New York
- LeDoux J 1995 Emotions: Clues from the brain. *Annual Review of Psychology* **46**: 209–35
- Levine S, Coe C L 1988 Psychosocial modulation of neuroendocrine activity. In: Pancheri P, Zichella A (eds.) *Biorhythms and Stress in the Physiopathology of Reproduction*. Hemisphere, New York
- Levine S, Ursin H 1991 What is stress? In: Brown M R, Koob G F, Rivier C (eds.) *Stress: Neurobiology and Neuroendocrinology*. Marcel Decker, New York
- Mason J W 1968 A review of psychoendocrine research on the pituitary-adrenocortical system. *Psychosomatic Research* **30**: 576–607
- Munck A, Guyre P M, Holbrook N J 1984 Physiological functions of glucocorticoids on stress and their relation to pharmacological actions. *Endocrinological Reviews* **5**: 25–44
- Plotsky P M, Meaney M J 1993 Early, postnatal experiences alters hypothalamic corticotropin-releasing factor (CRF) mRNA, medium eminence CRF content and stress induced release in adult rats. *Molecular Brain Research* **18**: 195–200
- Sawchenko P E 1991 The final common path: Issues concerning the organization of central mechanisms controlling cortico-

trophin secretion. In: Brown M R, Koob G F, River C (eds.) *Stress: Neurobiology and Neuroendocrinology*. Marcel Dekker, New York

Selye H 1956 *The Stress of Life*. McGraw-Hill, New York

Sokolov E N 1963 *Perception and the Conditioned Reflex*. Pergamon, Oxford, UK

Weiss J M 1972 Influence of psychological factors on stress induced pathology. In: Porter R, Knight J (eds.) *Physiology and Emotions in Psychosomatic Illness*. Elsevier, Amsterdam

S. Levine

Hypothalamus

1. Introduction

The hypothalamus is a small brain structure (0.4 percent of total brain volume) primarily of gray matter providing important functions which can be summarized to maintain homeostasis. From an evolutionary perspective, the hypothalamus can be considered a very old component of the central nervous system.

Historically, the first observations concerning hypothalamic functions were described for the control of the pituitary gland since endocrinological abnormalities (e.g., due to tumors) were more obvious than disturbances of autonomic functions, which are also under hypothalamic control. In 1930 Ernst and Berta Scharrer developed the concept of 'neurosecretory' hypothalamic cells derived from their morphological structure, and this dramatically stimulated subsequent research. It became obvious that the hypothalamus combines characteristics of glands and neuronal tissue. However, the focus of interest remains in investigations of hypothalamic control via the pituitary gland, which led to the discovery and chemical characterization of vasopressin and oxytocin in the early 1950s. The finding by Schally and Guillemin that the hypothalamus controls the secretion of other fundamentally important hormones such as TSH, FSH, ACTH, and LH was honored by the Nobel prize in 1977.

2. Neuroanatomical Localization and Structure

The hypothalamus is located in the diencephalon. It is laterally bounded by the subthalamus, medially by the third ventricle, rostrally by the lamina terminalis, caudally by the mesencephalon, and dorsally by the thalamus.

Although these boundaries are well defined, the anatomical structure of the hypothalamus is much more complex. From stimulation and lesion studies it is known that the hypothalamus consists of several

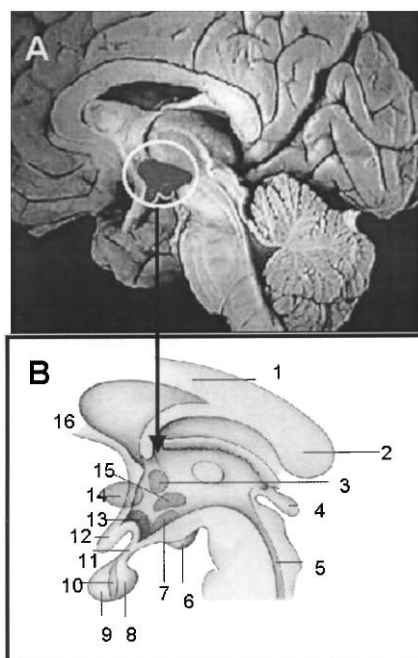


Figure 1

Localization of the hypothalamus in the brain (A) and hypothalamic nuclei (B). 1 Corpus callosum. 2 Splenium of the corpus callosum. 3 Paraventricular nuclei. 4 Pineal gland. 5 Aqueduct of mesencephalon. 6 Mammillary body. 7 Tuber cinereum. 8 Neurohypophysis. 9 Adenohypophysis. 10 Intermediate part. 11 Infundibulum. 12 Optic chiasma. 13 Supraoptic nucleus. 14 Preoptic nuclei. 15 Posterior hypothalamic nucleus. 16 Fornix.

neuronal clusters referred to as nuclei (see Fig. 1). Previous attempts at describing the anatomical relationships between these nuclei made use of a hexagon (see Fig. 2).

Based on the systematic topography of the hypothalamus given in Fig. 1, several areas can be distinguished. Most of the nuclei are located in the midline, medial, and lateral areas, which are themselves divided into supraoptic, tuberal, and mammillary sections. The preoptic nuclei are located in the preoptic region of the medial zone. The preoptic region is bounded by the septum in both directions. The anterior (supraoptic) region contains the anterior nucleus, the supraoptic nucleus, and the paraventricular nucleus. The intermediate (tuberal) region consists of the dorsomedial nucleus, the ventromedial nucleus, and the infundibular or arcuate nucleus. In the posterior region of the medial zone, the posterior nucleus and the mammillary nucleus can be identified. Finally, the lateral zone consists of the lateral hypothalamic nucleus. In the area of the periventricular zone many neurons have been identified as having

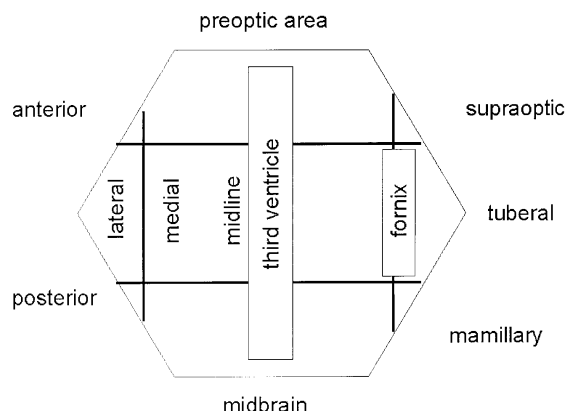


Figure 2

Schematic diagram showing the anatomical relationships between different nuclei of the hypothalamus as a hexagon (after Sladek and Sladek 1990)

fundamental importance for the secretion of releasing or inhibiting neurohormones. However, a clear distinction of the different nuclei has not been given for this area.

3. Functions

Although the hypothalamus can be considered a small brain structure, its functions are widespread and of fundamental importance. It is known that the hypothalamus is responsible for at least seven categories of functions. These are: the regulation of the anterior pituitary; the control of water balance; the regulation of autonomic responses; the regulation of feeding; the regulation of drinking; the regulation of reproductive functions; and the reinforcement and control of circadian functions. For an understanding of these functions, the characterization of hypothalamic inputs from other systems is essential (for a review, see Risold et al. 1997).

Retinal ganglion cell axons project through the retinohypothalamic tracts to the ventral regions of the suprachiasmatic nuclei (SCN). It is known that these nuclei couple the *biological rhythms* to the light/dark cycle. These rhythms can be divided into several categories depending on their frequency. Events repeated nearly every 24 hours are defined as circadian rhythms (e.g., changes in concentration of ACTH and cortisol). Periods with lower frequencies are considered circa-annual or infradian (e.g., season-dependent behavioral or metabolic changes or the menstrual cycle). Those with shorter frequencies (faster occurrence of periods) are defined as ultradian rhythms (e.g., respiration). The hypothalamus controls infradian and circadian rhythms. The important role of the SCN can be demonstrated by lesion studies in

which animals lost rhythmicity. The SCN projects to the paraventricular nucleus, which is connected to the spinal cord by efferent fibers. Noradrenergic fibers project to the pineal gland and affect the production and release of melatonin, concentrations of which markedly depend on the light/dark cycle. Moreover, recent studies have demonstrated that the SCN is involved in the regulation of autonomic responses (changes in heart rate) after brief stimulation by a flash of light.

Interestingly, the lateral hypothalamus seems to be a target for pheromonal stimuli derived from the main olfactory inputs. Moreover, projections from the medial amygdalar nucleus, as well as from the bed nuclei of the stria terminalis, provide accessory olfactory signals to some regions of the medial zone of the hypothalamus. This close connection to the vomeronasal organ is not fully understood with respect to its physiological function. However, two different pathways including amygdalar projections seem to be involved in the modulation of sexual and/or aggressive-defensive behavior.

Another source of 'information' is derived from viscerosensory and somatosensory signals. Most of these projections into the lateral hypothalamus are derived from the solitary tract and the parabrachial nuclei, which both receive visceral information from cranial nerves. However, there is evidence that the periventricular zone is an important target for projections providing 'visceral' signals derived from the brain stem involving catecholaminergic fibers. It is a likely assumption that these projections may provide information inducing the regulation of complex processes such as ingestive behavior. The lateral hypothalamus has been identified as a main target for spinal cord fibers which send information concerning sensations of touch or pain. The presence of fibers presenting auditory information has been demonstrated for the ventromedial nucleus of the hypothalamus. It has been suggested that these projections are important for regulation of lordosis and milk-ejection in lactating females. Somatosensory signals including nociceptive stimuli seem to reach the hypothalamus by activating neurons of the spinohypothalamic tract or involving the catecholamine system located in the brain stem. As a consequence, norepinephrine is released from the paraventricular nucleus (PVN) of the hypothalamus. In general, catecholaminergic projections to the PVN may play a pivotal role in stress responses by activating neurosecretory functions of the hypothalamus, as indicated by a reduced or blocked activation of the hypothalamus-pituitary-adrenal (HPA) axis after lesions of catecholaminergic fibers of the brain stem. There is agreement that the hypothalamus is one of the most important brain structures preparing the individual for appropriate reactions when faced with threat or other types of stress. One set of reactions concerns the cardiovascular system, which is controlled by the hypothalamus via the autonomic

nervous system, especially the vagus nerve and/or the initiation of hormone secretion. The fundamental role of the hypothalamus in orchestrating multiple influences on the pituitary will be discussed later.

Hypothalamic control of thermoregulation can be considered an excellent example for maintaining homeostasis. A center responsible for heat dissipation has been identified in the preoptic area and the anterior hypothalamus. The main physiological consequences of an activation of these neurons is suppression of shivering, panting, and vasodilation of cutaneous vessels with the common goal of reducing temperature. On the other hand, conservation of temperature seems to be mediated by the posterior hypothalamic region. Reactions to maintaining or increasing temperature are shivering, vasoconstriction, and suppression of panting. In his early work Bligh (1966) suggested the presence of two regulatory systems. One is mainly responsible for maintaining the normal variation of body temperature, whereas the other becomes activated as a consequence of more serious increases or decreases in body temperature. It is known that the neurons involved are sensitive to higher (37°C–40°C) and lower (32°C–36°C) than normal temperatures. Afferent pathways transport peripheral information (e.g., cooling or heating of the skin) to these neurons, initiating the regulatory responses. However, one has to realize that the hypothalamus is not the only structure involved in thermoregulatory responses. It is known that in 'spinal' animals (brain disconnected from spinal cord) some basic thermoregulatory responses are still intact, indicating the presence of lower centers for regulation. There is agreement that thermoregulation involves several regulatory circuits acting on different levels, with the hypothalamus the most sensitive one (Satinoff 1978).

Early lesion studies had revealed that the hypothalamus mediates ingestive behavior. Distinct areas responsible for feeding have been identified in the ventromedial nucleus as a kind of satiety center, and in the lateral hypothalamus as a feeding center. Lesion of the ventromedial nucleus results in hyperphagia and polydipsia, whereas stimulation leads to the opposite behavior. There is some evidence that the paraventricular nucleus affects the vagus nerve via its projections to the brain stem, and that the lesions result in a disturbed parasympathetic control of the liver and the pancreas. A closer look reveals that after lesion of the ventromedial nucleus, hyperphagia occurs with temporary dramatic weight gain. However, after a state of obesity has been reached, ingestive behavior becomes normalized and no further weight gain can be observed, although the hypothalamic function is still disturbed. This clearly demonstrates that further brain structures are involved in the regulation of ingestive behavior.

For the lateral hypothalamus, stimulation leads to hyperphagia, and lesions lead to aphagia. However, the mechanisms involved are much more complex

since animals with lesions of the lateral hypothalamus start to eat after an appropriate period of time. Again, other brain areas such as the amygdaloid nuclei, the frontal cortex, and the substantia nigra are involved in regulation of feeding. More recently it became clear that leptin produced by fat cells is important for the regulation of ingestive behavior. Leptin receptors have been identified in the lateral hypothalamus (Hakansson et al. 1998), the supraoptic arcuate, and the paraventricular nuclei. A low quantity of leptin can be considered a signal for increased fat intake. Peripheral (derived from the gut in response to food intake), as well as central (probably produced within the hypothalamus), cholecystokinin seems to play an important role as a satiety hormone. However, neuropeptide Y, known to stimulate appetite, is found in the arcuate nucleus, the medial preoptic area, and the paraventricular nuclei of the hypothalamus. Application of neurotransmitters or neuropeptides into the paraventricular nucleus leads to preferences for different food components. While norepinephrine is accompanied by intake of fat and carbohydrates, galanin stimulates the intake of fat. Leptin suppresses both stimulations of ingestive behavior induced by NPY and by galanin. Orexin has been discovered as a protein that is exclusively produced in the lateral hypothalamus. This neuropeptide markedly increases eating behavior after injection into the hypothalamus.

The most common reason for the induction of thirst is loss of water due to sweating, urination, or respiration. This osmotic thirst, rather than the thirst caused by hypovolemia, is regulated by the hypothalamus. The basis for this is the presence of osmoreceptors in hypothalamic cells. In fact, injection of salty fluid into the hypothalamus induces drinking. Neurons responding to changes in osmotic pressure have been identified in the preoptic area, the anterior hypothalamus, the supraoptic nucleus, and the circumventricular organ. As a consequence of stimulation, AVP (vasopressin, antidiuretic hormone) is released to stop further loss of water. Moreover, increased concentrations of angiotensin II and stimulation of baroreceptors are sensitive signals to the preoptic area resulting in drinking behavior.

Another important function of the hypothalamus is coordination and control of sexual behavior. For female behavior, it is known that estradiol, a sex steroid produced during the ovulatory cycle, specifically stimulates neurons of the ventromedial hypothalamus. These neurons exhibit receptors for estrogen and progesterone, and they are mainly responsible for inducing so-called lordosis behavior. This signals the receptivity of the female to the male population by several postural and physiological signs. It can be concluded that the ventromedial hypothalamus is important for controlling changes in steroid hormone concentrations and, furthermore, that it activates several pathways at the appropriate time during the ovulatory cycle to induce the lordosis response. It is

believed that estrogens can activate the hypothalamic receptors for progesterone. During pregnancy, when progesterone levels are very high, a down-regulation of these receptors appears with a decrease in sexual receptivity. The medial preoptic area of the hypothalamus is enriched with steroid-sensitive neurons, and evidence has been published that lesions in this area completely abolish male copulatory behavior in many species. The medial preoptic area is a target for pheromonal stimulation as well. Vomeronasal information reaches the amygdala, which sends signals to the medial preoptic area, indicating that this specific hypothalamic area plays a salient role in coordinating sensory and motor responses important for copulation. Both aspects are initiated by levels of testosterone which, interestingly, affect neuronal targets after aromatization into estrogen. The preoptic area of the hypothalamus is characterized by a sexual dimorphism, which may explain the differences in male vs. female sexual behavior. Taken together, the preoptic area of the hypothalamus is of fundamental importance for controlling sexual behavior, moreover, for sexual orientation as well. Interestingly, a clear-cut sexual dimorphic nucleus of the preoptic area could be identified as being larger in males than in females (Gorski et al. 1978). The reason for this dimorphism is brought about by testicular androgens (testosterone), which obviously influence the development of this specific nucleus within the preoptic area of the hypothalamus.

From animal studies, it is known that the medial preoptic area plays a key role in mediating maternal behavior (Numan 1978). Lesions in that area abolished nest-building and pup care in animals. Important pathways mediating these effects are the connections between the medial preoptic area and the ventral tegmental area. Disruptions of these connections lead to the same effects as described for lesions of the medial preoptic area. Estrogen receptors in this hypothalamic area are of major importance for regulating maternal behavior, as demonstrated by direct injections of estrogen or by suppression of this behavior after treatment with specific antagonists. Moreover, the stimulation of maternal behavior has been demonstrated for other hormones such as progesterone, oxytocin, and prolactin as well. However, more recent reports suggest the involvement of additional hypothalamic areas such as the ventromedial part for mediating maternal behavior.

A last behavioral component mediated in part by the hypothalamus is aggressive behavior. The first reports demonstrating that stimulation of the hypothalamus induces aggressive behavior were published about 70 years ago (Hess 1928). However, the precise mechanisms of these responses are still unknown. There is evidence that the lateral and frontal area of the ventromedial nucleus is of major importance. Hypothalamic aggression can be facilitated by testosterone, which indicates the initiation of offensive

Table 1
Hypothalamic-releasing hormones stimulating tropic hormones of the pituitary

Hypothalamus	Pituitary
Gonadotropin-releasing hormone (GnRH)	Follicle-stimulating hormone Luteinizing hormone
Thyrotropin-releasing hormone (TRH)	Thyroid-stimulating hormone (TSH)
Corticotropin-releasing hormone (CRH)	Adrenocorticotrophic hormone (ACTH)
Somatocrinin Somatostatin	Growth hormone (GH)
Prolactin-releasing peptide Prolactin-inhibiting factor	Prolactin

rather than defensive behavioral patterns, since androgens are ineffective in defensive behavior. Other data from the literature support the view that the hypothalamic type of aggression is offensive because behavior is not suppressed by benzodiazepines or alcohol (which is not true for defensive aggression) but by serotonin agonists, which are more potent in reducing offensive behavior. Interestingly, many peripheral hormones seem to stimulate offensive aggression by affecting hypothalamic neurons. Of special interest is the finding that injections of cortisol into the hypothalamus of golden hamsters induce aggressive behavior. This may explain why stress, which activates the HPA-axis, resulting in increased levels of glucocorticoids, is frequently associated with aggression.

One of the most important functions of the hypothalamus is the orchestration of hormone regulation. Nuclei from the hypothalamus, such as the supraoptic nucleus and the paraventricular nucleus, are sources for the two hormones oxytocin and arginine-vasopressin (often called antidiuretic hormone, ADH). These hypothalamic areas elicit neural impulses to the posterior pituitary in order to release the hormones into the circulation. The important function of vasopressin for water balance has already been discussed. Oxytocin has several functions concerning reproductive and parental behavior (see above). Furthermore, oxytocin stimulates contractions of the uterus and is involved in the milk ejection reflex. However, this hormone is present in males as well, and it has been suggested that oxytocin is of importance for social contacts and attachment.

Another fundamental influence on hormone secretion can be seen in the hypothalamic control of the anterior pituitary. The mechanism of regulating anterior pituitary hormone release is quite different from the hypothalamic influence on the posterior pituitary. While neurons from various hypothalamic nuclei transport the hormones along the axons into the posterior pituitary and release them after specific stimulation, neuroendocrine or neurosecretory cells of the hypothalamus influence the anterior pituitary. They release hormones and transport them into the

medium eminence; after stimulation by action potentials, these hormones are transported into the hypothalamic pituitary portal system. These releasing hormones reach the anterior pituitary and stimulate other cells to release tropic hormones into the blood stream; these are different from the releasing hormones provided by the hypothalamus. Table 1 shows the releasing hormones from the hypothalamus, and corresponding tropic hormones derived from the pituitary.

The stimulating effects of releasing hormones derived from the hypothalamus on secretory cells of the pituitary are of fundamental importance since they regulate energy metabolism (e.g., TRH), sexual and reproductive functions (e.g., GnRH), and growth, as well as behaviors related to adaptation (e.g., CRH) and homeostasis. The precise functions of these tropic hormones secreted in the pituitary, as well as their endocrine effects at the levels of target organs, are not within the scope of this entry. However, it should be noted, as described above, that the hypothalamus is the target for many of these peripheral hormones, leading to the modulation of behavior on the one hand and providing complex negative feedback loops for controlling further secretion of hormones on the other.

4. Summary

The hypothalamus can be considered to be a brain structure which integrates the complex interactions between hormonal and behavioral processes with the purpose of maintaining homeostasis and survival of the individual and the species (Freeman 1995). It is involved in many different fundamental processes—not only in the regulation of metabolism and hormones, but also in feeding, drinking, reproductive, maternal, and aggressive behavior. It has therefore been suggested that one of its major functions is organization of behavior.

See also: Hunger and Eating, Neural Basis of; Hypothalamic–Pituitary–Adrenal Axis, Psychobiology

of; Neuroendocrinology; Olfactory System; Sex Hormones and their Brain Receptors; Sexual Behavior and Maternal Functions, Neurobiology of; Thirst and Drinking

Bibliography

- Bligh J 1966 The thermosensitivity of the hypothalamus and thermoregulation in mammals. *Biological Reviews* **41**: 317–67
- Freeman M E 1995 The hypothalamus. In: Conn P M (ed.) *Neuroscience in Medicine*. J. B. Lippincott Company, New York
- Gorski R A, Gordon J H, Shryne J E, Southam A M 1978 Evidence for a morphological sex difference within the medial preoptic area of the rat brain. *Brain Research* **148**: 333–46
- Hakansson M L, Brown H, Ghilardi N, Skoda R C, Meister B 1998 Leptin receptor immunoreactivity in chemical defined target neurons of the hypothalamus. *Journal of Neuroscience* **18**: 559–72
- Hess W R 1928 Stammganglien-Reizversuche. *Berichte der gesamten Physiologie* **42**: 554–5
- Numan M 1978 Medial preoptic area and maternal behavior in the female rat. *Journal of Comparative and Physiological Psychology* **87**: 746–59
- Risold P Y, Thompson R H, Swanson L W 1997 The structural organization of connections between the hypothalamus and cerebral cortex. *Brain Research Reviews* **24**: 197–254
- Satinoff E 1978 Neural organization and evolution of thermal regulation in mammals. *Science* **201**: 16–22
- Sladek J R, Sladek C D 1990 Morphology of the endocrine brain, hypothalamus, and neurohypophysis. In: Becker K L (ed.) *Principles and Practice of Endocrinology and Metabolism*. J. B. Lippincott Company, New York

J. Hennig

Hypothesis Testing in Statistics

A *statistical hypothesis* is a statement about a population parameter, and the two complementary hypotheses in a hypothesis testing problem are called the *null hypothesis* and the *alternative hypothesis*. They are denoted by H_0 and H_1 , respectively.

If θ denotes a population parameter, the general format of the null and alternative hypotheses is $H_0: \theta \in \Theta_0$ and $H_1: \theta \in \Theta_0^c$ where Θ_0 is some subset of the parameter space and Θ_0^c is its complement. Typically, a hypothesis test is specified in terms of a test statistic $W(X_1, \dots, X_n) = W(\mathbf{X})$, a function of the sample. The sample space for W is partitioned into the *rejection region*, R , and its complement, the *acceptance region*. If $W \in R$ is observed, the null hypothesis H_0 is rejected and the decision is made that H_1 is true. If $W \notin R$, H_0 is accepted as true.

For example, in a study to examine the factors that influence a student's success at completing a four-year

college, one factor of interest might be socioeconomic status (SES). If, for simplicity, we just have two groups (high SES and low SES), and the success probabilities of the respective groups are p_h and p_l , then a hypothesis test of interest may be $H_0: p_h = p_l$ vs. $H_1: p_h > p_l$.

This is an example of a one-sided test, in which the alternative hypothesis specifies a direction. In contrast, we may consider the two-sided test $H_0: p_h = p_l$ vs. $H_1: p_h \neq p_l$ in which no direction is specified in the alternative. If we let $\hat{p}_h$ and $\hat{p}_l$ denote the observed success rates in the two SES classes, the one-sided test would reject the null hypothesis if $\hat{p}_h - \hat{p}_l$ is big, while the two-sided test would reject the null hypothesis if $\hat{p}_h - \hat{p}_l$ is either big or small.

A hypothesis test of $H_0: \theta \in \Theta_0$ versus $H_1: \theta \in \Theta_0^c$ might make one of two types of errors. If $\theta \in \Theta_0$ but the hypothesis test incorrectly decides to reject H_0 , then the test has made a *Type I error*. If, on the other hand, $\theta \in \Theta_0^c$ but the test decides to accept H_0 , a *Type II error* has been made. In our example, in testing $H_0: p_h = p_l$ versus $H_1: p_h > p_l$, if we conclude that $p_h > p_l$ but, in fact, $p_h \leq p_l$ we have made a Type I error.

1. Constructing Tests

There are many methods of deriving test statistics for a hypothesis test, a few of which follow.

1.1 Likelihood Ratio Tests

The likelihood ratio method of hypothesis testing is related to maximum likelihood estimators (MLE) (discussed in *Estimation: Point and Interval*). Suppose that we have a sample $X_1, \dots, X_n$, where the X_i are independent, each with distribution $f(x|\theta)$. The observed values of these random variables are denoted by $\mathbf{x} = (x_1, \dots, x_n)$, with likelihood function $L(\theta|\mathbf{x})$ given by

$$L(\theta|\mathbf{x}) = \prod_{i=1}^n f(x_i|\theta) \quad (1)$$

The *likelihood ratio test statistic* for testing $H_0: \theta \in \Theta_0$ vs. $H_1: \theta \in \Theta_0^c$ is

$$\lambda(\mathbf{x}) = \frac{\sup_{\Theta_0^c} L(\theta|\mathbf{x})}{\sup_{\Theta_0} L(\theta|\mathbf{x})} \quad (2)$$

A likelihood ratio test (LRT) is any test that has a rejection region of the form $\{\mathbf{x}: \lambda(\mathbf{x}) \leq k\}$, where k is any number satisfying $0 \leq k \leq 1$.

If we interpret the likelihood function as measuring how likely the values of θ are, then we see that the LRT

is comparing the plausibility of the θ values in the null hypothesis with those in the alternative. Small values of the LRT statistic are then interpreted as being evidence against H_0 , and lead to rejection.

If the null hypothesis consists of a single value θ_0 , and the alternative is everything else, then the LRT statistic is simply $\lambda = L(\theta_0 | \mathbf{x}) / L(\hat{\theta} | \mathbf{x})$, where $\hat{\theta}$ is the MLE of θ .

Example. Let $X_1, \dots, X_n$ be a random sample from a $n(\theta, 1)$ population. The LRT statistic for testing $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$ is

$$\lambda(\mathbf{x}) = \frac{L(\theta_0 | \mathbf{x})}{L(\bar{x} | \mathbf{x})} = \frac{(2\pi)^{-n/2} \exp[-\sum_{i=1}^n (x_i - \theta_0)^2 / 2]}{(2\pi)^{-n/2} \exp[-\sum_{i=1}^n (x_i - \bar{x})^2 / 2]} \quad (3)$$

If $T(\mathbf{X})$ is a sufficient statistic for θ then, as with maximum likelihood estimators, the LRT statistic is a function of T . That is, $\lambda(\mathbf{x})$ depends on $\mathbf{x}$ only through $T(\mathbf{x})$.

1.2 Bayesian Tests

The Bayesian paradigm prescribes that the sample information be combined with the prior information using Bayes' Theorem to obtain the posterior distribution $\pi(\theta | \mathbf{x})$. All inferences about θ are now based on the posterior distribution. In a hypothesis-testing problem, the posterior distribution may be used to calculate the probabilities that H_0 and H_1 are true.

One way a Bayesian hypothesis tester may choose to use the posterior distribution is to decide to accept H_0 as true if

$$\frac{P(\theta \in \Theta_0 | \mathbf{X})}{P(\theta \in \Theta_0^c | \mathbf{X})} \geq k \quad (4)$$

for some constant k , and to reject H_0 otherwise. Equivalently, we can reject H_0 if $P(\theta \in \Theta_0^c | \mathbf{X})$ is greater than a specified number.

Example. Let $X_1, \dots, X_n$ be independent identically distributed (i.i.d.) $n(\theta, \sigma^2)$ and let the prior distribution on θ be $n(\mu, \tau^2)$ where σ^2 , μ , and τ^2 are known. Consider testing $H_0: \theta \leq \theta_0$ vs. $H_1: \theta > \theta_0$ where we decide to accept H_0 if $P(\theta \in \Theta_0 | \mathbf{X}) \geq P(\theta \in \Theta_0^c | \mathbf{X})$. After some calculation, we find that H_0 will be accepted as true if

$$\bar{X} \leq \theta_0 + \frac{\sigma^2(\theta_0 - \mu)}{n\tau^2} \quad (5)$$

1.3 Union–Intersection and Intersection–Union Tests

In some situations, tests for complicated null hypotheses can be developed from tests for simpler null hypotheses. The union–intersection method of test construction might be useful when the null hypothesis is conveniently expressed as an intersection, say $H_0: \theta \in \bigcap_{\gamma \in \Gamma} \Theta_\gamma$, where Γ is an arbitrary index set. If tests are available for each of the problems of testing $H_{0\gamma}: \theta \in \Theta_\gamma$ vs. $H_{1\gamma}: \theta \in \Theta_\gamma^c$ where the rejection region for the test of $H_{0\gamma}$ is $\{\mathbf{x}: T_\gamma(\mathbf{x}) \in R_\gamma\}$, then the rejection region for the union–intersection test is

$$\bigcup_{\gamma \in \Gamma} \{\mathbf{x}: T_\gamma(\mathbf{x}) \in R_\gamma\} \quad (6)$$

The rationale is that if any one of the hypotheses $H_{0\gamma}$ is rejected, then H_0 must also be rejected.

A complementary method, the intersection–union method, may be useful if the null hypothesis is conveniently expressed as a union. Suppose we wish to test the null hypothesis $H_0: \theta \in \bigcup_{\gamma \in \Gamma} \Theta_\gamma$, and $\{\mathbf{x}: T_\gamma(\mathbf{x}) \in R_\gamma\}$ is the rejection region for a test of $H_{0\gamma}: \theta \in \Theta_\gamma$ vs. $H_{1\gamma}: \theta \in \Theta_\gamma^c$. Then the rejection region for the intersection–union test of H_0 versus H_1 is

$$\bigcap_{\gamma \in \Gamma} \{\mathbf{x}: T_\gamma(\mathbf{x}) \in R_\gamma\} \quad (7)$$

and H_0 is false if and only if all of the $H_{0\gamma}$ are false, so H_0 can be rejected if and only if each of the individual hypotheses $H_{0\gamma}$ can be rejected.

Example. Returning to our example about success in four-year schools, we might argue that there are measures of success other than finishing. For example, yearly income could be used to measure success, and denoting mean incomes in the high and low SES groups by θ_h and θ_l respectively results in the hypothesis test

$$H_0: \theta_h \leq \theta_l \text{ or } p_h \leq p_l \quad \text{vs.} \quad H_1: \theta_h > \theta_l \text{ and } p_h > p_l \quad (8)$$

where the groups are considered different only if H_1 is accepted.

If we have estimators $\hat{\theta}_h$, $\hat{\theta}_l$, $\hat{p}_h$ and $\hat{p}_l$, a rejection region for the intersection–union test could be given by

$$\{(\hat{\theta}_h, \hat{\theta}_l, \hat{p}_h, \hat{p}_l): \hat{\theta}_h - \hat{\theta}_l > k_1 \text{ and } \hat{p}_h - \hat{p}_l > k_2\} \quad (9)$$

Thus the intersection–union test decides that the high and low SES groups have different levels of success, that is, H_1 is true, if and only if it decides that each of the individual parameters are different.

The topic of acceptance sampling, in which one tests whether to accept a product based on a collection of

measurements, provides an extremely useful application of an intersection–union test (see Berger 1982).

There are many other methods available for constructing hypothesis tests, methods based on invariance, pivots, robust, or large sample arguments, to name a few. For more on hypothesis testing see Lehmann (1986).

2. Evaluating Tests

As mentioned previously, a hypothesis test might make one of two different kinds of errors. The probabilities of making these errors can be calculated using the power function. If R denotes the rejection region for a test, the power function is

$$P_{\theta}(\mathbf{X} \in R) = \begin{cases} \text{probability of a Type I error} & \text{if } \theta \in \Theta_0 \\ 1 - \text{the probability of a Type II error} & \text{if } \theta \in \Theta_0^c \end{cases} \quad (10)$$

A good test has power function near one for most $\theta \in \Theta_0^c$ and near zero for most $\theta \in \Theta_0$.

Example. Let $X_1, \dots, X_n$ be a random sample from a $n(\theta, \sigma^2)$ population, σ^2 known. The likelihood ratio test of $H_0: \theta \leq \theta_0$ vs. $H_1: \theta > \theta_0$ rejects H_0 if $(\bar{X} - \theta_0)/(\sigma/\sqrt{n}) > k$ and has power function

$$P_{\theta}(\mathbf{X} \in R) = P \left(Z > k + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}} \right) \quad (11)$$

where Z is a standard normal random variable.

After a hypothesis test is done, the conclusions must be reported in some statistically meaningful way. One method of reporting the results of a hypothesis test is to report the size ($\sup_{\theta \in \Theta_0} P_{\theta}(\mathbf{X} \in R)$), α , of the test used and the decision to reject H_0 or accept H_0 . The size of the test carries important information. If α is small, the decision to reject H_0 is fairly convincing, but if α is large, the decision to reject H_0 is not very convincing because the test has a large probability of incorrectly making that decision.

Another way of reporting the results of a hypothesis test, one that is data dependent, is to report the p -value. Typically, not one but an entire class of tests are constructed, a different test being defined for each value of α . The p -value for the sample point $\mathbf{x}$ is the smallest value of α for which this sample point will lead to rejection of H_0 (see *Significance, Tests of*).

Because rejection of H_0 using a test with small size is more convincing evidence that H_1 is true than rejection

of H_0 with a test with large size, the interpretation of p -values goes in the same way. The smaller the p -value the stronger the sample evidence that H_1 is true.

Many other types of evaluations of test can be done. The theory of most powerful tests shows how to construct best tests under a variety of conditions (see Lehmann 1986 or Casella and Berger 2001, Chap. 8). Hypothesis tests can also be evaluated using risk functions (see *Decision Theory: Classical*), as in Hwang et al. (1992).

3. Asymptotics

For the likelihood ratio test statistic (2), the following general theorem allow us to construct a large sample test.

Theorem 1. Let $X_1, \dots, X_n$ be a random sample from a probability distribution function or probability mean function $f(x|\theta)$. Under some regularity conditions on the model $f(x|\theta)$, if $\theta \in \Theta_0$ then the distribution of the statistic $-2 \log \lambda(\mathbf{X})$ converges to a chi-squared distribution as the sample size $n \rightarrow \infty$. The degrees of freedom of the limiting distribution is the difference between the number of free parameters specified by $\theta \in \Theta_0$ and the number of free parameters specified by $\theta \in \Theta$.

(The ‘regularity conditions’ are concerned mainly with the existence and behavior of the derivatives (with respect to the parameter) of the likelihood function, and the support of the distribution (it cannot depend on the parameter). See Casella and Berger (2001, Sect. 10.6) or Lehmann (1986, Sect. 8.8) for precise conditions.)

Rejection of $H_0: \theta \in \Theta_0$ for small values of $\lambda(\mathbf{X})$ is equivalent to rejection for large values of $-2 \log \lambda(\mathbf{X})$. Thus,

$$H_0 \text{ is rejected if and only if } -2 \log \lambda(\mathbf{X}) \geq \chi_{v, \alpha}^2$$

where v is the degrees of freedom specified in Theorem 1.

As might be expected, Theorem 1 has wide applicability. In particular, it is extremely useful in categorical data analysis (see *Multivariate Analysis: Discrete Variables (Loglinear Models)*).

Other large-sample test constructions are based on asymptotic normality of a point estimator (see *Estimation: Point and Interval*). Suppose we wish to test a hypothesis about a real-valued parameter θ , and $W_n = W(X_1, \dots, X_n)$ is a point estimator of θ , based on a sample of size n , that satisfies

$$\frac{W_n - \theta}{\sqrt{\text{Var } W_n}} \rightarrow Z$$

where $\text{Var } W_n$ is the variance of W_n and Z is a standard normal random variable. We now have the basis for an approximate test, for example, we would reject $H_0: \theta \leq \theta_0$ at level 0.05 if $(W_n - \theta_0)/\sqrt{\text{Var } W_n} > 1.645$. Note that $\text{Var } W_n$ could depend on θ_0 and we can still use it in the test statistic. This type of test, where we use the actual variance of W_n , is called a *score test*.

If $\text{Var } W_n$ also depends on unknown parameters we could look for an estimate S_n^2 of $\text{Var } W_n$ with the property that $(\text{Var } W_n)/S_n^2$ converges in probability to one. Then, using Slutsky's Theorem (see Casella and Berger 2001, Sect. 5.5), we can deduce that $(W_n - \theta)/S_n$ also converges in distribution to a standard normal distribution. The large-sample test based on this statistic is called a *Wald test*.

4. Conclusions

Hypothesis testing is one of the most widely used, and some may say abused, methodologies in statistics. Formally, the hypotheses are specified, an α -level is chosen, a test statistic is calculated, and it is reported whether H_0 or H_1 is accepted. In practice, it may happen that hypotheses are suggested by the data, the choice of α -level may be ignored, more than one test statistic is calculated, and many modifications to the formal procedure may be made. Most of these modifications cause bias and can invalidate the method. For example, a hypothesis suggested by the data is likely to be one that has 'stood out' for some reason, and hence H_1 is likely to be accepted unless the bias is corrected for (using something like Scheffé's method—see Hsu 1996).

Perhaps the most serious criticism of hypothesis testing is the fact that, formally, it can only be reported that either H_0 or H_1 is accepted at the prechosen α -level. Thus, the same conclusion is reached if the test statistic only barely rejects H_0 and if it rejects H_0 resoundingly. Many feel that this is important information that should be reported, and thus it is almost required to also report the p -value of the hypothesis test.

For further details on hypothesis testing see the classic book by Lehmann (1986). Introductions are also provided by Casella and Berger (2001) or Schervish (1995), and a good introduction to multiple comparisons is Hsu (1996); see also *Hypothesis Tests, Multiplicity of*.

See also: Explanation: Conceptions in the Social Sciences; Hypothesis Testing: Methodology and Limitations

Bibliography

Berger R L 1982 Multiparameter hypothesis testing and acceptance sampling. *Technometrics* **24**: 295–300

Casella G, Berger R L 2001 *Statistical Inference*, 2nd edn. Wordsworth/Brooks Cole, Pacific Grove, CA

Hsu J C 1996 *Multiple Comparisons, Theory and Methods*. Chapman & Hall, London

Hwang J T, Casella G, Robert C, Wells M T, Farrell R H 1992 Estimation of accuracy in testing. *Annals of Statistics* **20**: 490–509

Lehmann E L 1986 *Testing Statistical Hypotheses*, 2nd edn. Springer, New York

Schervish M 1995 *Theory of Statistics*. Springer, New York

G. Casella and R. L. Berger

Hypothesis Testing: Methodology and Limitations

Hypothesis tests are part of the basic methodological toolkit of social and behavioral scientists. The philosophical and practical debates underlying their application are, however, often neglected. The fruitful application of hypothesis testing can benefit from a clear insight into, the underlying concepts and their limitations.

1. The Idea of Hypothesis Testing

A test is a statistical procedure to obtain a statement on the truth or falsity of a proposition, on the basis of empirical evidence. This is done within the context of a model, in which the fallibility or variability of this empirical evidence is represented by probability. In this model, the evidence is summarized in observed data, which is assumed to be the outcome of a stochastic, i.e., probabilistic, process; the tested proposition is represented as a property of the probability distribution of the observed data.

1.1 Some History

The first published statistical test was by John Arbuthnot in 1710, who wondered about the fact that in human births, the fraction of boys born year after year appears to be slightly larger than the fraction of girls (cf. Hacking 1965). He calculated that this empirical fact would be exceedingly unlikely (he obtained a probability of $1/483\,600\,000\,000\,000\,000\,000\,000$) if the probability of a male birth were exactly 0.5, and argued that this was a proof of divine providence, since boys—some of whom will be soldiers—have a higher risk of an early death, so that a higher ratio of male births is needed to obtain an equal ratio of males among young adults. We see here the basic elements of a test: the proposition

that the male birth ratio is 0.5, related to data by regarding these as outcomes of stochastic variables, the calculation that the data would be unlikely if the proposition were true, and a further conclusion interpreting the falsity of the proposition.

One of the first statistical procedures that comes close to a test in the modern sense was proposed by Karl Pearson in 1900. This was the famous chi-squared test for comparing an observed frequency distribution to a theoretically assumed distribution. Pearson derived the now well-known statistic to test the proposition, or *hypothesis*, that the probabilities of the various possible (finitely many) outcomes of some random variable are given by certain preassigned numbers. Pearson proved that this statistic has (in a large-sample approximation) the chi-squared distribution; this distribution can therefore be used to calculate the probability that, if the hypothesis holds, the test statistic will assume a value equal to or larger than the value actually observed.

The idea of testing was further codified and elaborated in the first decades of the twentieth century, mainly by R. A. Fisher (e.g., 1925). In his significance tests the data are regarded as the outcome of a random variable X (usually a vector or matrix), which has a probability distribution which is a member of some family of distributions; the tested hypothesis, also called the null hypothesis, is an assertion which defines a subset of this family; a test statistic $T = t(X)$, which is a function of X , is used to indicate the degree to which the data deviate from the null hypothesis; and the significance of the given outcome of the test statistic is calculated as the probability, if the null hypothesis is true, to obtain a value of T which is at least as high as the given outcome. If the probability distribution of T is not uniquely determined by the null hypothesis, then the significance is the maximal value of this probability, for all distributions of T compatible with the null hypothesis. The significance probability is now often called the p -value (the letter p referring to probability). With Fisher originates the convention to consider a statistical testing result as 'significant' if the significance probability is 0.05 or less—but Fisher was the first to recognize the arbitrary nature of this threshold.

A competing approach was proposed in 1928 by J. Neyman and Egon Pearson (the son of Karl). They criticized the arbitrariness in Fisher's choice of the test statistic and asserted that for a rational choice of test statistic one needs not only a null hypothesis but also an alternative hypothesis, which embodies a proposition that competes with the proposition represented by the null hypothesis. They formalized the testing problem as a two decision problem. Denoting the null hypothesis by H_0 and the alternative H_1 , the two decisions were represented as 'reject H_0 ' and 'do not reject H_0 .' (The second decision was also called 'accept H_0 ,' but it will be discussed below that this is an unfortunate term.) Two errors are possible: rejecting a

true H_0 , and failing to reject a false H_0 . Neyman and Pearson conceived of the null hypothesis as a standard situation, the burden of proof residing with the researcher to demonstrate (if possible) the untenability of this proposition.

Correspondingly, they called the error of rejecting a true H_0 an error of the first kind and the error of failing to reject a false H_0 an error of the second kind. Errors of the first kind are considered more serious than errors of the second kind. The probability of—correctly—rejecting H_0 if H_1 is true, which is 1 minus the probability of an error of the second kind, given that the alternative hypothesis is true, they called the power of the test. Neyman and Pearson proposed the requirement that the probability of an error of the first kind, given that the null hypothesis is indeed true, do not exceed some threshold value called the significance level usually denoted by α . Further they proposed to determine the test so that, under this essential condition, the power will be maximal.

In the Neyman–Pearson formulation, we obtain richer results (namely, specific precepts for constructing good tests) at the cost of a more demanding model. In addition to Fisher's null hypothesis, we also need to specify an alternative hypothesis; and we must conceive the testing problem as a two-decision situation. This led to vehement debate between Fisher on the one hand, and Neyman and E. Pearson on the other. This debate and the different philosophical positions are summarized by Hacking (1965) and Gigerenzer et al. (1989), who also give a further historical account. The latter study also discusses how this controversy was resolved in the teaching and practice of statistics in the social sciences by a hybrid theory which combines elements of both approaches to testing, and which has been treated often as an objective precept for the scientific method, glossing over the philosophical controversies.

Examples of this hybrid character are that, in accordance with the Neyman–Pearson approach, the theory is explained by making references to both the null and the alternative hypotheses, and to errors of the first and second kind (although power tends to be treated in a limited and often merely theoretical way), whereas—in the spirit of Fisher—statistical tests are regarded as procedures to give evidence about the particular hypothesis tested and not merely as rules of behavior that will in the long run have certain (perhaps optimal) error rates when applied to large numbers of hypotheses and data sets. Lehmann (1993) argues that indeed a unified formulation is possible, combining the best features of both approaches.

Instead of implementing the hypothesis test as a 'reject/don't reject' decision with a predetermined significance level, another approach often is followed: to report the p -value or significance probability, defined as the smallest value of the significance level at which the observed outcome would lead to rejection of the null hypothesis. Equivalently, this can be defined

as the probability, calculated under the null hypothesis, of observing a result deviating from the null hypothesis at least as much as the actually observed result. This deviation is measured by the test statistic, and the p -value is just the tail probability of the test statistic. For a given significance level α , the null hypothesis is rejected if and only if $p \leq \alpha$.

2. Hypothesis Tests in Empirical Research

2.1 The Role of the Null Hypothesis

In the social and behavioral sciences, the standard use of hypothesis testing is directed at single research questions, practically always fragments of larger investigations, about the existence of some or other effect. This effect could be a difference between two group means, the effect of some explanatory variable on some dependent variable as expressed by a regression coefficient in multiple linear regression analysis, etc. Expressed schematically, the researcher would like to demonstrate a research hypothesis which states that the effect in question, indeed, exists. The negation of this hypothesis, then, is understood as the proposition that the effect is absent, and this proposition is put up as the null hypothesis. The research hypothesis, stating that the effect exists, is the alternative hypothesis. Rejecting the null hypothesis is interpreted as support for the existence of the hypothesized effect. In this way, the burden of proof rests with the researcher in the sense that an error of the first kind is to state that there is evidence supporting the existence of the effect if, in reality, the effect is absent. The usual welcoming phrase for the rejection of the null hypothesis is the statement that a significant result has been obtained.

2.2 Example: The t -test

Consider, as a major example, the proposition that two groups, arbitrarily labeled A and B , are different with respect to some numerical characteristic X . Available data are measurements X_{Ai} of this characteristic for some individuals i in group A and measurements X_{Bi} of the characteristics for other individuals i in group B . The first step in modeling this is to consider the observed values X as outcomes of random variables, usually with the assumption that they are stochastically independent and have a probability distribution not depending on the individual i . The next step is, usually, to focus on the expected values, i.e., population means in the two groups, denoted here by μ_A and μ_B . The tested proposition is formalized as the statement that μ_A and μ_B are different (two-sided alternative hypothesis) or as the statement that one (e.g., μ_A) is bigger than the other (e.g., μ_B) (one-sided alternative hypothesis). The null hypothesis H_0 is defined as the statement that μ_A and μ_B are equal.

The most commonly used test for this testing problem is *Student's t -test*, called after the pseudonym of W. Gosset, who laid the mathematical basis for this test in 1908. The test statistic is

$$T = \frac{M_A - M_B}{\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right) S^2}}$$

where M_A and M_B are the two sample means, S^2 is the pooled within-group variance, and n_A and n_B are the two sample sizes. (Definitions of these quantities can be found in any statistics textbook.) This formula illustrates the property of many test statistics that an observed effect (here the difference between the two sample means) is compared to a measure of variability (based here on the within-group variance). Student/Gosset showed that, if the variable X has a normal distribution with the same variance in the two groups, then T has, if the null hypothesis is true, the so-called t distribution on $df = n_A + n_B - 2$ degrees of freedom. For the two-sided alternative hypothesis, H_0 is rejected for large values of the absolute value of T , for the one-sided hypothesis H_0 is rejected for large values of T itself. The threshold beyond which H_0 is rejected is called critical value, and is determined from the t -distribution in such a way that the significance level has the pre-assigned value. The one-sided threshold at significance level α is denoted $t_{df, \alpha}$, so that the 'rejection region' for the one-sided t -test is given by $\{T > t_{df, \alpha}\}$. The power of the one-sided test is larger than that of the two-sided test for $\mu_A > \mu_B$ (corresponding to the one-sided alternative hypothesis) and smaller for $\mu_A < \mu_B$. This is, in the Neyman-Pearson formulation, the reason for using the one-sided test if the alternative hypothesis is one-sided (in which case parameter values $\mu_A < \mu_B$ are not taken into consideration).

The conventional value of the significance level in the social and behavioral sciences is 0.05. For large values of the combined sample size $n_A + n_B$, the critical value of the t -test approximates the critical value that can be computed from the standard normal distribution. This is because the fact that the variance is estimated and not known beforehand becomes less and less important as the combined sample size gets larger; if the population variance was known beforehand and substituted for S^2 , the test statistic would have the standard normal distribution, and the test would be the so-called z -test. For this test, the critical value is 1.645 for the one-sided test and 1.960 for the two-sided test. The power depends on many quantities: the sample sizes, the population means and variances, and the significance level. As examples, the power of the one-sided z -test for $\alpha = 0.05$ is equal to 0.50 for

$$\frac{\mu_A - \mu_B}{\sigma} = 1.645 \sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

where σ is the within-group standard deviation. The power is equal to 0.95 if $(\mu_A - \mu_B)/\sigma$ is equal to twice this value.

2.3 The Role of Assumptions

The probability statements that are required for statistical tests do not come for free, but are based on certain assumptions about the observations used for the test. In the two-sample *t*-test, the assumptions are that the observations of different individuals are outcomes of statistically independent, normally distributed, random variables, with the same expected value for all individuals within the same group, and the same variance for all individuals in both groups. Such assumptions are not automatically satisfied, and for some assumptions it may be doubted whether they are ever satisfied exactly. The null hypothesis H_0 and alternative hypothesis H_1 are statements which, strictly speaking, imply these assumptions, and which therefore are not each other's complement. There is a third possibility: the assumptions are invalid, and neither H_0 nor H_1 is true. The sensitivity of the probabilistic properties of a test to these assumptions is referred to as the lack of robustness of the test. The focus of robustness studies has been on the assumptions of the null hypothesis and the sensitivity of the probability of an error of the first kind to these assumptions, but studies on robustness for deviations from assumptions of the alternative hypothesis have also been done, cf. Wilcox (1998).

One general conclusion from robustness studies is that tests are extremely sensitive to the independence assumptions made. Fortunately, those assumptions are often under control of the researcher through the choice of the experimental or observational design. Traditional departures from independent observations are multivariate observations and within-subject repeated measures designs, and the statistical literature abounds with methods for such kinds of dependent observations. More recently, methods have been developed for clustered observations (e.g., individual respondents clustered within groups) under the names of multilevel analysis and hierarchical linear modeling.

Another general conclusion is that properties of tests derived under the assumption of normal distributions, such as the *t*-test, can be quite sensitive to outliers, i.e., single, or a few, observations that deviate strongly from the bulk of the observations. Since the occurrence of outliers has a very low probability under normal distributions, they are 'assumed away' by the normality assumption. The lack of robustness and sensitivity to outliers have led to three main developments.

First, there are non-parametric tests, which do not assume parametric families of distributions (such as the normal). Most of these are based on the ranks of

the observations rather than their numerical values. They are standard fare in statistical textbooks. Second, robust tests have been developed, based on numerical values of the observations, which are less sensitive to outliers or heavy-tailed distributions, e.g., by some kind of automatic downweighting of outlying observations (e.g., Wilcox, 1998). The purpose of such tests is to retain a high power as much as possible while decreasing the sensitivity to deviations from the normal distribution or to departures from other assumptions.

Third, diagnostic means have been developed to spot single observations, or groups of observations, with an undue influence on the result of the statistical procedure (e.g., Cook and Weisberg 1999, Fox 1991). The idea behind most of these diagnostics is that most of the observations come from a distribution which corresponds well enough to the assumptions of the statistical model, but that the observations could be contaminated by a small number of poorly fitting observations. Ideally, these observations should also be recognizable by close inspection of the data or data collection process. After deleting such poorly fitting observations one can proceed with the more traditional statistical procedures, assuming that the remaining data do conform to the model assumptions.

3. Confidence Intervals

Confidence intervals are closely related to hypothesis tests. They focus on a parameter in the statistical model. Examples of such parameters are, in the two-sample situation described above, the difference of the two population means, $\mu_A - \mu_B$, or the within-group standard deviation, σ . To frame this in general terms, consider a one-dimensional statistical parameter denoted by θ . A null hypothesis test is about the question whether the data is compatible with the hypothesis that the parameter θ (in this case $\mu_A - \mu_B$) has the particular value 0. Instead, one could put forward the question: what are the values of θ with which the data is compatible? This question can be related to hypothesis testing by considering the auxiliary problem of testing the null hypothesis $H_1: \theta = \theta_0$ against the alternative hypothesis $H_1: \theta \neq \theta_0$, for an arbitrary but fixed value θ_0 . The data may be said to be compatible with any value of θ_0 for which this null hypothesis would not be rejected at the given level of significance. Thus, the confidence interval is the interval of non-rejected null hypotheses.

Another way of viewing the confidence interval is as follows. A confidence coefficient $1 - \alpha$ is determined by the researcher (a conventional value is 0.95). The confidence interval is an interval with lower boundary L and upper boundary U , calculated from the data and therefore being random variables, with the property that the probability that $L \leq \theta \leq U$, i.e., the

interval contains the true parameter value, is at least $1 - \alpha$. It can be proved mathematically that the interval of non-rejected null hypotheses has precisely this property.

The confidence interval can be inverted to yield a hypothesis test for any null hypothesis of the form $H_0: \theta = \theta_0$ in the following way. For the most usual null hypothesis that $\theta = 0$, the hypothesis is rejected if the value 0 is not included in the confidence interval. More generally, the null hypothesis $H_0: \theta = \theta_0$ is rejected whenever θ_0 is outside the confidence interval. This shows that a confidence interval is strictly more informative than a single hypothesis test.

4. Problems in the Interpretation and Use of Hypothesis Tests

While hypothesis tests have been applied routinely by hosts of social and behavioral scientists, there have been continuing debates, sometimes vehement and polemic, about their use, among those inclined to philosophical or foundational discussion. Many contentious issues can be found in Harlow et al. (1997), Nickerson (2000), and Wilkinson and TFSI (1999). Within the context of the present contribution I can only give a personally colored discussion of some main points which might contribute to a better understanding and a more judicious application of hypothesis tests.

The reason that there has been so much criticism of the nature and use of hypothesis tests, is in my view the difficulty of reasoning with uncertain evidence and the natural human tendency to take recourse to behavioral rules instead of cogitate afresh, again and again, about how to proceed. The empirical researcher would like to conclude unequivocally whether a theory is true or false, whether an effect is present or absent. However, experimental and observational data on human behavior are bound to be so variable that the evidence produced by these data is uncertain. Our human tendency is to reason—if only provisionally—as if our conclusions are certain rather than tentative; to be concise rather than accurate in reporting our arguments and findings; and to use simple rules for steering how we make inference from data to theories. We are tempted to reason as if ' $p > \alpha$ ' implies that the investigated effect is nil, and the presumed theory is false, while ' $p \leq \alpha$ ' proves that the effect indeed exists and the theory is true.

However, such cognitive shortcuts amount to serious misuse of hypothesis testing. Much of the criticism of the use of hypothesis testing therefore is well-founded. In my view this implies that better interpretations of hypothesis testing should be promoted, and that tests should be used less mechanically and combined with other argumentations and other statistical procedures. In any case, the user of tests (or

other statistical procedures) should realize that these procedures are based on the assumption that there is random variability in the data which cannot be completely filtered out of the results. Whether a test result is significant or not depends partly on chance, and each researcher obtaining a significant result should be aware that this could be an error of the first kind, just as a non-significant result could be an error of the second kind.

4.1 Some Misinterpretations

The following list tries to correct misinterpretations that have haunted applications of statistical tests. A more extensive discussion can be found, e.g., in Nickerson (2000).

(a) A common misinterpretation is that non-rejection implies support for the null hypothesis. Nonrejection should be interpreted, however, as an undecided outcome: there is not enough evidence against the null hypothesis, but this does not imply that there is evidence for the null hypothesis. Maybe the sample size is small, or error variability is large, so that the data does not contain much information anyway. Usually the alternative hypothesis contains probability distributions that approximate the null distribution; e.g. when testing $\mu_A - \mu_B = 0$ against $\mu_A - \mu_B \neq 0$, the difference between the two population means can be tiny, while the alternative hypothesis is still true. In other words, the power of practically all tests is quite low in a sizeable part of the alternative hypothesis, so if one of the distributions in this part would prevail, chances would be low of rejecting the null hypothesis. Therefore, nonrejection provides support for those parts of the alternative practically as strongly as for the null hypothesis itself, and non-rejection may not be interpreted as support for the null hypothesis and against the alternative hypothesis.

(b) If one wishes to get more information about whether a nonsignificant result provides support for the null hypothesis, a power study is not the answer. Statistical power is the probability to reject the null hypothesis, if a given effect size obtains. Its interpretation cannot be inverted as being the degree of support of the null hypothesis in the case of nonsignificance. Power studies are important in the stage of planning a study. Once the study has been conducted, and one wishes to see in more detail to what extent the null and alternative hypotheses are supported, a confidence interval is the appropriate procedure (also see Wilkinson and TFSI, 1999, p. 596).

(c) Test results tell us nothing about the probabilities of null or alternative hypotheses. The probabilities known as the significance level or the power of the test are probabilities of certain sets of outcomes given the condition that the null or, respectively, the alternative hypothesis is true.

(d) A significant result is not necessarily important. Very low p -values do not in themselves imply large effect sizes. A small estimated effect size still can yield a low p -value, e.g., if the corresponding standard error is very small, which can be caused by a large sample or by good experimental control. Investigation of effect sizes is important, and is a different thing than hypothesis testing. The importance of reporting effect sizes is stressed in Schmidt (1996) and Wilkinson and TFSI (1999).

(e) The alternative hypothesis is not the same as a scientific theory. If a researcher is investigating some theory empirically, this can be based on tests of which the alternative hypotheses are deduced from the theory, but always under the assumption that the study was well designed and usually under additional technical assumptions such as, e.g., normality of the distributions. Since the alternative hypothesis is only a consequence and not a sufficient condition for the theory, it is possible that the alternative hypothesis is true but the theory is not. To find support for theories, one has to find corroborations of many consequences of the theory under various circumstances, not just of one of its implications. On the other hand, it is possible that the theory is true but the alternative hypothesis deduced from it is not, e.g., because the experiment or observation was ill-designed or because the auxiliary technical assumptions are not satisfied.

(f) The null hypothesis is just what it says: a hypothesis. The test is not invalidated by the mere fact that there are other reasons for thinking, or knowing, that the null hypothesis is wrong. Even if there may be other reasons for which the null hypothesis is wrong, still it can be sensible to check whether the data at hand are, or are not, compatible with it.

4.2 Limitations of the Neyman–Pearson Approach to Hypothesis Testing

The Neyman–Pearson (1928) formulation changed the concept of hypothesis testing because it provided a rational basis for the choice of a test statistic. This was obtained at the cost of a quite narrow frame for the notion of a hypothesis test: a two-decision problem in which one error ('type I') is considered much more serious than the other error ('type II'). This model for decision-making is often inappropriate and, when it is useful, it is more commonly a useful approximation than a faithful representation of the actual decision situation. Some of its limitations are discussed briefly in this section.

(a) Even when a dichotomous decision must be taken as a result of a scientific investigation, it is possible that both types of error should be considered equally serious. This is the case, e.g., when two therapies are compared and the most promising one must be chosen for further study or for practical application. If the costs of both kinds of error (choose

therapy B , while A is better; or choose therapy A , while B is better) are about the same, then there is no asymmetry between the two competing hypotheses, and a significance test is not in order.

(b) Often there are more than two outcomes of the decision situation. For example, in the two-sample situation discussed above, where the population means are μ_A and μ_B , and where the researcher entertains a theory implying that $\mu_A > \mu_B$, it is natural to define three decision outcomes defined as ' $\mu_A > \mu_B$ (support for the theory),' 'undecided,' and ' $\mu_A < \mu_B$ (evidence against the theory).' Usually, however, such a situation is represented by a Neyman–Pearson testing problem with null hypothesis H_0 : ' $\mu_A = \mu_B$ ' and alternative H_1 : ' $\mu_A > \mu_B$.' If the t -statistic for $\mu_A - \mu_B$ yields a strongly negative outcome ('significance in the wrong direction') then the researcher only concludes that the null hypothesis is not rejected, whereas the data conveys the message that the alternative hypothesis should be rejected. The data then have to be considered as evidence against the theory embodied in H_1 . Such a three-decision formulation is closer to the actual purpose of hypothesis tests in many situations where a one-dimensional parameter is being tested, cf. Leventhal (1999).

(c) The Neyman–Pearson approach is formulated as if the data at hand is the only evidence available about the null and alternative hypotheses: it is an *in vacuo* formulation. This is almost never realistic. Often the propositions under study are theoretically plausible or implausible, there is earlier empirical evidence, there may be converging evidence (as in the case of triangulation studies) in the same investigation, and other evidence may become available in the future. In the light of this surrounding evidence, it is unrealistic—perhaps arrogant—to presume that the researcher has to take either of two decisions: 'reject H_0 ' or 'do not reject H_0 .' It is more realistic to see the data at hand as part of the evidence that was and will be accumulated about these hypotheses. This is a major reason why presenting the result of the test as a p -value is more helpful to other researchers than presenting only the 'significant–nonsignificant' dichotomy (also see Wilkinson and TFSI 1999).

There are several ways in which the results obtained from the data at hand, perhaps summarized in the p -value, can be combined with other evidence. In the first place, an informal combination can be and often will be made by a sensible, nonformalized weighing of the various available pieces of evidence. One formal way of combining evidence is provided by the Bayesian paradigm (e.g., Gelman et al. 1995). This paradigm presupposes, however, that the prior and current ideas of the researcher about the plausibility of all possible values of the statistical parameter under study be represented in a probability distribution for this parameter, a very strong requirement indeed. Other formal ways of combining evidence have been developed and are known collectively by the name of

meta-analysis (e.g., Cook et al. 1992, Schmidt 1996). These methods require that the several to-be-combined hypothesis tests address the same substantive hypothesis, or that they can be regarded as tests of the same substantive parameter.

5. Conclusion

The statistical hypothesis test is one of the basic elements of the toolbox of the empirical researcher in the social and behavioral sciences. The difficulty of reasoning on the basis of uncertain data has led, however, to many misunderstandings and unfortunate applications of this tool. The Neyman–Pearson approach is a mathematically elegant formulation of what a hypothesis test could be. Its precise formulation rarely applies literally to the daily work of social and behavioral scientists, very often it is a useful approximation to a part of the research question, sometimes it is inappropriate—but sacrosanct it is not. A hypothesis test evaluates the data using a test statistic set up to contrast the null hypothesis with the alternative hypothesis, and the p -value is the probability to obtain, *if* the null hypothesis is true, outcomes of the test statistic that are at least as high as the outcome actually observed. Low values therefore are evidence against the null hypothesis, contrasted with the alternative hypothesis. It is more conducive to the advance of science to report p -values than merely whether the hypothesis was rejected at the conventional 0.05 level of significance.

In the interpretation of test outcomes one should be aware that these are subject to random variability; and that the probability calculations are based on assumptions which may be in error. Incorrect assumptions can invalidate the conclusions seriously. Nonparametric and robust statistical procedures have been developed which depend less on these kinds of assumption, and diagnostics have been developed for checking some of the assumptions. When reporting empirical results, it is usually helpful to report not only tests but also estimated effect sizes and/or confidence intervals for important parameters. To sum up: ‘It seems there is no escaping the use of judgment in the use and interpretation of statistical significance tests’ (Nickerson 2000, p. 256).

See also: Hypothesis Testing in Statistics; Hypothesis Tests, Multiplicity of; Model Testing and Selection, Theory of; Significance, Tests of

Bibliography

- Cook R D, Weisberg S 1999 *Applied Regression Including Computing and Graphics*. Wiley, New York
- Cook T D, Cooper H, Cordray D S, Hartmann H, Hedges L V, Light R J, Louis T A, Mosteller F 1992 *Meta-analysis for Explanation. A Casebook*. Russell Sage Foundation, New York
- Fisher R A 1925 *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, UK
- Fox J 1991 *Regression Diagnostics*. Sage, Newbury Park, CA
- Gelman A, Carlin J B, Stern H S, Rubin D B 1995 *Bayesian Data Analysis*. Chapman & Hall, London
- Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Krüger L 1989 *The Empire of Chance*. Cambridge University Press, New York
- Hacking I 1965 *Logic of Statistical Inference*. Cambridge University Press, Cambridge, UK
- Harlow L L, Mulaik S A, Steiger J H (eds.) 1997 *What if there were no Significance Tests?* Erlbaum, Mahwah, NJ
- Lehmann E L 1993 The Fisher, Neyman–Pearson theories of testing hypotheses: one theory or two? *Journal of the American Statistical Association* **88**: 1242–9
- Leventhal L 1999 Updating the debate on one- versus two-tailed tests with the directional two-tailed test. *Psychological Reports* **84**: 707–18
- Neyman J, Pearson E S 1928 On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* **20A**: 175–240, 263–94
- Nickerson R S 2000 Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods* **5**: 241–301
- Schmidt F L 1996 Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods* **1**: 115–29
- Wilcox R R 1998 The goals and strategies of robust methods (with discussion). *British Journal of Mathematical and Statistical Psychology* **51**: 1–64
- Wilkinson L, Task force on statistical inference 1999 Statistical methods in psychology journals. *American Psychologist* **54**: 594–604

T. A. B. Snijders

Hypothesis Tests, Multiplicity of

Standard procedures for controlling the probability of erroneous statistical inferences apply only for a single comparison or determination from a given set of data. Multiplicity is encountered whenever a data analyst draws more than one inference from a data set. With multiplicity, standard procedures must be adjusted for the number of comparisons or determinations that actually or potentially are performed. If some adjustment is not made, the probability of erroneous inferences can be expected to increase with the degree of multiplicity. In other words, the data analyst loses control of a critical feature of the analysis.

Questions arising from problems of multiplicity raise a diversity of issues, issues that are important, difficult, and not fully resolved. This account empha-

sizes the issues, tries to avoid being trapped in undue detail, and provides examples of the application of a recommended procedure.

It is helpful to recognize several general guidelines:

(a) Two approaches to the same data, with different purposes or with different questions being posed, need not yield compatible results.

(b) Exact equality of long-run values is very rare and should not play an appreciable role either in one's thinking or one's practice.

(c) When posing questions about data, greater realism will be very valuable.

(d) The situations faced must not be characterized either as completely 'fixed' or completely 'random.'

(e) In selecting a bouquet of techniques for the next decade or the next millennium, the selection must not be cast in stone; further evolution is to be expected (in later months, years, or centuries).

Issues of multiplicity appear in a wide variety of settings, including

(a) outcomes, each measured with variability, in several situations:

(i) comparison of all pairs of situations ('multiple comparisons'); and

(ii) comparison between each situation and a standard ('multiple determination');

(b) selection of one or more elements from a set of alternatives:

(i) selection of how variables are to be expressed for analysis (i.e., what re-expression, if any, should be used?);

(ii) selection of which of the available variables are to participate in an analysis (subset selection); and

(iii) reduction in size of a set of alternatives in an attempt to reduce their number without losing the best.

1. Illustrative Settings, Error Rates, and Procedures

Some important issues arise for all of these alternatives. These issues are considered and illustrated for one example of multiple determination and one example of multiple comparisons.

Consider k independent estimates $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_k$ of, respectively, $\mu_1, \mu_2, \dots, \mu_k$ and an estimate of variability applicable to all these estimates. This is typical enough of multiplicity problems to render conclusions that are broadly helpful.

1.1 Multiple Determination

Defining a standard, λ , and for each μ_i , drawing conclusions about $\mu_i - \lambda$ based on $\bar{y}_i - \lambda$ is, by definition, making a multiple determination. (In some applications, μ_i is a mean change from one time to another.

Often in that case, $\lambda = 0$.) Experience and common sense suggest strongly that μ_i and λ are certain to differ in some decimal place, so that a focus on a null hypothesis that $\mu_i = \lambda$ is quite unrealistic and misleading.

With an interest in $\mu_i - \lambda$ for each value of i , it is reasonable to ask for one of these conclusions: (1) to act as if $\mu_i < \lambda$; (2) to act as if the sign of $\mu_i - \lambda$ is not (yet) determined; or (3) to act as if $\mu_i > \lambda$. Here (1) or (3) is a 'definite' result, while (2) is 'indefinite.'

1.2 Multiple Comparisons

Consider the $k(k-1)/2$ individual pair-wise comparisons $y_i - y_j$. Using these to draw conclusions about the set of $\mu_i - \mu_j$ is, by definition, making multiple comparisons, although the definition is more general, and applies to any set of linear contrasts among the μ_i . Because μ_i and μ_j are certain to differ in some decimal place, there is no good reason to focus on the null hypothesis that $\mu_i = \mu_j$. Instead, it should be asked, for each (i, j) pair, whether one should (1) act as if $\mu_i < \mu_j$; (2) act as if the sign of $\mu_i - \mu_j$ were not (yet) determined; or (3) act as if $\mu_i > \mu_j$.

1.3 In General

Notice the use of ' $>$ ' or ' $<$ ' and the avoidance of ' $=$ ' and ' $\neq$.' 'Accepting the null hypothesis' is usually at best misleading. Confounding ' $>$ ' and ' $<$ ' as ' $\neq$ ' is unresponsive and confusing. Thus, focusing on direction is important when dealing with only one comparison (or determination), and becomes vital when dealing with several or many comparisons.

In these formulations, an 'error' is asserting one direction when the (unknown) truth is the other direction. (One should forget about the 'null case' of equality.) Asserting that the direction is not yet established may be a wasted opportunity, but is not an 'error.' The task is to control error while minimizing wasted opportunity, i.e., minimizing indefinite results.

Two classical choices for controlling errors consider expected error rates over replications of the study. (See, for example, Hochberg and Tamane 1987, p. 7.) They are:

$$E \left\{ \frac{\text{number of errors}}{\text{number of comparisons}} \right\} \leq .025$$

('individual,' 'unadjusted,' or 'per comparison')

and

$$E \left\{ \frac{\text{number of errors}}{\text{number of families}} \right\} \leq .025$$

('simultaneous,' 'Bonferroni,' or 'per family').

Here, 0.025 could easily be replaced by another choice. (For consistency with commonly used notation, the chosen value will be denoted by $\alpha/2$.) ‘Family’ refers to an identified set of comparisons whose simultaneous error rate is to be controlled. Normally, ‘number of families’ would be taken equal to unity. These two choices for controlling errors differ by a factor of k for multiple determinations and of $k(k-1)/2$ for multiple comparisons, so the difference between the two can be quite large.

There has been a feeling for decades (e.g., Tukey 1994/1953) that ‘individual’ or ‘unadjusted’ is too soft while ‘simultaneous’ is too severe. More recently, Benjamini and Hochberg (1995) introduced a compromise that is receiving increasing acceptance.

The two parts of their innovation are a definition of an alternative error rate, the false discovery rate (*FDR*), and a moderately simple procedure that, under the simplest circumstances, controls the *FDR*. The definition is the expectation of the ratio of the number of erroneous assertions of direction to the total number of assertions of direction, i.e.,

false discovery rate =

$FDR =$

$$E \left\{ \frac{\text{number of errors}}{\max\{1, \text{number of definite results}\}} \right\}$$

Let p_{crit} be the tail area (usually for each of two tails) of the null sampling distribution of the test statistic for any single comparison being judged by a multiplicity-respecting procedure. Each such procedure will stipulate a probability of error or average fraction of error that is bounded by $\alpha/2$ when judging confidence of direction. The value of p_{crit} will depend on the sort of confidence to be attained. Let K be the number of determinations or comparisons in the family under consideration, where $K = k$ for multiple determinations and $K = k(k-1)/2$ for all pair-wise multiple comparisons. Let $h = 1, \dots, K$ be the rank of the p -value associated with the test statistic for the comparison concerned when ordered from smallest to largest, so that the observed p -values— $p_{(h)}$ for the h th comparison—are weakly increasing from $h = 1$ to $h = K$. Three distinct procedures, two classical choices and one that controls the *FDR*, are defined as follows:

Unadjusted: be confident of the observed direction for comparison h if

$$p_{(h)} \leq p_{\text{crit}} = p_{\text{UNA}} = \alpha/2$$

FDR (Benjamini and Hochberg 1995): be confident of the observed direction of the h th comparison when $p_{(h)}$ is the largest p -value such that

$$p_{(h)} \leq p_{\text{crit}} = p_{B-H}(h) = (h/K)\alpha/2.$$

Moreover, be confident of the observed direction for all comparisons g with

$$p_{(g)} \leq p_{(h)}, \text{ i.e., } g = 1, \dots, h.$$

Bonferroni (per family): be confident of the observed direction for comparison h if

$$p_{(h)} \leq p_{\text{crit}} = p_{\text{BON}} = (1/K)\alpha/2.$$

Because

$$(1/K) \leq (h/K) \leq 1$$

it must be true that

$$p_{\text{BON}} \leq p_{B-H}(h) \leq p_{\text{UNA}}$$

Specifically, $p_{B-H}(h) = (h)p_{\text{BON}}$ and $p_{\text{UNA}} = (K)p_{\text{BON}}$. The three collections of confident directions corresponding to these three approaches are nested, with the collection for the unadjusted procedure the largest and that for the Bonferroni the smallest.

2. Examples: The NAEP Trial State Assessment

Data from the National Assessment of Educational Progress (NAEP) Trial State Assessment were subjected to two separate analyses by Williams et al. (1999). The data for the first analysis consist of average eighth-grade mathematics proficiency scores for the 34 states that participated in both the 1990 and 1992 NAEP assessments; data for the second analysis consist of average eighth-grade mathematics proficiency scores for the 41 states that participated in 1992 (Johnson et al. 1993).

2.1 Multiple Determination

One way of assessing year-to-year change in state mathematics achievement involves directly testing the significance of change in each of the 34 states. An ordinary Bonferroni adjustment for control of Type I error per family, as was used by NAEP (Johnson et al. 1993), results in an effective critical p -value of $p_{\text{BON}} = 0.000735$ for $\alpha/2 = 0.025$. Table 1 contains the computed difference between 1990 and 1992 mathematics achievement means for each state and the computed pooled standard error. It also includes the statistical decision about confidence in direction (with confident direction indicated by ‘ b ’ in each column) for each state under the three multiplicity treatments. For this example, critical p values are used as a basis for comparison of the adjustment procedures. The cutoff

Table 1

Mean eighth-grade mathematics achievement change, 1990 to 1992 (and standard error), by state, and t , p value, and p_{crit} values for three multiple comparison adjustments, $k = 34^a$

State	$X_{92} - X_{90}$	(s)	t	p value	p_{UNA}	$p_{\text{B-H}} (h)$	p_{BON}
GA	-0.323	(1.77571)	-0.182	0.42814	0.025	0.025000	0.000735
AR	-0.777	(1.48529)	-0.523	0.30141	0.025	0.024265	0.000735
AL	-1.568	(2.01745)	-0.777	0.22004	0.025	0.023529	0.000735
NJ	1.565	(1.92728)	0.812	0.20999	0.025	0.022794	0.000735
NE	1.334	(1.52772)	0.873	0.19320	0.025	0.022059	0.000735
ND	1.526	(1.68552)	0.905	0.18445	0.025	0.021324	0.000735
DE	1.374	(1.34651)	1.020	0.15581	0.025	0.020588	0.000735
MI	2.215	(1.84727)	1.199	0.11761	0.025	0.019853	0.000735
LA	2.637	(2.07943)	1.268	0.10482	0.025	0.019118	0.000735
IN	2.149	(1.63556)	1.314	0.09694	0.025	0.018382	0.000735
WI	2.801	(1.96269)	1.427	0.07936	0.025	0.017647	0.000735
VA	2.859	(1.92992)	1.481	0.07187	0.025	0.016912	0.000735
WV	2.331	(1.39639)	1.669	0.05013	0.025	0.016176	0.000735
MD	3.399	(1.92320)	1.767	0.04113	0.025	0.015441	0.000735
CA	3.777	(2.11460)	1.786	0.03956	0.025	0.014706	0.000735
OH	3.466	(1.85022)	1.873	0.03295	0.025	0.013971	0.000735
NY	4.893	(2.53195)	1.933	0.02901	0.025	0.013235	0.000735
PA	4.303	(2.20545)	1.951	0.02786	0.025	0.012500	0.000735
FL	3.784	(1.93266)	1.958	0.02745	0.025	0.011765	0.000735
WY	2.226	(1.09641)	2.030	0.02339	0.025 ^b	0.011029	0.000735
NM	2.334	(1.14816)	2.033	0.02325	0.025 ^b	0.010294	0.000735
CT	3.204	(1.53443)	2.088	0.02052	0.025 ^b	0.009559	0.000735
OK	4.181	(1.75467)	2.383	0.01018	0.025 ^b	0.008824	0.000735
KY	4.327	(1.61804)	2.674	0.00482	0.025 ^b	0.008088 ^b	0.000735
AZ	4.994	(1.85110)	2.698	0.00452	0.025 ^b	0.007353 ^b	0.000735
ID	2.956	(1.06775)	2.769	0.00374	0.025 ^b	0.006618 ^b	0.000735
TX	5.645	(1.88770)	2.990	0.00202	0.025 ^b	0.005882 ^b	0.000735
CO	4.326	(1.38868)	3.115	0.00141	0.025 ^b	0.005147 ^b	0.000735
IA	4.811	(1.48805)	3.233	0.00100	0.025 ^b	0.004412 ^b	0.000735
NH	4.422	(1.35399)	3.266	0.00090	0.025 ^b	0.003676 ^b	0.000735
NC	7.265	(1.58701)	4.578	0.00001	0.025 ^b	0.002941 ^b	0.000735 ^b
HI	5.550	(1.17134)	4.738	0.00001	0.025 ^b	0.002206 ^b	0.000735 ^b
MN	6.421	(1.35226)	4.748	0.00001	0.025 ^b	0.001471 ^b	0.000735 ^b
RI	5.097	(0.94844)	5.374	0.00000	0.025 ^b	0.000735 ^b	0.000735 ^b
t_{crit}					2.00	2.47	3.33

a Source: Williams et al. 1999. Copyright 1999 by the American Educational Research Association and the American Statistical Association; reproduced with permission from the publisher. b Confident direction of change.

critical t value for each procedure is also displayed in the bottom line of Table 1.

The unadjusted per comparison approach is confident about the direction of 15 differences, whereas the Benjamini–Hochberg (B–H) FDR procedure is confident about 11 of these, and the Bonferroni technique is confident about the direction of only four of the differences.

2.2 Multiple Comparisons

All pair-wise mean differences between the states' 1992 eighth-grade mathematics achievement scores were considered. Because 41 states participated in the 1992

assessment, the family size is $K = 41 \times 40/2 = 820$ comparisons. Using the Bonferroni adjustment, there are 480 confident directions between states, and the use of the B–H technique results in an additional 159. The unadjusted analysis increases the number of confident directions beyond the B–H technique by only six.

2.3 For Both Examples

For each example above, the B–H FDR procedure has led to a greater number of confident directions than the Bonferroni procedure. For the examples, we may expect that no more than 0.025 of the confident

directions identified by the B–H procedure are in error. (This is what it means to control the *FDR*.) As is generally the case, the relative advantage of the *FDR* procedure becomes greater with increasing numbers of comparisons (Kesselman et al. 1999).

3. General Considerations

3.1 Qualitative Status

In a situation where only a very few of the results are definite, the B–H false discovery procedure performs much like the Bonferroni simultaneous approach. Because these situations include those where ‘definite’ is purely accidental, as in extreme instances of ‘data mining,’ it is important that the approach selected resembles a severe procedure, and the *FDR* approach does so.

In a situation where the total number of definite results is a large fraction of the total number of comparisons an algorithm that controls the *FDR* behaves like the unadjusted procedure. If many comparisons deserve a definite result, it is reasonable that the selected approach should not lead to excessively many errors, and again, the *FDR* approach satisfies the criterion.

3.2 Family Size

One of the many problems associated with multiplicity is a frequent strong dependence of result on the number of comparisons in the family. Applying ‘simultaneous’ to Vermont vs. Maine as New England states can give results fairly different from applying it to Vermont vs. Maine as two of the ‘lower 48,’ with a family more than 75 times as large in the latter case. Controlling *FDR* turns out to be much less dependent on family size than the use of simultaneous control.

If the individual elements, e.g., the individual determinations, are stochastically independent, then it is known (Benjamini and Hochberg 1995) that using the *FDR* algorithm does control the rate of false discoveries, and may be unnecessarily conservative. It is reasonable to believe that the result of applying the *FDR* algorithm to elements lightly correlated in the sort of way arising from multiple comparisons will be even more conservative.

For making multiple comparisons in the special case where the true values follow an idealized Gaussian distribution, simulations have shown (Lewis and Tukey 2000) that the algorithm can be made less conservative while still protecting the rate of false discoveries, at least when the number of values being compared is no greater than 25. The modified algorithm follows the B–H approach, but uses $(h/K)\sqrt{(h+1)/2}$ in place of (h/K) in the formula for $p_{\text{B-H}}(h)$, with a caveat that enlargement should stop when the raw modification would otherwise be more liberal than the unadjusted analysis. This modifi-

cation, here denoted by ‘TSR,’ results in the following expression for the series of critical values:

$$p_{\text{TSR}}(h) = \min\{1, (h/K)\sqrt{(h+1)/2}\}(\alpha/2)$$

Simulations indicate that a substantial improvement in power may be attained with this modification to the B–H procedure in the situations for which it is appropriate.

For additional recent work on procedures that control the *FDR*, see Shaffer (1999) and Benjamini and Hochberg (2000).

3.3 ‘Fixed’ and ‘Variable’

A major source of debate associated with multiplicity and factorial design has to do with ‘fixed’ vs. ‘variable,’ i.e., fixed versus random. A significant part of the confusion and conflict can be traced to an unwarranted belief, in tests of hypotheses, that mean squares themselves (or what mean squares estimate) are important. With interest just in directionality, each of the individual questions and answers in multiple comparisons involve $\bar{y}_i - \bar{y}_j$, where i and j are specified, and there is *no sense in asking* whether i and j are fixed or variable. (It may make sense, as shown below, to ask if something else is fixed or variable.)

3.4 A Prototype

The simplest multiple-comparison situation that illustrates the issues is a two-way table of responses with some replication in the cells and questions directed to differences between rows. Starting with k rows and m columns, each comparison involves only two rows. Replacing two rows of measurements by one row of estimated differences yields a 1-by- m table with some replication. (Readers may choose their favorite identities for rows and for columns; in particular, columns = blocks and rows = varieties or treatments gives appropriate realizations.)

If the columns and calculations are fixed, then the uncertainty about the answer is a consequence of whatever makes replicates within a cell different. An estimate of the corresponding variance is, say, an f -error term.

If the columns or calculations would be different were the data to be collected again, in particular, if different applications would require different weights, then the uncertainty about the overall results would reflect two sources of variability: whatever makes replicates differ, and the long-run differences (for a selected comparison) between one column and another.

(The second would not appear if the columns, etc., would be the same.) The resulting error term will be larger than the f -error term; just how much larger will depend on the extent to which differences in columns and calculations are reflected in the overall comparison.

In one crisply specified case where the columns are a random sample from a specifiable infinite population of columns, (and the calculations are fixed), there is a well-determined error term, say, the *r*-error term. Clearly

$$r\text{-error term} > f\text{-error term}$$

In practice, it is not likely to have exactly either columns and calculations fixed, or columns that constitute a simple random sample from a population.

Two alternatives, ‘fixed’ and ‘variable,’ are not enough.

A good way to provide a reasonable amount of realism is to define ‘*c*’ by

appropriate error term

$$= (c)[r\text{-error term}] \\ + (1 - C)[f\text{-error term}]$$

so that ‘everything fixed’ corresponds to $c = 0$ and ‘random’ is a particular case of $c = 1$. It pays then to learn as much as possible about values of ‘*c*’ in the real world.

3.5 *Mathematization vs. Realism*

Those who believe that data analysis can and should be mathematized are likely to be neither happy about the introduction of ‘*c*’ nor highly tolerant of a major role for such a quantity that is not estimable from data. When they recognize, however, that ‘*c*’ is about the data analyst or the data analyst’s client—about just what question is being asked or about what is the most realistic question that can reasonably be asked—then they may be more willing to think about ‘*c*’ and even to work with its default values. Those who recognize the need for more realism today (and even more tomorrow), however, must be prepared for a growing seriousness about ‘*c*’ and a growing impact of nonzero values of ‘*c*’ on the conclusions to be drawn from the analysis of data. Some may find it useful to give ‘*c*’ a Bayesian interpretation, but this is not a necessity.

The only context in which something like being serious about ‘*c*’ has passed into a subject-matter field seems to be climatology. The usefulness of a climatological result is clearly measured by its performance over the set of years from its publication until the publication of its next revision. It will be based on what happened, quantitatively, in a set of years before its construction. Weather is a nontrivial time series (surely not a white noise), so neither $c = 0$ nor $c = 1$ can be expected to do well. Empirical study has led to using a base period for climatology of roughly the last 25 years. This is not directly related to a ‘*c*’ but is of a similar flavor.

Most data analysis intended to illuminate the real world is rather like climatology. The data, whether experimental or historical, has been collected under some circumstances and in the past, while its analysis has been carried out to estimate what will or what would happen in the future. The strength of the link from past to future is important; judging it effectively is not likely to be easy. Difficult issues need to be faced. If analyses are to be increasingly realistic, decisions about those issues will be required.

3.6 *Plausible Defaults for ‘c’*

To begin to face the tasks, some values of ‘*c*’ can be set forth for some crudely described situations. Using such a list as a basis for choosing a ‘*c*’ in real data analyses is likely to be better than doing what has become customary, i.e., to assume either that $c = 0$ or $c = 1$. Values on this list will be likely to change over time as more is learned. (Here ‘columns’ refers to whatever versions are involved.)

- $c = 0.0$ fixed columns and calculation
- $c = 0.1$ fixed columns and somewhat variable calculation
- $c = 0.5$ fixed columns and illustrative weights
- $c = 1.0$ random columns and fixed weights
- $c = 1.4$ somewhat illustrative columns
- $c = 1.6$ illustrative columns
- $c = 2.0$ very illustrative columns

For simple random sampling without replacement from a finite population, the value of ‘*c*’ is $1 - (\text{sample size})/(\text{population size})$.

3.7 *More Quantitative Conclusions*

The focus so far has been on directionality, an appropriate focus when not dealing with crisp measurements, when the hope is to be convinced by the analysis about the definiteness of at least some directions. Especially in some branches of physical science and engineering, the appropriate interest is in the sizes of differences, sizes that may be measured to a small fraction of their nominal value. Instead of directions, the focus will then be upon confidence intervals.

If the concern really is with all the pair-wise differences among k varieties, treatments, or circumstances, or if the pair-wise differences we are concerned about are selected independently of the numbers to be differenced, the Benjamini and Hochberg approach has to regard the results for all differences as definite so that the *FDR* approach could be taken to coincide with the individual unadjusted approach.

If it really is desired to make a simultaneous statement about ‘all differences’ or to validate a process of focusing on the single difference or the few differences that appear to be most different from a reasonable anticipation, the only viable alternative is

to take a simultaneous approach, presumably using percentage points for a Studentized-range distribution.

These two procedures need not yield compatible results. A difference may have a definite direction by *FDR* but a simultaneous confidence interval that includes long-run differences of both signs. Such incompatibility should be regarded only as evidence that the analyst needs to be clear which question is being asked. Is the analyst looking for a few definite directions (or for all the directions that can reasonably be believed to be definite) or is the interest in providing intervals covering the worst that is likely to have happened for any difference?

3.8 Other Kinds of Multiplicity

The broad issues presented here in discussing multiple determination and multiple comparisons appear in most of the other situations where multiplicity matters. Details may vary, but the broad issues are parallel. For more comprehensive treatments of multiplicity, see, for example, Miller (1981), Hochberg and Tamane (1987), and Hsu (1996).

Bibliography

- Benjamini Y, Hochberg Y 1995 Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57**: 289–300
- Benjamini Y, Hochberg Y 2000 On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* **25**: 60–83
- Hochberg Y, Tamane A C 1987 *Multiple Comparison Procedures*. John Wiley & Sons, New York
- Hsu J C 1996 *Multiple Comparisons: Theory and Methods*. Chapman & Hall, New York
- Johnson E G, Mazzeo J, Kline D L 1993 *Technical Report of the NAEP 1992 Trial State Assessment Program in Mathematics*. Report 23-ST05. National Center for Education Statistics, Washington, DC
- Keselman H J, Cribbie R, Holland B 1999 The pairwise multiple comparison multiplicity problem: An alternative approach to familywise and comparisonwise Type I error control. *Psychological Methods* **4**: 58–69
- Lewis C, Tukey J W 2000 Improved multiple comparison procedures for controlling the false discovery rate. Manuscript in preparation
- Miller R G 1981 *Simultaneous Statistical Inference*, 2nd edn. Springer-Verlag, New York
- Shaffer J P 1999 A semi-Bayesian study of Duncan's Bayesian multiple comparison procedure. *Journal of Statistical Planning and Inference* **82**: 197–213
- Tukey J W 1994/1953 The problem of multiple comparisons. In: Braun H I (ed.) *The Collected Works of John W. Tukey, Multiple Comparisons, 1948–1983*. Chapman & Hall, New York, Vol. VIII, pp. 1–300

Williams V S L, Jones L V, Tukey J W 1999 Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics* **24**: 42–69

L. V. Jones, C. Lewis, and J. W. Tukey

Hysteria

The concept of hysteria is almost as old as medicine itself. Explanations of the complaints and behavior of the predominantly female patients suffering from this enigmatic condition have changed many times in response to changing ideas about the causes of disease, but none of these explanations has ever been entirely convincing. For at least the last 100 years physicians have been puzzled by the varied manifestations of the condition, dismayed by its ambiguities, and often exasperated by the behavior associated with it. As a result the term is now fading into history, though the various syndromes it embraced still exist and still await a satisfactory explanation.

1. The History of Hysteria

The concept of hysteria (from the Greek word *hystera* for womb) originated in the second millennium BC. Egyptian priest physicians attributed the paroxysmal symptoms of their female patients to wandering of the womb, and treated them either by the local application of fragrant potions to entice the uterus back to its proper station, or by the oral administration of nauseous draughts to drive it back. These assumptions were inherited by Hippocrates and Galen, and so was born the tenacious idea of a disease of women manifested by recurrent fits and other dramatic symptoms of sudden onset.

The European literature of the last 400 years is summarized by Merskey (1979). Thomas Sydenham was probably the first to recognize explicitly the role of emotion in the genesis of hysteria, for in 1697 he said in his *Discourse* that 'The *External Causes* of this Disease are either violent motion of the Body, or more frequently vehement Agitations of the Mind by the Assaults of Grief, Anger, Fear and such like Passions.' Sydenham was, of course, allowing for 'internal causes' as well, but the identification of these remained elusive. Sir Benjamin Brodie established in the 1830s that the muscles and peripheral nerves of patients with hysterical paralyses or anaesthesias appeared quite normal when examined postmortem, and by the 1870s Sir James Paget was convinced that the problem was one of volition rather than of neurological pathology. As he observed succinctly, 'They say "I cannot"; it looks like "I will not"; but it is "I cannot will."' Even

so, the great French neurologists Charcot and Janet still assumed at the end of that century that the fundamental cause of the suggestibility and hypnotizability of hysterics was a hereditary nervous degeneration. The assumption that hysteria occurred only in women also persisted well into that era. Landouzy stated firmly in 1846 that it was '*exclusive au sexe féminin*' and Charcot regarded ovarian tenderness as one of the stigmata of the condition and attached great significance to the precipitation of hysterical symptoms by pressure on the ovary. Only in World War I, when large numbers of soldiers in all the combatant armies developed florid hysterical symptoms in response to the appalling stresses of trench warfare, was the assumption of an exclusively female illness finally abandoned.

The idea that hysteria was a purely psychological disorder was introduced by Breuer and Freud (1896) in their famous *Studies on Hysteria* and contemporary opposition to their views owed almost as much to the abandonment of neurological pathology as it did to their novel and disturbing ideas about infantile sexuality. According to the corpus of psychoanalytic theory that developed from this revolutionary treatise, hysteria was a specific clinical entity arising from sexual conflicts originating in the oedipal phase of psychosexual development. Failure to relinquish the incestuous tie to the opposite sexed parent that develops during infancy leads to sexual conflicts in adult life because the sexual drive, or libido, retains its forbidden incestuous quality. The drive is, therefore, repressed and its energy converted into a hysterical symptom, which not only protects the patient from conscious awareness of the incestuous wishes, and hence from the anxiety that would otherwise result (primary gain), but simultaneously provides a symbolic expression of it (the rhythmic movements of a hysterical convulsion, for example, expressing the patient's coital fears or phantasies). In addition, the symptom itself often provided other benefits (secondary gain) by rendering the subject visibly ill, or making crucial activities impossible.

Despite the lack of supporting empirical evidence, and the wartime evidence that widespread and florid hysterical symptoms were precipitated by overwhelming stress rather than by infantile sexual experiences or phantasies, this psychoanalytic concept of hysteria remained dominant until the 1960s.

2. Criticisms of the Concept and its Multiple Meanings

The English physician Skey commented as long ago as 1867 that 'One good reason that may be assigned to the persistent employment of the term Hysteria—a term we all know to be objectionable—is the difficulty of finding a substitute for it.' Those objections and that difficulty have persisted ever since. Indeed, the

objections have been expressed forcefully almost continuously from the early years of the twentieth century by German, English, and even French psychiatrists. The most consistent criticisms of hysteria have centered on the multiple meanings of the term and the wide variety of symptoms and behaviors to which it has been applied. At least seven different meanings can be distinguished.

2.1 Neurological Conversion Symptoms

Muscular weakness or paralysis, affecting one, two, or all four limbs, or resulting in loss of speech; convulsions simulating epilepsy; or a loss of sensation (commonly blindness or tunnel vision, deafness, or a circumscribed loss of cutaneous sensation) assumed to depend on the psychoanalytic mechanism of conversion. As the extent of the paralysis or sensory loss reflects the patient's ideas rather than anatomical facts, and as it can often be demonstrated neurologically that the relevant motor or sensory pathways are undamaged, the diagnosis of hysteria can usually be made with some confidence.

2.2 Hysterical Pain and other Psychophysiological Reactions

Visceral symptoms like vomiting or otherwise unexplained pain assumed to be produced by the same conversion mechanism. Unlike neurological conversion symptoms, there is usually no easy way of demonstrating that such symptoms must be psychogenic and as a result the diagnosis of hysteria remains uncertain.

2.3 Dissociative Reactions

Episodes of amnesia, fugue states, and rare cases of multiple personality assumed, after Janet, to be based on a restriction or dissociation of consciousness. Like conversion symptoms, hysterical amnesias and fugues usually develop in highly stressful situations and have the effect of providing at least a temporary respite from that stress.

2.4 Anxiety Hysteria

An obsolete psychoanalytic term for phobic anxiety, implying that repressed anxiety is displaced onto a neutral object or situation instead of being converted into a bodily symptom.

2.5 Epidemic or Mass Hysteria

Since the 'dancing manias' of fourteenth-century Europe, over 100 outbreaks of mass hysteria have been recorded, usually in schools, convents, or other communities of girls or young women, and usually at times of obvious stress (Sirois 1974). Shared appre-

hensions and suggestion are clearly important as those affected are always in close contact with one another and develop almost identical symptoms. Influential members of the community are generally also the first to be affected.

2.6 Briquet's Syndrome

A well-defined pattern of behavior, first described by the French physician Briquet in 1859 and seen almost exclusively in young women. They repeatedly consult doctors with multiple and constantly changing somatic symptoms for which no adequate cause can be found. They also have unsatisfactory sexual lives and are prone to develop mood disturbances and conversion symptoms. This behavior persists into middle age or beyond, and as a result such women acquire voluminous medical files, have multiple abdominal and pelvic operations, and often become dependent on analgesics or tranquillisers (Guze and Perley 1963).

2.7 The Hysterical Personality

A personality type, first clearly described by Janet at a time when hysteria was still regarded as an almost exclusively feminine condition, which is assumed to be particularly susceptible to develop conversion and other hysterical symptoms. As described, the characteristic features are attention-seeking behavior, shallow, labile emotions, and self-centered, demanding, interpersonal behavior. A variety of tactics are used to attract and retain the attention of others, including striking or sexually provocative clothing, flirtatious behavior, flattery, dramatized accounts of their past lives or current circumstances, and extravagant gestures.

Despite the linkages between these seven meanings the inevitable consequence of this diversity of usage was confusion and serious ambiguity. The situation was summarized trenchantly by the American psychiatrist Chodoff (1974): 'Entities that are clinically quite different are being held together artificially by little more than the authority of an ancient name' he asserted; 'the diagnosis (of hysteria) has become a balloon filled with air rather than substance ... it is a fossil encrusted with and obscured by successive layers of meaning.' For good measure he reminded his audience that the term 'hysteric' was used widely by young male psychiatrists to describe almost any attractive female patient.

3. Hysteria Discarded

Although similar criticisms had been voiced many times before, Chodoff's strictures struck home, and when the American Psychiatric Association produced the revolutionary 3rd edition of its *Diagnostic and Statistical Manual* in 1980 the term hysteria had been

discarded (see *Nosology in Psychiatry*). Most of its constituent syndromes survived, however, though with revised titles. Conversion symptoms became conversion disorder and anxiety hysteria became phobic disorder, and both were shorn of their original psychodynamic explanations. Dissociative reactions likewise became dissociative disorders. Briquet's syndrome was renamed somatization disorder, and the hysterical personality was renamed histrionic personality disorder. Mass or epidemic hysteria was not mentioned, presumably because it was assumed that each individual would be categorized separately under the appropriate syndrome. When the World Health Organization produced the next (10th) revision of the International Classification in 1992 its treatment of hysteria was almost identical. The only significant difference was that dissociative and conversion disorders were combined in a single category. Quite suddenly, therefore, one of the oldest, most evocative, and most reviled diagnoses in medicine's nomenclature had been expunged, at least for all official purposes.

4. The Limitations of the 'Medical Model'

Discarding the term hysteria has not, of course, solved the underlying problem of why the patients previously so labeled behaved in the way they did. A large part of that problem is that physicians have traditionally been conditioned by a set of mainly unwritten assumptions—what social scientists call the 'medical model'—to assume that every *bona fide* patient must have a disease of some kind. If the patient's complaints can be related to a pathology of some sort no problem arises. But if they cannot, most doctors tend to oscillate uneasily between two alternative attitudes: either there is nothing wrong with them or they are psychiatrically ill; and both alternatives are often unsatisfactory. The former may be true in organic disease model terms, but it provides no explanation for the patient's behavior and no relief of their suffering. The latter is easily taken to imply that the patient must have some kind of mental illness with a discrete cause of its own. Traditionally, hysteria has been one of these mental illnesses, but only one of them. For physicians have also invoked the concepts of hypochondriasis, neurasthenia, psychophysiological reaction, psychosomatic reaction, 'functional overlay,' and malingering in an attempt to explain their patients' behavior, and all these concepts, except possibly the last, are inadequately defined and overlap with one another.

5. The Sick Role and Illness Behavior

It is increasingly well recognized that the question to be asked when patients consult doctors with complaints that appear to be symptoms of bodily illness but in whom no evidence of illness can be found is often not 'What is the matter with them?' but 'Why are

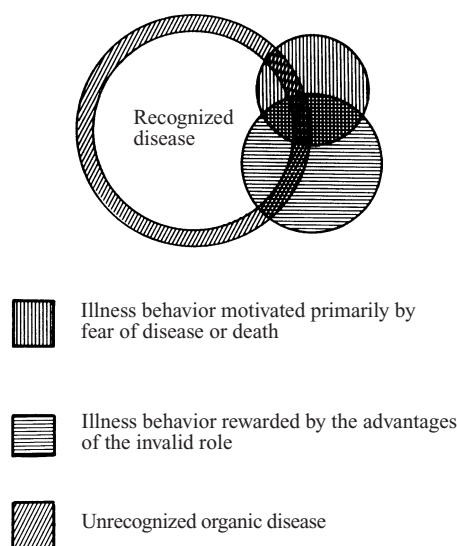


Figure 1

A tentative classification of illness behavior (reproduced by permission of Cambridge University Press from Kendell R E 1983 *Hysteria and the hysterical (histrionic) personality*. In: Russell G F M, Hersov L A (eds.) *The Neuroses and Personality Disorders*. Cambridge University Press, Cambridge, UK)

they behaving in this way?' If the issue is formulated in this way the concepts of the 'sick role' and 'illness behavior' originally formulated by Parsons (1951) and Mechanic (1962) provide the starting point for a better understanding of the problem (Pilowsky 1969). As such writers have emphasized, the sick role in contemporary society involves many privileges. Invalids are exempted from many normal social obligations; children don't have to go to school and adults don't have to go to work, and often continue to be paid nonetheless. They are also exempted from responsibility for their incapacity, and so cannot simply be told to pull themselves together. Indeed, others are under an obligation to be kind and sympathetic, and sometimes to take over their responsibilities as well. The only reciprocal obligation on the patient is to seek and accept appropriate treatment, so that they utilize the privileges of the sick role for as briefly as possible.

Patterns of behavior that are rewarded tend to increase in frequency at the expense of those that are not, and the rewards of the sick role are so substantial, and are often experienced so frequently in childhood, that it is hardly surprising that not all those who consult doctors with bodily symptoms have objective evidence of disease. Against this conceptual background, Kendell (1983) proposed a tentative classification of invalid behavior in the form of a Venn diagram (Fig. 1). Although reality is undoubtedly

more complex, the relationships between the three overlapping populations in this diagram illustrate some basic principles. Most illness behavior is generated by recognized disease (including well-defined psychiatric syndromes like schizophrenia and phobic anxiety). But substantial parts of the whole are generated either by fear of disease or by the 'positive reinforcement' provided by the rewards of the invalid role. There are large and important overlaps between these three populations, and the main recognized disease population is surrounded by a penumbra of unrecognized but genuine disease (mainly people with accepted but so far undiagnosed illnesses, like occult cancers, but perhaps also including subpopulations with as yet unrecognized syndromes, like people with narcolepsy 100 years ago). The management of individual patients depends on which of the 11 populations formed by the intersections of these four circles they come from, rather than on which of a few mutually exclusive diagnostic categories they belong to. Broadly speaking, illness behavior generated by fear of disease corresponds to hypochondriasis and illness behavior motivated by the attractions of the invalid role to what used to be called hysteria.

Essentially, the sick role is attractive, and so liable to be adopted, whenever the balance of its advantages and disadvantages outweighs those of health. To some people the demands of everyday life are so onerous—either because they lack the energy or ability to cope successfully, or because only when ill do they receive sufficient affection, sympathy or attention from other people—that the sick role is preferable most of the time. To many more, the sick role becomes attractive only in situations in which their responsibilities are abnormally heavy (soldiers in battle, for example), or when they are in trouble (impending exams, debt, the threat of blackmail or prosecution, and so on), or when the prospect of financial gain increases the attractions of illness still further (so-called compensation neurosis). The very high incidence of hysterical phenomena in exhausted combat troops suggests that everyone is potentially capable of developing hysterical symptoms, just as they are capable of developing epilepsy, but vary in their liability to do so. This variation presumably depends partly on constitutional variables like suggestibility and capacity to dissociate, partly on past experience, and partly on the subject's current social situation.

6. Contemporary Understanding of Hysterical Behaviors

The social and behavioral perspective described above is now accepted widely and implies that hysteria should be regarded not as an illness, or even as a group of related illnesses, but as a set of ways of reacting to stressful situations. It also provides a plausible ex-

planation of what is known of the epidemiology of hysterical symptoms and of the choice and outcome of individual symptoms, and has important implications for diagnosis and treatment.

6.1 Implications for the Epidemiology of Hysteria

Few community surveys of the incidence or prevalence of psychiatric syndromes have generated satisfactory measures of hysteria, because the range of phenomena embraced by the term is so vague and variable, and because most hysterical phenomena are short-lived. Consequently, most of the epidemiological 'facts' referred to below are derived, *faute de mieux*, from the opinions and experience of clinicians. Subject to this proviso, the reason the incidence of conversion hysteria is falling both in the industrial and the developing world (in India there is epidemiological evidence of this) is probably that in both settings levels of education and sophistication are rising and life becoming less stressful. The reason hysterical symptoms are seen mainly in adolescents and young adults is probably because the adoption of a dependent role comes most readily to those who have only recently forgone the privileges of childhood and are least experienced at coping with the demands of the adult world. The predominance of conversion symptoms in women almost certainly has a similar explanation. In most cultures women have always been dependent on their husbands or fathers. As a result, they have been forced to utilize the few opportunities provided by that dependent role to get their own way in a male dominated world. Only in the last two or three generations, and only in industrial countries, has this slowly ceased to be so, which is probably why the number of young women presenting to doctors with conversion symptoms has fallen steadily since Charcot's day and why the female predominance has now fallen below 2:1. The frequent coexistence of hysteria and organic disease is also readily explained, for those who are genuinely ill or handicapped—by epilepsy or multiple sclerosis, for example—are the very people likely to have the most difficulty coping with the demands of everyday life, and with the most vivid and recent awareness of the rewards of the sick role. Finally, the predominance of left-sided limb paralyses and anesthetics, first remarked upon by Briquet, is explained, for in a right-handed person this minimizes the handicap without reducing the social impact.

6.2 Implications for Choice of Symptom and Outcome

According to psychoanalytic theory the choice of symptom is determined by its symbolic meaning for

the patient, and doubtless this is sometimes so. Usually, though, it is determined by cultural expectations and the subject's own experience of illness, in themselves or others. The nurse who develops hysterical blindness, for example, had severe conjunctivitis as a child, and the young man who loses the use of his legs had a grandfather who was housebound for years after a stroke.

Just as the development of hysterical symptoms is encouraged strongly by situations in which the balance of the advantages and disadvantages of sickness outweigh those of health, so recovery is encouraged by a reversal of that balance of advantage. This is why the symptom usually fades away if the situation that provoked its appearance resolves itself. If that situation persists, however, or if the inadequacies of the subject are paramount, the advantages and disadvantages of the chosen symptom become important. This is why disabling symptoms, like blindness or inability to stand or walk, have a comparatively good prognosis, whereas those which can be switched on or off as occasion demands, like fits or vomiting, or cause little disability, like tremors, are more intractable.

6.3 Implications for Investigation and Treatment

There are two implications for the investigation of patients with suspected hysterical symptoms, both well known. The first is that medical attention, admission to hospital, and glamorous investigations may easily potentiate the attractions of invalidism. (Indeed, many psychiatrists suspect that the fascinating phenomenon of multiple personality is largely the product of excessive medical interest.) For this reason it is important that the investigations needed to make sure that the patient's symptoms are indeed hysterical are completed as quickly as possible, and not repeated without good reason. The second is that it is immaterial whether or not the patient is aware that he or she is in some sense feigning illness, because the treatment is the same either way. Although the dilemma usually was avoided by assuming unconscious motivation, the traditional medical or psychoanalytic approach should require a distinction between conscious motivation (which is malingering, and so subject to strong disapproval) and unconscious motivation (which is an illness—hysteria—and so worthy of sympathy and treatment) in spite of the fact that it is generally impossible to distinguish the two. Indeed, the patient's degree of awareness often varies from time to time, so that what is initially unconscious becomes conscious, and vice versa.

There are always two components to the management of hysterical symptoms: an attempt to form a psychotherapeutic relationship with the patient, and an assessment of the elements of their social environment which are encouraging or discouraging

invalid behavior. The former is never straightforward, and inexperienced therapists often allow themselves to be manipulated by flattery or coquetry, or end up angrily rejecting the patient because they resent being manipulated, or indeed do both these things in succession. Occasionally it is possible by gentle degrees to lead the patient into a discussion of the emotional predicament from which their symptoms arose, and of the relationship between that predicament and their symptoms. More often there is a sustained denial of emotional difficulties of any kind, and it becomes increasingly clear that the subject is more interested in being a patient than in being cured, and curiously uninterested in the origins of their symptoms.

Arriving at an adequate understanding of the conflicting emotional and social pressures on the patient, both those encouraging and discouraging invalidism, and assessing the possibilities for strengthening the latter, usually involves talking to other key figures in their life as well as the patient. If the patient no longer needs to be ill, they either get better spontaneously, at least in the short term, or do so in response to any treatment that allows them to recover without loss of face. If, on the other hand, there are cogent reasons why invalidism is still more attractive than health, attempts have to be made to modify the attitudes and behavior of those with most influence on the balance of advantage and disadvantage referred to above, and at times this involves devoting more time to a key relative than to the patient. In the past hypnosis was used widely in the treatment of hysteria, and there is no doubt that, thanks to the suggestibility of people with hysterical symptoms, hypnosis is often dramatically successful in the short run. But unless the attractions of invalidism are also tackled, or the underlying predicament has resolved itself, success is usually short-lived.

6.4 The Limitations of Contemporary Understanding

Although this contemporary social and behavioral view of hysterical behaviors explains many otherwise puzzling clinical observations, it provides no

explanation of the mechanisms of conversion and dissociation. The explanatory theories proposed originally by Freud and Janet over 100 years ago have been abandoned, but nothing has yet been offered in their place and the neurophysiology of both phenomena remains a mystery. Indeed, shorn of the original psychoanalytic assumption that unconscious anxiety was being converted into a neurological deficit, the term conversion disorder has lost all meaning. And hysteria itself has been discarded. But as Lewis observed a generation ago, 'a tough old word like hysteria dies very hard.' It may yet come back.

See also: Anxiety and Anxiety Disorders; Anxiety and Fear, Neural Basis of; Culture as a Determinant of Mental Health; Freud, Sigmund (1856–1939); Mental and Behavioral Disorders, Diagnosis and Classification of; Psychoanalysis, History of; Somatoform Disorders

Bibliography

- American Psychiatric Association 1980 *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn. (DSM-III). American Psychiatric Association, Washington, DC
- Breuer J, Freud S 1896 Studies on hysteria. In: *Standard Edition of the Complete Psychological Works of Sigmund Freud*, Vol. 2. Hogarth Press, London
- Chodoff P 1974 The diagnosis of hysteria: An overview. *American Journal of Psychiatry* **131**: 1073–8
- Guze S B, Perley M J 1963 Observations on the natural history of hysteria. *American Journal of Psychiatry* **119**: 960–5
- Kendell R E 1983 Hysteria and the hysterical (histrionic) personality. In: Russell G F M, Hersov L A (eds.) *The Neuroses and Personality Disorders*. Cambridge University Press, Cambridge, UK
- Lewis A 1975 Survival of hysteria. *Psychological Medicine* **5**: 9–13
- Mechanic D 1962 The concept of illness behavior. *Journal of Chronic Diseases* **15**: 189–94
- Merskey H 1979 *The Analysis of Hysteria*. Bailliere Tindall, London
- Parsons T 1951 *The Social System*. Free Press, New York
- Pilowsky I 1969 Abnormal illness behavior. *British Journal of Medical Psychology* **42**: 347–51
- Sirois F 1974 Epidemic hysteria. *Acta Psychiatrica Scandinavica*, **51**: 7–44

R. E. Kendell

Copyright © 2001 Elsevier Science Ltd.
All rights reserved.